

MPRA

Munich Personal RePEc Archive

Stochastic choice, systematic mistakes and preference estimation

Breitmoser, Yves

28 July 2016

Online at <https://mpra.ub.uni-muenchen.de/72779/>
MPRA Paper No. 72779, posted 31 Jul 2016 04:45 UTC

Stochastic choice, systematic mistakes and preference estimation

Yves Breitmoser*
Humboldt University Berlin

July 28, 2016

Abstract

Individual choice exhibits “presentation effects” such as default, ordering and round-number effects. Using existing models, presentation effects bias utility estimates, which suggests instability of preferences and obscures behavioral patterns. This paper derives a generalized model of stochastic choice by weakening logit’s axiomatic foundation. Weakening the axioms implies that focality of options is choice-relevant, alongside utility, which entails presentation effects. The model is tested on four well-known studies of dictator games exhibiting typical round-number patterns. The generalized logit model captures the choice patterns reliably, substantially better than existing models: it robustly predicts and controls for the round-number effects, thus provides “clean” utility estimates that are stable and predictive across experiments.

JEL-Code: D03, C10, C90

Keywords: stochastic choice, systematic mistakes, axiomatic foundation, utility estimation, dictator game

*I thank Nick Netzer, Martin Pollrich, Sebastian Schweighofer-Kodritsch, Georg Weizsäcker and audiences at the BERA workshop in Berlin and at THEEM 2016 in Kreuzlingen for many helpful comments. Financial support of the DFG (project BR 4648/1) is greatly appreciated. Address: Spandauer Str. 1, 10099 Berlin, Germany, email: yves.breitmoser@hu-berlin.de, Telephone/Fax: +49 30 2093 99408/5619.

1 Introduction

Economic studies analyze individual choice in order to understand preferences, amongst others social, time and risk preferences. These analyses make for a large and important literature, as a reliable understanding of preferences is required as basis for the theoretical analyses and counterfactual predictions underlying policy recommendations. Preferences often appear to be unstable, however, as small details in presentation substantially affect choice patterns. Examples include default effects (McKenzie et al., 2006; Dinner et al., 2011), left-digit effects (Pollock and Schwartz, 1984; Lacetera et al., 2012), round-number effects (Heitjan and Rubin, 1991; Manski and Molinari, 2010), and positioning effects (Dean, 1980; Miller and Krosnick, 1998; Feenberg et al., 2015). Absent a model of such presentation effects that would allow researchers to control for them, they induce biased and incoherent utility estimates. Thus, presentation effects may be a major reason for the failure to reach a consensus on preference theories that has been plaguing behavioral analyses (see, e.g., the recent critique by Levine, 2012). Moreover, without a predictive model of how choice and welfare depend on presentation, the effects of "nudging" interventions remain elusive.

The present paper shows that presentation effects can be modeled similarly to stochastic mistakes and in fact represent a generalization thereof. I axiomatically derive a model of stochastic choice with presentation effects by weakening the axioms underlying multinomial logit. Put briefly, in this generalized model two attributes of options affect choice, utility and focality, and focality captures the choice-relevant implications of presentation. I then apply this "focal choice adjusted logit" model (FOCAL) to data from controlled experiments to test the above intuition: First, do existing models indeed yield significantly biased estimates that suggest preference instability? Second, does the FOCAL model capture presentation effects to the point that utility estimates are coherent across studies? Based on an analysis of predictive accuracy and robustness across experiments on generalized dictator games, both empirical questions will be answered in the affirmative: Existing models indicate preference instability even across simple dictator games, while controlling for focality factors out presentation effects and yields stable, even predictive estimates. This exercise showcases how the FOCAL model is able to both (i) facilitate the emergence of a consensus in behavioral modeling and (ii) reliably predict presentation effects, as envisaged by various "nudging" interventions.¹

As point of departure, let us revisit the axiomatic foundation of logit (McFadden, 1974), as analyzed in Breitmoser (2016). Consider a decision maker (DM) with utility function u who chooses $x \in B$ with probability $\Pr(x|u, B)$. If the choice probabilities satisfy positivity and independence of irrelevant alternatives (IIA), they are functions of choice

¹The analysis of predictive adequacy relates the paper to the literature studying predictive adequacy in e.g. decision under risk (Harless and Camerer, 1994; Wilcox, 2008; Hey et al., 2010) and learning (Camerer and Ho, 1999). These studies analyze binomial choice, which presumably exhibits negligible presentation effects. Multinomial choice, and in particular numerical choice, generally exhibits strong presentation effects (most notably round-number effects). Establishing predictive adequacy in this context is novel. The relation to studies of behavioral welfare economics, e.g. Kőszegi and Rabin (2008), is discussed below.

“propensities” v such that

$$\Pr(x|u, B) = \frac{\exp\{v(x, y|u)\}}{\sum_{x' \in B} \exp\{v(x', y|u)\}} \quad \text{with} \quad v(x, y|u) = \log \left(\frac{\Pr(x|u, B)}{\Pr(y|u, B)} \right)$$

for all $x \in B$ in relation to a benchmark option $y \in B$. Formally, v represents the log-odds of the choice between x and y . Without further restrictions, v may be any function of $u(x)$ and $u(y)$. McFadden (1974) assumes $v(x, y|u) = u(x) - u(y)$ in Axiom 3 (“Irrelevance of Alternative Set Effect”). Given v ’s definition, this is equivalent to assuming

$$\frac{\Pr(x|u, \{x, y\})}{\Pr(y|u, \{x, y\})} = \exp\{u(x) - u(y)\} \quad \Leftrightarrow \quad \Pr(x|u, \{x, y\}) = \frac{\exp\{u(x)\}}{\exp\{u(x)\} + \exp\{u(y)\}},$$

i.e. to assuming that binomial choice is logit. Thus, binomial logit is assumed by axiom and itself not independently founded. (IIA merely extends binomial logit to multinomial choice.) Breitmoser (2016) shows that logit is founded in axioms called “narrow bracketing” and “absence of systematic mistakes”. Dropping the latter axiom implies that choice probabilities take the generalized logit form

$$\Pr(x|u, B) = \frac{\exp\{\lambda \cdot u(x) + w(x)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u(x') + w(x')\}}.$$

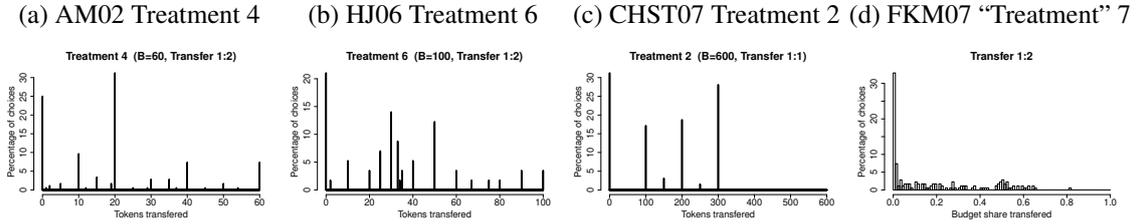
In general, that is, choice propensities are predicted to depend on two option attributes, $u(x)$ and $w(x)$. By assumption, the former is DM’s utility, and the second component therefore induces systematic deviations from maximizing utility—i.e. it captures “systematic mistakes”. Thus, the systematic mistakes in choice are theoretically predicted, but had been unknowingly assumed away in logit. Since $w(x)$ is an attribute of option x that is independent of utility, it is intuitively interpreted as x ’s focality—capturing a notion debated at least since Schelling and relating intimately to the received understanding of systematic mistakes in choice. For example, $w(x)$ may reflect that default options are relatively focal, that top-left or bottom-right positioning elevates focality, that labeling, coloring, or level of “roundness” affects focality, and so on, thus inducing deviations from utility maximization and capturing the various manifestations of presentation effects.

Building on this observation of Breitmoser (2016), note first that generalized logit offers a framework, but not yet an applicable model of choice. While $w(x)$ intuitively relates to focality, the formal relation is indeterminate. If a focality index ϕ is given, physically or neuro-economically measured, assumed based on previous work, or to be estimated econometrically, then $w(x)$ may be an arbitrary function of $\phi(x)$. Section 2 shows that an axiom capturing “relativity of focality” implies that the choice probabilities are uniquely represented by “focal choice adjusted logit” (FOCAL),

$$\Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x) + \kappa \cdot \phi(x)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u(x') + \kappa \cdot \phi(x')\}}.$$

Given focality ϕ , FOCAL offers a simple, single-parameter extension of logit that is readily

Figure 1: Dictator choice across experiments



Note: Andreoni and Miller (2002, AM02) and Harrison and Johnson (2006, HJ06) analyze generalized dictator games with numerical choice entry and budgets up to 100 tokens; choices are predominantly multiples of 10. The difference between AM02 and HJ06 is that the Leontief choice is not generally a multiple of 10 in HJ06. Cappelen et al. (2007, CHST07) analyze dictator games with numerical choice entry and budgets up to 1600 tokens; choices are predominantly multiples of 100. Fisman et al. (2007, FKM07) analyze generalized dictator games with graphical user interface and budgets up to 100 tokens; choice exhibit no round-number effects. In FKM07, treatments as such are not defined, as budget sets and transfer rates are individually randomized; the plot above therefore shows the budget share transferred. Plots for all treatments of all experiments are provided as supplementary material.

applicable using existing methods—while being theoretically attractive: It is derived from comparably general axioms reflecting standard practice even in least-squares analyses, it captures a general intuition about choice debated at least since Schelling, and it is a formal representation of informal arguments often made in view of presentation effects.²

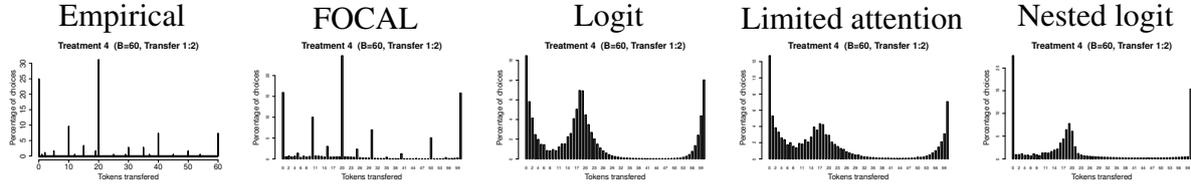
We do not know the focality index ϕ , however. This may suggest that controlling for focality might in general not improve robustness of estimates, as we would likely overspecify the focality index, in which case we risk misspecification, or underspecify it, in which case we risk overfitting. We need to specify ϕ only up to linear transformation, though, which extends the range of appropriate specifications and largely resolves these concerns. For example, if focality is bimodal, either high or low as in default or positioning effects, the focal option may be assigned focality 1 and the other option focality 0 without any loss of adequacy. In other cases, the qualitative properties of focality are understood, as in ordering and round-number effects. Since ϕ needs to be defined only up to linear transformation, such a qualitative understanding seems sufficient to control for focality at least in substantial manner.³ To test this hypothesis, a case where focality is qualitatively understood but certainly not trivial will be analyzed in detail.

Specifically, I cross-analyze existing data from multiple influential experimental studies exhibiting one of the most prominent presentation-related biases: round-number effects. For a transparent analysis of utility in relation to focality, data sets on “generalized dictator games” are particularly suitable: there is consensus on the functional form of util-

²For example, if there is a default option, this option has relatively high focality $\phi(x)$, which by FOCAL induces the default effect. If options are ordered and say the first option is “focal” (i.e. has high relative focality), ordering effects result, and similarly other presentation effects are usually related to focality.

³The focality index ϕ may contain flexible parameterization to verify the qualitative understanding. If focality is relevant but not qualitatively understood, focality can be studied the same way as utility is studied.

Figure 2: The adequacy of the choice models in capturing numerical choice: AM02 treatment 4, with budget 60 and transfer ratio 1 : 2. See Figure 3 for further plots



Note: These plots depict the predicted choice distributions for AM02’s treatment 4 after fitting the model parameters to all data from AM02. The econometric methodology is standard and described in Section 4 and Appendix A; all models are described in Section 2. Besides FOCAL and logit, I consider more two benchmark models prominently discussed in the literature: Limited attention (Manzini and Mariotti, 2014; Echenique et al., 2014) and the cross-nested logit model allowing for similarity effects between proximate options (Ordered GEV, Small, 1987). More plots are provided below.

ities (CES), consensus on the nature of the presentation effects (round-number effects), and no additional influences due to e.g. risk or uncertainty. Further, the literature on generalized dictator games is unusually rich, containing four extensive within-subject studies of dictator choice varying only presentation—including a graphical experiment without round-number effects (Fisman et al., 2007). This allows me to analyze reliability directly. Figure 1 provides an overview of the data sets, Section 3 describes them in detail.

To put the results into context, I relate FOCAL to three benchmark models capturing the main ideas in the existing literature: multinomial logit, limited attention, and nested logit (capturing similarity effects in choice). Section 4 analyzes the models’ adequacies to replicate the basic choice patterns in-sample. This clarifies how the models represent the observations internally and highlights their respective strengths and weaknesses. The differences are striking, see Figure 2. FOCAL reproduces the choice patterns rather accurately. Logit, despite having just one parameter less, fails in this respect and predicts choice patterns that do not resemble the observed ones: by ignoring the possibility of systematic mistakes, logit assumes differences in choice probabilities indicate differences in utility, which is inadequate in the presence of round-number effects and yields unstable preference estimates. Limited attention and nested logit are in-between FOCAL and logit, but only slightly improving on logit (detailed discussions follow).

FOCAL’s ability to capture the choice patterns suggests that the theoretically predicted separation of utility and focality is adequate, that the standard assumption of CES utilities is adequate, and that FOCAL’s preference estimates may thus be reliable. Section 5 investigates these hypotheses in an analysis of predicting preferences, precision and choice bias across experiments. They all receive overwhelming support: separating utility and focality yields reliable utility estimates not just theoretically, but also in typical experimental data sets. Even the extent of the choice bias, i.e. parameter κ , is robust and predictable across experiments. In turn, existing models such as logit, nested logit and limited attention yield unreliable estimates and indicate significant differences of preferences

even between simple dictator game experiments.

For its first test-case of social preference theory, FOCAL provides a rather positive perspective. If preferences were found to vary from application to application unpredictably, even across simple dictator game experiments, then differences to other games would be fairly uninformative and reliable application would be impossible. Having seen that these differences are likely due only to the inadequacy of prior choice models in capturing presentation effects and that the differences are resolved using adequate choice models suggests that a robust understanding of (social) preferences and the emergence of a consensus on their modeling are attainable. The results also indicate that experimental analyses need to analyze and control for (systematic) mistakes even if mistakes are not of explicit interest to the analyst. Systematic mistakes due to presentation are at the center of interest in behavioral welfare economics, and in this respect, FOCAL provides a first general model for econometric and theoretical analyses of nudging. The concluding Section 6 discusses these results and some implications for econometric and experimental analyses e.g. regarding graphical user interfaces following Fisman et al. (2007). The supplementary material contains robustness checks and some additional analysis.

2 Modeling choice with systematic mistakes

2.1 Utility and focality

The decision maker (DM) chooses an option $x \in B$ from a finite budget $B \subset X$. Each option induces an outcome vector of dimensionality $n \geq 1$ (say a payoff vector), denoted as $\pi : X \rightarrow \mathbb{R}^n$. Further, each option is presented to DM in some way, as described by the injective function $l : X \rightarrow L$ (not to be specified further) capturing the choice-relevant aspects of e.g. the default setting, positioning, coloring, or the number representing x . Utility $u : \mathbb{R}^n \rightarrow \mathbb{R}$ maps outcome vectors to utilities, and focality $\phi : L \rightarrow \mathbb{R}$ maps the presentation of options to their focality in the eyes of DM. Thus, utility and focality are orthogonal in the sense that changing the payoffs associated with options does not affect their focality and vice versa.⁴ For later reference, let me formalize this notion.

Definition 1 (Orthogonality). Utility u and focality ϕ are orthogonal if for any pair of decision tasks $\langle B, \pi, l \rangle$ and $\langle B, \pi', l' \rangle$,

$$\pi = \pi' \quad \Rightarrow \quad u(\pi(x)) = u(\pi'(x)) \text{ for all } x \in B, \quad (1)$$

$$l = l' \quad \Rightarrow \quad \phi(l(x)) = \phi(l'(x)) \text{ for all } x \in B. \quad (2)$$

Keeping orthogonality in mind, we may simplify notation by dropping references to outcome vectors and presentation, writing $u(x) = u(\pi(x))$ and $\phi(x) = \phi(l(x))$ for all x .

⁴Focality in this sense captures presentation effects that are independent of payoffs, but alternatively, payoffs as such may also affect focality, see e.g. Bordalo et al. (2012) and Kőszegi and Szeidl (2013).

Given u and ϕ , the probability that DM chooses $x \in B$ from budget $B \subseteq X$ is denoted as $\Pr(x|u, \phi, B)$. As point of departure, assume that budget B is derived from a “rich” choice environment X , that choice probabilities are positive and satisfy IIA, and that choice exhibits a form of narrow bracketing (following Breitmoser, 2016).

Axiom 1 (Richness). There exist $x, x' \in X$ such that $u(x) \neq u(x')$. For all $x, x' \in X$ and all $\lambda \in [0, 1]$, there exists $x'' \in X$ such that $u(x'') = \lambda u(x) + (1 - \lambda) u(x')$.

Axiom 2 (Positivity). $\Pr(x|u, \phi, B) > 0$ for all $x \in B \subseteq X$.

Axiom 3 (Independence of Irrelevant Alternatives, IIA). For all $B, B' \subseteq X$,

$$\frac{\Pr(x|u, \phi, B)}{\Pr(y|u, \phi, B)} = \frac{\Pr(x|u, \phi, B')}{\Pr(y|u, \phi, B')} \quad \text{for all } x, y \in B \cap B'.$$

Axiom 4 (Narrow bracketing). For all $r \in \mathbb{R}$ and $x \in B \subseteq X$, $\Pr(x|u, \phi, B) = \Pr(x|u+r, \phi, B)$.

These axioms either reflect standard practice or are technically innocuous. Positivity and Richness are innocuous in that they are neither restrictive nor falsifiable (as probabilities can be arbitrarily low). Jointly with positivity and richness, IIA implies that choice probabilities are functions of choice propensities V or log-propensities $v := \log V$ as in

$$\Pr(x|u, \phi, B) = \frac{V(x|u, \phi)}{\sum_{x' \in B} V(x'|u, \phi)} \quad \text{or} \quad \Pr(x|u, \phi, B) = \frac{\exp\{v(x|u, \phi)\}}{\sum_{x' \in B} \exp\{v(x'|u, \phi)\}}.$$

IIA is innocuous in the sense that the relationship of propensities V (or, v) and utility or focality is entirely unrestricted, i.e. $v(x)$ may be an arbitrary function of $u(x)$ and $\phi(x)$. IIA is considered restrictive in the presence of similarity effects, which may lead DM to partition the options into “nests” and then to first choose a nest and second an option from a nest (see e.g. the red-bus/blue-bus problem of Debreu, 1960). Such hierarchical choice violates IIA and can be modeled by “nested logit” as discussed below. In case similarity affects choice, FOCAL can be generalized straightforwardly to allow for nesting. Studies of numerical choice rarely relax IIA, however, noting that IIA is also obeyed in least squares analyses, which renders IIA in this context standard practice. I further discuss a model relaxing IIA below. Finally, narrow bracketing (Read et al., 1999) holds if DM considers the analyzed choice tasks in isolation, i.e. independently of his level of utility outside the experiment, which again is standard practice in experimental analyses.

Given orthogonality of utility and focality, these axioms imply that choice probabilities are generalized logit (Breitmoser, 2016), i.e. log-propensities are linear in u .

Definition 2 (Generalized logit). Consider a DM with utility function u and focality function ϕ . The choice probabilities are generalized logit if there exist $\lambda \in \mathbb{R}$ and $w : X \rightarrow \mathbb{R}$ such that for all finite $B \subseteq X$ and all $x \in B$,

$$\Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x) + w(x|\phi)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u(x') + w(x'|\phi)\}}. \quad (3)$$

Briefly, let me clarify the main relations to the literature (for a detailed discussion, see Breitmoser, 2016). Generalized logit results if we assume choice satisfies positivity, IIA and narrow bracketing (besides richness of the environment). All three assumptions are equally satisfied in least squares analyses. Applications of least squares make additional assumptions on the irrelevance of utility differences between options and on the relevance of the options' quadratic distances to the utility maximizer. These assumptions are not supported empirically, and they are made on top of logit's assumptions. Thus, least squares makes strictly more assumptions than generalized logit, and hence is strictly less general, despite the seemingly specific functional form of generalized logit.

Generalized logit posits that utility and focality are additively separable. Their distinction is not identified based on choice data alone, i.e. we cannot say if w actually is part of u or not. This reproduces a central result in behavioral welfare economics (Kőszegi and Rabin, 2008; Gul and Pesendorfer, 2008; Bernheim and Rangel, 2009). While it is not a concern in positive analyses of choice, it will be relevant in normative analyses of say nudging. Relatedly, if DM does not know the utilities associated with his options, but may learn about utilities by studying options, he has to decide which options to study and when to stop. If DM does so rationally, choice has a generalized logit representation similar to above (Matejka and McKay, 2015), but this generalization of logit captures similarity effects including the red-bus/blue-bus problem. Thus, both the above axiomatic analysis capturing presentation effects and the rational-inattention analysis capturing similarity effects yield the same structure of choice propensities, and both reproduce the behavioral welfare result that true utility u and bias w cannot be separated based on choice data alone. This observation, that three fundamentally different approaches to choice analysis converge in the generalized logit model, appears to be rather reassuring.

2.2 Focal choice adjusted logit

The remainder of the paper uses these general observations of Breitmoser (2016) to develop and test a model of choice with systematic mistakes. As indicated, generalized logit predicts that options have two choice-relevant attributes, utility $u(x)$ and an attribute $w(x)$ relating to focality $\phi(x)$. So far, however, $w(x)$ may be an arbitrary function of $\phi(x)$.

To fix ideas, let me define a focality index ϕ capturing relative focality of round numbers. Econometric models of rounding have been proposed in a growing literature starting with Heitjan and Rubin (1991) for consumption data and Manski and Molinari (2010) for subjective beliefs.⁵ This literature fairly consistently finds that 100, 50, 10, 5, 1, 0.5, 0.1, ... exhibit decreasing levels of roundness (Battistin et al., 2003; Whynes et al., 2005; Covey and Smith, 2006). Thus, let us say that the focality index of a number x is the level of the highest number in this sequence that divides x .

⁵This literature studies continuous-choice models to be applied in analyses of survey data. Utility functions are not discussed and utility differences between options are choice-irrelevant by assumption, as these models are not intended (and thus not appropriate) to estimate utilities from dictator games.

Definition 3 (Focality index ϕ). Given the smallest notable difference $\varepsilon > 0$, a potentially negative power of 10, define $\lfloor x \rfloor_\varepsilon$ as x rounded down to the nearest multiple of ε . Further, define an index $i : \mathbb{Z} \rightarrow \mathbb{R}$ such that $i(0) = 1$, $i(1) = 5$, and $i(k) = 10 \cdot i(k-2)$ for all $k \in \mathbb{Z}$. Then, for all $x \neq 0$: $\phi(x) = \max\{k \in \mathbb{Z} \mid i(k) \text{ divides } \lfloor x \rfloor_\varepsilon\}$, and $\phi(0) = \phi(10)$.

This definition is very stylized, by assuming equidistant levels and by equating the level of “0” with that of 10. These assumptions can easily be relaxed, e.g. by introducing parameters to be estimated econometrically, but for transparency I restrict the degrees of freedom to a minimum. This freedom also concerns the definition of the base level of focality, i.e. the options x that have zero focality $\phi(x) = 0$. The above definition states that plain integers have zero focality, but there does not seem to be a rational justification for any such assumption, as focality inherently is relative. Therefore, I formally require that choice probabilities be invariant with respect to changes of the base level.

Axiom 5 (Relativity). For all $r \in \mathbb{R}$ and $x \in B \subseteq X$, $\Pr(x|u, \phi, B) = \Pr(x|u, \phi + r, B)$

Relativity of focality is shown to imply that choice probabilities are linear in focality,

$$\Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x) + \kappa \cdot \phi(x) + c(x)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u(x') + \kappa \cdot \phi(x') + c(x')\}}. \quad (4)$$

The constant $c(x)$ is independent of both utility u and focality ϕ and has a fairly intuitive explanation. Technically, we have not yet clarified that we captured all (known) factors of choice, and $c(x)$ offers the possibility to include additional sources of say focality. As multiple sources of focality may as well be captured using a single focality index, this is technically not necessary, but it is logically consistent and illustrates how further extensions of logit can be achieved. The following axiom implies that choice is fully described by utility and focality.

Axiom 6 (Decision utility = true utility + focality). Given any $B \subseteq X$ and any bijective function $f : B \rightarrow B$, $\Pr(f(x)|u, \phi, B) = \Pr(x|u \circ f, \phi \circ f, B)$ for all $x \in B$.

Technically, Axiom 6 yields $c(x) = \text{const}$ by requiring a form of permutation invariance. This implies that all $c(x)$ cancel out and thus focal choice adjusted logit.

Definition 4 (Focal choice adjusted logit, FOCAL). Consider a DM with utility function u and focality function ϕ . The choice probabilities are FOCAL if there exist $\lambda, \kappa \in \mathbb{R}$ such that for all finite $B \subseteq X$ and all $x \in B$,

$$\Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x) + \kappa \cdot \phi(x)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u(x') + \kappa \cdot \phi(x')\}}. \quad (5)$$

The following proposition formally establishes the result.

Theorem 1. Consider a decision maker with utility u and focality ϕ satisfying orthogonality, with choice probabilities $\Pr(\cdot)$ satisfying positivity (Axiom 2) in a choice environment

satisfying richness (Axiom 1). Then, $\Pr(\cdot)$ satisfies Axioms 3, 4, 5, and 6 if and only if the choice probabilities are FOCAL.

Proof. I proof that the axioms imply that choice probabilities are FOCAL; it is straightforward to verify that FOCAL satisfies the axioms. By orthogonality, Axioms 2–4 imply that choice probabilities satisfy (see Breitmoser, 2016)

$$\Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x) + w(x|\phi)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u(x') + w(x'|\phi)\}} = \frac{\exp\{\lambda \cdot u(x)\} \cdot \exp\{w(x|\phi)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u(x')\} \cdot \exp\{w(x'|\phi)\}}. \quad (6)$$

Hence, there exists a function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$\Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x)\} \cdot f(\phi(x))}{\sum_{x' \in B} \exp\{\lambda \cdot u(x')\} \cdot f(\phi(x'))}, \quad (7)$$

and by Axiom 5, for all $r \in \mathbb{R}$,

$$\Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x)\} \cdot f(\phi(x) + r)}{\sum_{x' \in B} \exp\{\lambda \cdot u(x')\} \cdot f(\phi(x') + r)}. \quad (8)$$

The probabilities are invariant in r only if there exists a function $g : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$\Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x)\} \cdot f(\phi(x)) \cdot g(r)}{\sum_{x' \in B} \exp\{\lambda \cdot u(x')\} \cdot f(\phi(x')) \cdot g(r)}, \quad (9)$$

i.e. $f(y + r) = f(y) \cdot g(r)$ for all r , which implies $f'(y) = f(y) \cdot g'(0)$ and thus $f(y) = \exp\{a \cdot y + b\}$ for some $a, b \in \mathbb{R}$. Hence, there exist $\kappa \in \mathbb{R}$ and $w : X \rightarrow \mathbb{R}$ such that

$$\Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x)\} \cdot \exp\{\kappa \cdot \phi(x) + c(x)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u(x')\} \cdot \exp\{\kappa \cdot \phi(x') + c(x')\}}. \quad (10)$$

Finally, by Axiom 6 this implies that the choice probabilities are FOCAL. For contradiction, assume the opposite, i.e. Eq. (10) and Axiom 6 are satisfied, but choice probabilities are not FOCAL, i.e. $c(x) \neq \text{const}$. Fix $x, y \in X$ such that $c(x) \neq c(y)$, $B \subseteq X$ such that $x, y \in B$, and a bijection $f : B \rightarrow B$ such that $f(y) = x$ and $f(x) = y$. Thus,

$$\frac{\Pr(x|u, \phi, B)}{\Pr(y|u, \phi, B)} = \frac{\Pr(f(y)|u, \phi, B)}{\Pr(f(x)|u, \phi, B)}, \quad (11)$$

and by Axiom 6

$$\frac{\exp\{\lambda \cdot u(x) + \kappa \cdot \phi(x) + c(x)\}}{\exp\{\lambda \cdot u(y) + \kappa \cdot \phi(y) + c(y)\}} = \frac{\exp\{\lambda \cdot u(f(y)) + \kappa \cdot \phi(f(y)) + c(y)\}}{\exp\{\lambda \cdot u(f(x)) + \kappa \cdot \phi(f(x)) + c(x)\}}, \quad (12)$$

which implies $c(x) = c(y)$, the contradiction. \square

2.3 Research hypotheses and benchmark models

FOCAL generalizes logit through capturing focality and by design, it thus appears more adequate than logit to analyze choice with presentation effects. For later reference, by logit, a DM with utility u chooses x with probability

$$\text{Logit: } \Pr(x|u, \phi, B) = \frac{\exp\{\lambda \cdot u(x)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u(x')\}}. \quad (13)$$

While the theoretical adequacy of FOCAL to capture effects relating to focality is immediate, it is unclear whether FOCAL improves on logit given the specific patterns in actual data. Further, it is unclear whether any such improvement implies that utility estimates are significantly more reliable given the comparably small size of typical (experimental) data sets. This yields the research hypotheses to be analyzed in Sections 4 and 5, respectively.

Hypothesis 1 (Model adequacy). *FOCAL captures choice with presentation effects significantly more accurately than logit.*

Hypothesis 2 (Reliability and consistency). *FOCAL's estimates are significantly more reliable across "similar tasks" than logit's estimates.*

These hypotheses are analyzed using data sets introduced in Section 3. The econometric analysis also allows me to clarify the relation to two other prominent models in the existing literature that intuitively apply to numerical choice and round-number effects.

Limited attention Round-number effects can be interpreted two ways: subjects either focus on some options or neglect other options. Masatlioglu et al. (2012) generalize revealed preference to account for DMs not considering all their options, Manzini and Mariotti (2014) generalize this idea to stochastic choice, and Echenique et al. (2014) generalize the model further by allowing for a weak "perception ordering": first all options at the highest perception level are considered, second the options at the next-highest level, and so on. This Perception Adjusted Luce Model (PALM) straightforwardly applies to focality effects, first the most focal options are considered, next the second layer, and so on, and hence it constitutes a natural benchmark for FOCAL. Formally, DM with utility u , focality ϕ , precision λ , and choice bias $\kappa \in [0, 1]$, chooses $x \in B$ with

$$\text{PALM: } \Pr(x|u, \phi, B) = \mu(x, X) \cdot \prod_{k > \phi(x)} \left(1 - \kappa \cdot \sum_{x' \in X: \phi(x')=k} \mu(x', X) \right) \quad (14)$$

where $\mu(x, X) = \text{Logit}(x) = \exp\{\lambda u(x)\} / \sum_{x' \in X} \exp\{\lambda u(x')\}$. The focality index ϕ used here will (of course) be equivalent to the one used for FOCAL (see Def. 3). While Echenique et al. define and axiomatize PALM only for $\kappa = 1$, I allow for the whole spectrum down to $\kappa = 0$ (which is logit). Further, I rescale the choice probabilities so they add up to 1, following Manzini and Mariotti's suggestion for cases without "outside options".

Similarity/Nested logit Choice violates IIA in the presence of “similarity” effects, and intuitively proximate numbers are more similar than distant numbers. Such similarity effects can be expressed by nested logit (McFadden, 1976) where DM first chooses a “nest” of options and secondly makes his final choice from this nest. Small (1987) introduces a cross-nested logit model (with overlapping nests) for choice from ordered sets, called *Ordered GEV*,⁶ which intuitively captures possible similarity effects in numerical choice. Here, DM first makes a tentative choice $y \in B$ and then reconsiders the neighborhood of y to make the final choice $x \in [y - w, y + w]$. To clarify the relevance of nesting and similarity effects, I include Ordered GEV as benchmark model. Formally, DM with utility u , precision λ , degree of correlation κ , bandwidth parameter $M < |X|$, and options represented by their integer ranks $s = 1, 2, \dots$, the choice probabilities are

$$\text{OGEV: } \Pr(s) = \sum_{r=s}^{s+M} \frac{w_{r-s} \exp\{\lambda u(s)/\kappa\}}{\exp\{I_r\}} \cdot \frac{\exp\{\kappa I_r\}}{\sum_{t=0}^{B+M} \exp\{\kappa I_t\}}$$

with $I_r = \ln \sum_{s' \in B_r} w_{r-s'} \exp\{\lambda u(s'|\alpha, \beta)/\kappa\}$. (15)

3 Testing FOCAL: Preliminary remarks

The data selected for the analysis of the research hypotheses are from studies on generalized dictator games.

Definition 5 (Generalized dictator game). DM chooses an option $x \in \{0, 1, \dots, B\}$. Given x , the dictator’s payoff is $\pi_1(x) = \tau_1 \cdot (B - x)$ and the recipient’s payoff is $\pi_2 = \tau_2 \cdot x$.

The data sets chosen satisfy the following requirements. First, there is consensus on the subjects’ true utilities u , there are unambiguous presentation effects, and the choice task does not involve risk or (strategic) uncertainty. This enables an unconfounded analysis of focality and utility. Second, the data sets are from controlled experiments, which gives us perfect information of the choice environment, and they are from experiments run to estimate utility, so we avoid re-analyzing experiments out-of-context. Third, the data sets are representative for literally hundreds of experimental analyses (Engel, 2011), and in all cases, dictator games are analyzed to understand altruism. Thus, analyzing reliability of utility estimates is also relevant in the wider context. Finally, the data sets contain multiple observations per subject, which allows us to disentangle preferences, precision, and choice bias (i.e. to estimate FOCAL) while allowing for subject heterogeneity, and there exist even four such experiments, examining essentially equivalent choice tasks varying only presentation. This allows us to study reliability of counterfactual predictions.

Table 1 provides an overview of the data sets and Figure 1 above provides selected histograms of observed choices. In conjunction, these studies provide a fairly comprehen-

⁶All cross-nested logit models are compatible with random utility if utility perturbations have a generalized extreme value distribution (GEV), hence the name Ordered GEV.

Table 1: The data sets

	#Treatments	#Options	#Observations	Transfer ratios
<i>“Numerical” dictator games</i>				
AM02 (Andreoni and Miller, 2002)	8	41–101	176 × 8	3 : 1, ..., 1 : 3
HJ06 (Harrison and Johnson, 2006)	10	41–101	57 × 10	1 : 1, ..., 1 : 4
CHST07 (Cappelen et al., 2007)	6	401–1601	96 × 2	1:1
<i>“Graphical” dictator games</i>				
FKM07 (Fisman et al., 2007)	50	500–1000	76 × 50	4 : 1, ..., 1 : 3

Table 2: Distribution of choices across “round” numbers

Treat	Percentage of choices with greatest factor ...												
	0	1000	500	250	100	50	25	10	5	2.5	1	0.5	other
AM02	39				1	9	7	33	6	0	4		
HJ06	22				3	4	7	39	14	0	10		
FKM07	25				0	0	0	0	2	1	4	5	63
CHST07	30	0	5	1	62	1	0	0	0	0	0		

Note: For each experiment, these percentages are pooled (and averaged) across treatments. The numbers do not always add up to exactly 100 due to rounding errors.

sive picture of dictator choice. Fisman et al. (2007, FKM07) use a graphical user interface which reliably prevents round-number effects (see Figure 1), while all other studies require subjects to enter numbers directly and thus reproduce the prevalent round-number patterns. Between those, Cappelen et al. (2007, CHST07) allow for budgets up to $B = 1600$, and subjects primarily choose multiples of 100, while Andreoni and Miller (2002, AM02) and Harrison and Johnson (2006, HJ06) allow for budgets up to $B = 100$ and choices mainly are multiples of 10. In AM02, the Leontief (payoff-equalizing) choice is generally a multiple of 10 or 25, but in HJ06, it is often a plain integer. The latter drastically affects the relative frequency of the Leontief choice, see Figure 1b: the Leontief choice is frequent if and only if it is a round number. Thus we may drastically under- or overstate precision and inference on utility estimates if we do not control for focality.⁷

Regarding utility, following Andreoni and Miller (2002) and Fisman et al. (2007), analyses of dictator games mostly use utility functions exhibiting constant elasticity of substitution (CES) between dictator income π_1 and recipient income π_2 , i.e.

$$u_i(\pi_i, \pi_j) = ((1 - \alpha) \cdot (1 + \pi_i)^\beta + \alpha \cdot (1 + \pi_j)^\beta)^{1/\beta}. \quad (16)$$

Here, α represents the degree of altruism and β represents the degree of efficiency con-

⁷The intuition is simple. Assume the Leontief choice is to transfer 10 out of 20. Without controlling for focality, we assume DM picked 10 over all alternatives, including say 5–9 and 11–15. If subjects seriously considered only 5 and 15 due to the low focality of plain integers, however, the scope for inference is much weaker. Theoretically, controlling for focality captures this effect, the practical side is tested below.

cerns. Subjects are efficiency concerned with $\beta = 1$ and equity concerned with $\beta \rightarrow -\infty$.

Finally, regarding the round-number effects, Table 2 provides a detailed overview of the numbers chosen. In the experiments with numerical choice, subjects rarely choose plain integers. Mostly, they choose multiples of 10 and 100, as described above. In addition, subjects tend to choose multiples of 5 and 50 reasonably regularly, while multiples of 2.5 and 25 are chosen a little less frequently. In this sense, the numbers chosen in numerical DG experiments do not differ substantially from those observed in survey responses, i.e. there is no indication that using the focality index reflecting the patterns in survey responses, introduced in Definition 3, would be inadequate. There may be room for improvement, but then, the adequacy of FOCAL would only be underestimated.

4 Assessing model adequacy

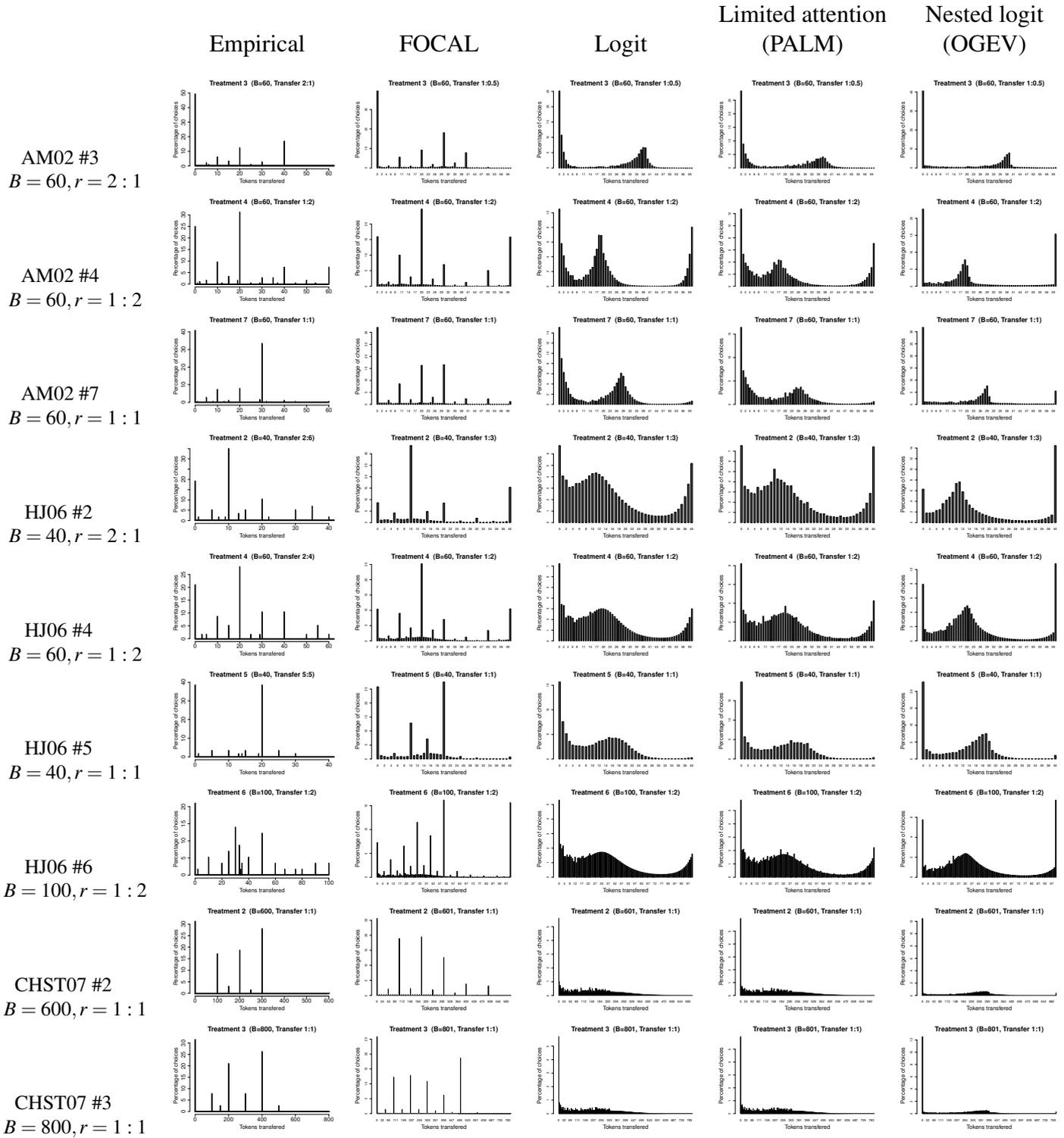
The present section analyzes how the models reproduce the choice patterns in-sample. This “descriptive adequacy” identifies strengths and weaknesses of the models, but potentially involves overfitting. Overfitting will be verified in the next section by analyzing reliability of estimates across experiments (“predictive adequacy”).

The basic picture Figure 3 provides plots of the model fits in-sample. The underlying econometric model is simple and transparent: utilities exhibit CES, Eq. (16), preferences are heterogeneous between subjects (random parameters), error variance and choice bias are identical across subjects, and parameters are estimated by maximum likelihood.⁸ This specification is leaned on the well-known “mixed effects” regression models; the main analysis will use a specification allowing for heterogeneous error variance and choice bias. The figure contains a superset of the treatments already presented in Figure 1, skipping FKM07 where all models fit similarly due to the absence of round-number effects.

The differences between models are largely in line with the intuition. To begin with, logit interprets differences in choice probabilities to indicate differences in utilities, and utilities are continuous by assumption. Thus logit does not comprehend the mass points at round numbers, yielding a “blurry” representation of behavior. The contrast to FOCAL is striking: by adding just one parameter to separate utility and focality, FOCAL reproduces the general shape of choice distributions. Parameters are fitted to all treatments per experiment, and thus there is some noise left in each treatment, but the overall accuracy is encouraging. Limited attention (PALM) and nested logit (OGEV) similarly generalize logit by adding one parameter, but without much of an effect. OGEV captures spikes at utility maximizers, most prominently zero transfers and Leontief transfers, but this misrepresents that spikes actually relate to round numbers. Limited attention, improves on

⁸Such mixed logit models are standard practice in analyses of consumer demand since Berry et al. (1995), and for ten years also in analyses of social preferences (Cappelen et al., 2007; Bellemare et al., 2008) and risk preferences (Harrison et al., 2007; Andersen et al., 2008).

Figure 3: The precision of the choice models in capturing numerical choice



Note: These plots depict the “predictions” of various models in the respective treatments after fitting the model parameters to all treatments from the respective experiments. The plots shown here represent a fairly representative selection, and the full list of plots is provided as supplementary material (as are the underlying parameter estimates). FKM07 is left out here, as Treatments as such are not defined (budget B s and transfer rates are individually randomized) and the differences between models are very minor in any case.

logit in the “right” direction, being based on the focality index ϕ (see Definition 3) capturing “roundness” of numbers, but the effect is hardly visible. Round-number effects are much stronger than limited attention admits, which will be discussed below. This shows that the focality index itself is not capturing behavior, suggesting that FOCAL’s apparent adequacy stems from its separation of utility and focality in choice propensities.

Econometric analysis Table 3 provides the information on the quantitative differences between the models. As indicated, all models are used exactly as defined in Section 2, equipped with CES utilities and estimated by maximum likelihood. All parameters $(\alpha, \beta, \lambda, \kappa)$ are allowed to be heterogeneous across subjects⁹ and the incomes in the CES utility function are the experimental tokens earned by the subjects.¹⁰ Likelihoods are computed by numerical integration using quasi-random numbers (Train, 2003) and maximized in a three-step approach, first using a gradient-free approach (NEWUOA, Powell, 2006), secondly using a Newton-Raphson method to ensure convergence, and finally using extensive cross-testing of estimates across data sets and models to ensure that global maxima are found. Appendix A provides the standard procedural details.

Table 3 reports both the Bayes information criterion (BIC) and the pseudo- R^2 . The former quantifies the model adequacy, the latter clarifies how much of the observed variance is explained. This relates the model to the benchmarks *clairvoyance*, i.e. correctly predicting the observed choice distributions, and *naivete*, i.e. uniform randomization.¹¹ In addition, relation signs indicate significance of differences, evaluated in likelihood ratio tests robust to both misspecified and arbitrarily nested models (Schennach and Wilhelm, 2014). I distinguish two levels of significance, namely the conventional level of 0.05, and the level of 0.005. The latter roughly implements the Bonferroni correction given the simultaneous tests of four models on four data sets. I will focus on the latter level, but the distinction does not affect the main results (as Table 3 indicates). Finally, I pool the results for AM02 and HJ06 due to their similarity, both entailing numerical choice from up to $B = 100$ with 8 or 10 observations per subject. This way, Table 6 presents the results by type of data set: numerical choice with many observations per subject (“large numerical”), with few observations per subject (“small numerical”), and graphical choice (“graphical”).

The results in Table 3 confirm Figure 3. In numerical DGs, the models can be partitioned into three equivalence classes, with logit and FOCAL as the extremes, and PALM and OGEV in-between, but improving only slightly on logit. In large numerical exper-

⁹ α is truncated normal on $[-0.5, 0.5]$, β is normal, λ and κ are log-normal.

¹⁰The supplement provides extensive robustness checks for both assumptions. Allowing for subject heterogeneity in all dimensions allows for slightly more robust fit for all models, without obstructing identification. The token-based utility function used here fits best assuming the status quo model logit. The main alternative to the latter assumption is contextual utility (Wilcox, 2011), which tends to fit best for FOCAL and PALM, but these tendencies are overall minor and therefore not discussed here. The main results are comparably strong and perfectly robust to changing these assumptions, as shown in the supplement.

¹¹ Given a model’s log-likelihood, the BIC is defined as $BIC(\mathbf{d}) = |ll(\mathbf{d})| + \log(\#obs) \cdot \#par/2$ (Schwarz, 1978) and given the log-likelihoods of the “clairvoyant” model and the naive model, denoted as ll_{\max} and ll_{\min} respectively, the pseudo- R^2 is defined as $R^2 = (BIC - ll_{\min}) / (ll_{\max} - ll_{\min})$ (Nagelkerke, 1991).

Table 3: Precision of the choice models in-sample (BIC: less is better; R^2 : more is better)

	Value range		Focality (FOCAL)	Similarity (OGEV)	Limited Attention (PALM)	IIA (Logit)
	Clairvoyance	Naivete				
<i>Large numerical DG experiments (AM02, HJ06; 8 and 10 observations per subject)</i>						
BIC	2812.1	8137	3371.9	\ll	4420.9	\approx 4492.6 \ll 4690.3
R^2	1	0	0.895	\gg	0.698	\approx 0.684 \gg 0.647
<i>Small numerical DG experiments (CHST07; 2 observations per subject)</i>						
BIC	260.8	1271.4	430.4	\ll	920	\approx 910.5 \approx 931.1
R^2	1	0	0.832	\gg	0.348	\approx 0.357 \approx 0.337
<i>Graphical DG experiment</i>						
BIC	10021.8	23249.2	15103.9	\approx	15087.7	\approx 15119.4 \approx 15123.9
R^2	1	0	0.616	\approx	0.617	\approx 0.615 \approx 0.614

Content: For each choice model (Logit, OGEV, PALM, and FOCAL), the descriptive accuracy (BIC and pseudo- R^2 in-sample) is reported for data from the three groups of experiments. Significance of differences is (following Schennach and Wilhelm, 2014) is denoted as follows: \approx indicates p -values above 0.05, $>$, $<$ indicate p -values between 0.005 and 0.05, and \gg , \ll indicate p -values below 0.005.

iments, FOCAL's adequacy (pseudo- R^2) is 89%, logit is around 65%, and PALM and OGEV are around 69%. All differences are statistically significant and easily visible (see Figure 3). In the small numerical experiment, FOCAL's adequacy is 83% and the other models are around 35%. In the graphical experiment, all models are around 62% adequacy. Aside from the high statistical significance, two observations stand out. First, FOCAL attains a substantially higher pseudo- R^2 in numerical DGs than in graphical ones. This is intuitive if a model captures round-number effects, as choices cover relatively few options in numerical DGs and thus indeed are more predictable than in graphical DGs. The other models are equally adequate in the (large) experiments with and without round-number effects, confirming the optical impression that they do not comprehend round-number effects. Second, FOCAL's adequacy is largely similar in the large and small numerical experiments, being 89% and 83%, respectively. These experiments differ in the relative strength of round-number effects: CHST07 implements the largest number of options, up to $B = 1600$, but the smallest number of different options actually gets chosen by the subjects (see also the discussion of entropy following shortly). In this sense, the round-number effects are strongest in CHST07, and the observation that FOCAL's adequacy is similar indicates that it is robustly adequate. In turn, the adequacy of the other models drops substantially in CHST07, to around 35%: these models do not capture round-number effects, and thus, the stronger the round-number effects, the lower their adequacy.

Result 1 (Basic adequacy). *FOCAL captures numerical choice effectively, explaining 88% of the observed variance. All models capture graphical choice equally well.*

Since OGEV fits only slightly better than logit, similarity effects appear to be of minor relevance in numerical choice. As for limited attention (PALM), the observation that it barely improves upon logit in the present context appears to relate to its axiom “Hazard Rate IIA” (Echenique et al., 2014), which is an implication of “I-Asymmetry” and “I-Independence” in Manzini and Mariotti (2014). To illustrate its impact, consider a dictator having to choose from $\{1, 2, \dots, 100\}$, assuming multiples of 10 have high focality (perception) and all other options have low focality. Let us say the 10% of the options with high focality attract 20% of the choices under logit. Then, Hazard Rate IIA implies that choice probabilities of the low-perception options are discounted by at most these 20%. Round-number effects are much stronger, discounting choice probabilities of low-focality options by close to 100%. Thus, Hazard Rate IIA limits PALM in capturing numerical choice, and weakening it may improve its adequacy substantially.

Explaining entropy The most palpable stylized fact related to round-number effects is the relatively low number of different options being chosen. Statistically, this is captured by the entropy of choices. Using $\Pr(x)$ as the relative frequency of x , the Shannon-entropy $H = -\sum_{x:\Pr(x)\neq 0} \Pr(x) \log(\Pr(x))$ measures the information contained in a set of observations, and $\exp(H)$ quantifies the number of different options being chosen “systematically”. To provide intuition, $\exp(H)$ is exactly 1 if all subjects choose the same option, it is equal to the number of options if the observations are distributed uniformly, and in the analyzed experiments, $\exp(H)$ is approximately equal to the number of different options that are (minimally) required to cover 90% of all choices. The estimates of $\exp(H)$ are around 5–10 in the numerical DG experiments AM02, HJ06, CHST07, and around 30 in the graphical DG of FKM07.¹² Thus, subjects consistently focus on 5–10 options in numerical DGs, although the total number of options ranges from 41 to 1601.

This observation is intuitively linked to both, limited attention and focal choice. Limited attention presumes that only a subset of options is considered, and focal choice presumes that a (limited) number of options have elevated focality. To evaluate whether this stylized fact is explained by the models, I use the estimates obtained above and compute the predicted entropy in the various treatments. The significance of differences between predicted and observed entropies is evaluated in Wilcoxon matched pairs tests. The results are reported in Table 4 and confirm that both limited attention (PALM) and Ordered GEV slightly improve on logit in capturing the observed choice patterns, but are far from being compatible with the extent of the round-number effects. In turn, FOCAL is compatible with it in the sense that the predicted entropy, while being slightly too large,¹³ is not significantly different from the observed entropy.

Result 2 (Entropy). *FOCAL consistently captures entropy in numerical choice.*

¹²Detailed overviews of these statistics are provided in the supplementary material. To compute these numbers for FKM07 (where budget sets are random between subjects), choices x_i are transformed to shares transferred, i.e. $x_i / \max x$, for each decision and rounded to multiples of 0.01.

¹³Intuitively, this relates to the noise in the data, which is unpredictable ex-ante but manifests as specific choices ex-post, and to the simplicity of the assumed focality index.

Table 4: Is entropy reliably captured?

	Empirical	FOCAL		OGEV		PALM		Logit
All numerical DGs	1.99	2.48*	≪	3.58**	≈	3.61**	≈	3.63**
AM02 + HJ06	2.1	2.53	<	3.27**	≈	3.29**	≈	3.32**
CHST07	1.65	2.34	<	4.51**	≈	4.57**	≈	4.58**

Content: This table relates the empirical entropy (estimated entropy averaged across all treatments in the respective experiments) to the respective predictions of the four choice models. The results are either pooled across all numerical DG experiments or reported for the large/small DG experiments (AM02 + HJ06 or CHST07, resp.). Differences are evaluated by Wilcoxon tests, the notation of relation signs is as in Table 3. The “stars” indicate that the model’s prediction differ significantly from the empirical estimate; significance at 0.005 is indicated by ** and significance at 0.05 is indicated by *.

5 Reliability and consistency

The basic hypothesis formulated above was that if the choice model does not capture the presentation effects, preference estimates are biased and inconsistent across studies. For, presentation affects focality, which in turn affects choice propensities—and if focality is not controlled for, its effect on choice propensities will bias utility estimates. The predicted bias can be econometrically tested, as it diminishes the reliability of predictions and the consistency of estimates across experiments, which is the scope of the present section.

The evaluation of predictive adequacy is the arguably most widely accepted approach to model validation, being an integral part of the scientific method, and has a tradition also in behavioral economics. Most notably, risk preferences tend to be evaluated by predictive adequacy, e.g. Harless and Camerer (1994), Wilcox (2008), Dave et al. (2010), and Hey et al. (2010), but similarly so models of learning (Camerer and Ho, 1999), time preferences (Keller and Strazzera, 2002), and strategic reasoning (Camerer et al., 2004). In the context of social preferences, however, the only study attempting an assessment of predictive adequacy appears to be Breitmoser (2013). This is somewhat surprising, as the instability of estimates of social preferences is notorious, but predicting say dictator behavior requires prediction of numerical choice. Thus, any result on differences across experiments would have to be ascribed to say round-number effects if those are not adequately captured, rendering analyses of predictive adequacy in dictator games uninformative. This implies, however, that we cannot assess the reliability of preference measurement so far.¹⁴

To give an idea of reliability, Table 5 provides the estimated mean degrees of altruism and efficiency concerns.¹⁵ The mean degree of efficiency concerns is similar across models and data sets, being close to zero (i.e. Cobb-Douglas). The mean degree of altruism varies substantially across data sets. The estimates of logit, PALM and OGEV range from roughly -0.1 to 0.5 , which is volatile given the general bounds of -0.5 and 0.5 . The ex-

¹⁴In addition, behavior is qualitatively robust across dictator game experiments (Camerer, 2003), fueling the prior that preferences are constant across experiments and reducing the need for statistical analysis of their actual stability. Such analyses are required to assess the instability of preferences across games, though.

¹⁵The full lists of parameter estimates and standard errors are provided in the supplementary material.

Table 5: Mean degrees of altruism α and efficiency concerns β

	FOCAL		OGEV		PALM		Logit	
	α	β	α	β	α	β	α	β
AM02	0.08	0.17	-0.13	-0.16	-0.27	0.25	-0.07	-1.69
HJ06	0.23	-0.67	0.13	-0.6	0.17	-0.65	0.17	-0.65
FKM07	0.02	0.26	0.15	-0.17	0.01	0.24	0.01	0.24
CHST07	0.16	-0.14	0.5	-0.62	0.5	-0.62	0.5	-0.62

Content: This table reports the estimated means of the degree of altruism (α) and the degree of equity concerns (β). The estimates are given for each choice model (Logit, OGEV, PALM, FOCAL) and each data set. The estimated standard errors are mostly rather close to zero and skipped for readability of this table. They are reported in the tables toward the end of the supplementary material.

treme estimates result in the experiments with the most pronounced round-number effects (i.e. with the lowest entropy), -0.1 in AM02 and 0.5 in CHST07. In contrast, FOCAL’s estimates are fairly robust, ranging from 0.02 to 0.23 .¹⁶ The main question is now: Do these estimates differ across data sets? Standard errors are skipped in Table 5 and relegated to the supplement, as they are not informative to answer the question. An answer requires the joint evaluation of four preference parameters and it requires us to control for both precision λ and choice bias κ . These concerns are addressed in likelihood ratio tests, and as before, I use the robust test of Schennach and Wilhelm (2014).

5.1 Counterfactual reliability of estimates

Assume we have a data set examining behavior in one set of conditions and wish to predict behavior in related conditions. The reliability of such counterfactual predictions is critical for policy recommendations about say tax rates and choice interfaces (nudging). In applications we might have more or less information about the target environment, i.e. about precision λ and bias κ of DMs in the counterfactual scenario. To reflect this possibility, and to check robustness in this respect, I distinguish counterfactual reliability to three degrees. The first degree is the reliability of the preferences estimates as such (distributions of α and β), while precision and choice bias (distributions of λ and κ) are assumed to be known for the target environment. This yields an understanding of the reliability of preference estimates in isolation. The second degree requires that precision is to be predicted as well, and only choice bias is assumed to be known, and the third degree requires out-of-sample prediction also of the choice bias (i.e. of all parameters). For clarity, let me formalize this.

Definition 6 (Counterfactual reliability). Given data sets D_{in} and D_{out} , let $(\alpha_{in}, \beta_{in}, \lambda_{in}, \kappa_{in})$ and $(\alpha_{out}, \beta_{out}, \lambda_{out}, \kappa_{out})$ denote the respective ML estimates (e.g. α_{in} is the duple of mean and variance of α in-sample). Given a parameter vector, let $BIC(\alpha, \beta, \lambda, \kappa | D)$ denote its

¹⁶Note that this is the case for the assumption that utility is based on the tokens earned in the experiment, which favors logit. FOCAL suggests that contextual utility (Wilcox, 2011) is more adequate, and then FOCAL’s estimates are virtually identical in all data sets (see the supplementary material).

Table 6: Counterfactual reliability of estimates: How reliable are predictions of behavior in numerical and graphical experiments, based on estimates from the other experiments?

	Value range		Focality (FOCAL)	Similarity (OGEV)	Limited Attention (PALM)	IIA (Logit)			
	Clairvoyance	Naivete							
First degree (CR_1)									
Large numerical	2812.1	8137	3632.4	\ll	5359.1	\ll	5548.6	\ll	5779.6
Small numerical	260.8	1271.4	464.7	\ll	1042.8	\approx	1075.2	\approx	1090.9
Graphical	10021.8	23249.2	15491.3	\ll	17217.5	\approx	17785.8	\approx	17633.5
Second degree (CR_2)									
Large numerical	2812.1	8137	3609.1	\ll	5416	\ll	5551.4	\ll	5777
Small numerical	260.8	1271.4	476.8	\ll	1064	\approx	1087.1	\approx	1089.5
Graphical	10021.8	23249.2	15581	\ll	17562.7	\approx	17964.1	\approx	17812.7
Third degree (CR_3)									
Large numerical	2812.1	8137	3746.8	\ll	5525.9	\ll	5809.2	\ll	6041.4
Small numerical	260.8	1271.4	506.9	\ll	1064.3	\approx	1097.9	\approx	1085.4

Content: This table evaluates the accuracy of predicting behavior in either “large numerical” experiments ($D_{out} = \text{AM02, HJ06}$), “small numerical” experiment ($D_{out} = \text{CHST07}$) or “graphical” experiment ($D_{out} = \text{FKM07}$), using estimates from the respective other studies. As before, \approx indicates p -values above 0.05, $>$, $<$ indicate p -values between 0.005 and 0.05, and \gg , \ll indicate p -values below 0.005.

Bayes Information Criterion evaluated on data set D , using the correct number of parameters fitted to data set D . Then, the three degrees of counterfactual reliability are

$$1. \text{ Preferences: } CR_1(D_{in}, D_{out}) = BIC(\alpha_{in}, \beta_{in}, \lambda_{out}, \kappa_{out} | D_{out}) \quad (17)$$

$$2. \text{ Pref \& Prec: } CR_2(D_{in}, D_{out}) = BIC(\alpha_{in}, \beta_{in}, \lambda_{in}, \kappa_{out} | D_{out}) \quad (18)$$

$$3. \text{ All parameters: } CR_3(D_{in}, D_{out}) = BIC(\alpha_{in}, \beta_{in}, \lambda_{in}, \kappa_{in} | D_{out}) \quad (19)$$

To determine counterfactual reliability to the first and second degree, I analyze predictions across numerical and graphical experiments. Counterfactual reliability to the third degree entails prediction of choice bias κ , i.e. predicting the extent of round-number effects. It cannot be evaluated between numerical and graphical experiments, as the latter had been explicitly designed to neutralize the choice bias. Hence, I focus on the numerical experiments in the evaluation of counterfactual reliability to the third degree. For convenience, the results are aggregated to present the average accuracy of predicting a given experiment using estimates from the other experiments. Formally, given degree k , the reliability in predicting D_{out} is $CR_k(D_{out}) = \sum_{D_{in} \neq D_{out}} CR_k(D_{in}, D_{out})/3$. The classes of experiments are labeled large numerical, small numerical, and graphical, as above.

Table 6 presents the results. The models roughly sort into the same equivalence classes as in-sample: FOCAL and logit are at the extremes, PALM and OGEV are in-between, but improving only slightly on logit. This pattern robustly holds for all three degrees of counterfactual reliability. Even the information criteria and the underlying pseudo- R^2 are similar to before, being close to 0.85 for FOCAL in environments with nu-

merical choice, and mostly between 0.4 and 0.5 for the other models in case of numerical choice. In addition to the high statistical significance, two observations specifically suggest that FOCAL captures choice and thus measures preferences reliably. On the one hand, the models are differently adequate in predicting graphical choice. Predicting graphical choice represents the arguably cleanest test of counterfactual reliability: graphical choice does not exhibit round-number effects, all models are equally adequate in-sample, and thus all differences in reliability stem from inaccurate measurement of preferences in the other data sets. The higher adequacy of FOCAL in predicting graphical choice thus shows that its preference estimates are more reliable.

On the other hand, the reliability of counterfactual predictions depends only marginally on our knowledge of the target environment. In particular, the precision of subjects is predicted reliably, i.e. it represents a robust facet of behavior, and accuracy drops only slightly (between two and four percentage points in the pseudo- R^2) if we have to predict even the extent of the choice bias. In applications, neither precision nor choice bias therefore need to be known for the target environment. This is largely trivial for all models but FOCAL, as those do not capture the choice bias. As for FOCAL, however, this shows that its decomposition of behavior into utility and focality provides a measurement of choice factors that is both accurate and robust across choice environments.

Result 3. *Counterfactual predictions are most reliable for FOCAL, for all user interfaces and regardless of how much we know about the target environment.*

5.2 External consistency of estimates

Now assume we have two data sets and wish to examine if preferences in one data set differ from those in the other data set. I refer to this as test of the consistency of preferences. Consistency of estimates across experiments on a given game is a necessary condition for reliably understanding differences between games. If estimates are inconsistent, e.g. if preferences are found to significantly depend on presentation, results on differences of estimates between classes of games would not be informative—they could either reflect actual changes in utility or pure presentation biases.

To evaluate estimate consistency, the relevant likelihood ratio is the difference of log-likelihoods in-sample and out-of-sample. That is, we take estimates from a data set D_{in} , predict D_{out} , and compare the respective likelihood (or, BIC) to the one achieved in-sample in D_{out} . As above, I distinguish consistency to the three degrees: consistency of preference estimates in isolation, consistency of preference and precision estimates jointly, and consistency of all estimates jointly.

Definition 7 (Estimate consistency). Given data sets D_{in} and D_{out} , let $(\alpha_{in}, \beta_{in}, \lambda_{in}, \kappa_{in})$ and $(\alpha_{out}, \beta_{out}, \lambda_{out}, \kappa_{out})$ denote the respective ML estimates (e.g. α_{in} is the duple of mean and variance of α in-sample). Given a parameter vector, let $BIC(\alpha, \beta, \lambda, \kappa|D)$ denote its Bayes Information Criterion evaluated on data set D , using the correct number of parameters

Table 7: Estimate consistency: How different are estimates between experiments?

	Value range		Focality (FOCAL)	Similarity (OGEV)	Limited Attention (PALM)	IIA (Logit)			
	Clairvoyance	Naivete							
First degree (EC_1)									
Large numerical	0	22737.7	531.4	\ll	1601.6*	<	2188.6**	>	1655*
Small numerical	0	12910.4	1377.9*	\ll	7111**	\ll	8738.5**	\approx	8890.6**
Graphical	0	5606	137.4	\ll	859.6**	>	733.7**	\approx	730.4**
Second degree (EC_2)									
Large numerical	0	22737.7	854.9	\ll	1825.5**	<	2327.2**	>	1755.7**
Small numerical	0	12910.4	1275.7*	\ll	8108.3**	\ll	9252.4**	\approx	9359.2**
Graphical	0	5606	151.4	\ll	908.3**	\gg	660.4**	\approx	686.7**
Third degree (EC_3)									
Large numerical	0	6447	235.5*	\ll	806.3**	\approx	746**	\approx	703.3**
Small numerical	0	4765.1	667**	\ll	1692.2**	\ll	2262**	\approx	2307.5**

Content: This table evaluates consistency of estimates from either “large numerical” experiments ($D_{in} = \text{AM02, HJ06}$), “small numerical” experiment ($D_{in} = \text{CHST07}$) or “graphical” experiment ($D_{in} = \text{FKM07}$), in relation to estimates from the respective other studies. As before, \approx indicates p -values above 0.05, $>$, $<$ indicate p -values between 0.005 and 0.05, and \gg , \ll indicate p -values below 0.005. Further, “stars” indicate the significance of inconsistency; significance at 0.005 is indicated by ** and significance at 0.05 by *.

Note: The third degree does not involve predictions of the graphical experiment (FKM07). Thus, the numbers are not directly comparable to the other degrees.

fitted to data set D . Then, the three degrees of estimate consistency are

$$EC_1(D_{in}, D_{out}) = BIC(\alpha_{in}, \beta_{in}, \lambda_{out}, \kappa_{out} | D_{out}) - BIC(\alpha_{out}, \beta_{out}, \lambda_{out}, \kappa_{out} | D_{out}) \quad (20)$$

$$EC_2(D_{in}, D_{out}) = BIC(\alpha_{in}, \beta_{in}, \lambda_{in}, \kappa_{out} | D_{out}) - BIC(\alpha_{out}, \beta_{out}, \lambda_{out}, \kappa_{out} | D_{out}) \quad (21)$$

$$EC_3(D_{in}, D_{out}) = BIC(\alpha_{in}, \beta_{in}, \lambda_{in}, \kappa_{in} | D_{out}) - BIC(\alpha_{out}, \beta_{out}, \lambda_{out}, \kappa_{out} | D_{out}) \quad (22)$$

The results are presented in Table 7, aggregated again based on user interface. Now, given degree k , we evaluate the consistency of the estimates from D_{in} , which is defined as $EC_k(D_{in}) = \sum_{D_{out} \neq D_{in}} EC_k(D_{in}, D_{out})$. That is, for a given vector of estimates, we analyze how it explains behavior in other experiments. This is complementary to the analysis of counterfactual reliability, which is the reliability of predicting a given data set using estimates from other data sets. The inversion feels natural in evaluating consistency and provides a complementary perspective on the results, facilitating their interpretation.

In particular, this regards the results on CHST07 (“small numerical” DG experiment in Table 7). The earlier analysis of counterfactual reliability shows that estimates from other experiments allow to predict behavior in CHST07 fairly accurately, close to achieving in-sample accuracy—but Table 7 shows that the estimates from CHST07 are significantly inconsistent in predicting other data. The significance of inconsistency is indicated by “stars” in Table 7; “one star” indicates weak significance (0.05) and “two stars” indicate significance robust to the Bonferroni correction (0.005). FOCAL’s estimates from

CHST07 are weakly inconsistent in this sense with respect to consistency to the first and second degree, strongly significant with respect to consistency in the third degree, and the CHST07-estimates of the other models are strongly inconsistent in all cases. Overall, that is, CHST07 can be predicted well but it is not predictive. This suggests that the subjects do not act differently in CHST07 than in the other experiments, but the two observations per subject in CHST07 do not suffice to reliably identify preferences, precision, and focality. Notably, FOCAL is least vulnerable in this respect.

Now looking at the consistency across all data sets, the overall pattern is similar: FOCAL's estimates are mostly consistent (in 7 out of 8 cases), while the estimates of logit, PALM and OGEV are consistent in at most 1 out of 8 cases each. This is notable, as consistency measures how much lower out-of-sample accuracy is than in-sample accuracy, which is in principle independent of whether in-sample accuracy is high. That is, intuitively a model that does not fit exceptionally well may still be particularly robust due to being "simpler" (Hey et al., 2010), e.g. by making less or fewer inadequate assumptions or by being less flexible and thus avoiding overfitting. By the above results, FOCAL makes the least or fewest inadequate assumptions and fits well without being overly flexible.

This may suggest that FOCAL, being consistent in 7 out of 8 cases, is "more consistent" than the other models, which are consistent in at most 1 of the 8 cases. Such a claim is potentially misleading, as FOCAL may be just above the threshold and the other models may be just below it. To clarify the comparative properties of the models, Table 7 also provides information on their relative consistency. This is depicted by the relation signs. As Table 7 shows, FOCAL's estimates are highly significantly "more consistent" than the other estimates, while there is no robust ranking between logit, PALM and OGEV with respect to their estimate consistency. Between these three models, every one of them is most consistent in one context and least consistent in another context. Estimates from large numerical experiments are most consistent for logit and OGEV, estimates from the small numerical experiment (CHST07) are most consistent for OGEV, and estimates from graphical experiment (FKM07) are most consistent for PALM and logit.

The volatile (and low) consistency of these models again indicates that their structural assumptions are inadequate in capturing numerical choice, while the high consistency of FOCAL across contexts lends further support to its hypothesized adequacy.

Result 4. *Parameter estimates are consistent across experiments (only) with FOCAL.*

The result relates to the discussion of out-of-sample reliability of structural models. Structural models are considered reliable if the structural assumptions are adequate and the model is identified (Keane, 2010; Rust, 2010). Both adequacy and identification are often questioned. The above results show that such concerns are not generally unfounded: the two observations per subject in CHST07 do not suffice to reliably identify models of (numerical) choice, and the estimates of the models other than FOCAL, which are visibly inadequate (Figure 3), are significantly inconsistent. The results also show, however, that reliable identification is attained in the data sets with at least eight observations per subject, assuming the choice model is adequate in capturing the choice patterns.

6 Discussion

Individual choice exhibits presentation effects due to default settings, positioning, coloring and labeling of options. Intuitively, these facets of presentation affect the focality of options, which is a choice-relevant attribute alongside utility. The implied systematic deviations from utility maximization (“systematic mistakes”) have been assumed away in analyses of choice—transparently in models of rational choice and nontransparently in models of stochastic choice. The axiomatic foundation of the best-known model of stochastic choice, multinomial logit, assumes binomial logit by axiom, which in turn rules out systematic mistakes. Binomial logit is founded in axioms on narrow bracketing and absence of systematic mistakes, and dropping the latter yields a generalized logit model where choice depends on both utility and focality. This confirms the intuition that focality affects choice and allows us to capture systematic mistakes due to presentation effects.

Based on this observation of Breitmoser (2016), the present paper derives a “focal choice adjusted logit” model (FOCAL) allowing for both utility and focality as choice-relevant attributes and applies it to re-analyze dictator game data with unambiguous presentation effects. FOCAL is founded in axioms reflecting standard practice and provides a single-parameter generalization of logit. It is tractable using standard methods and allows us to study focality exactly the way utility has been studied for decades. The econometric analysis shows that utility estimates in the presence of round-number effects strikingly depend on the choice model used, and that utility estimates are inconsistent and unreliable if the model does not capture the round-number effects. In contrast, FOCAL effectively captures the choice patterns, its estimates are consistent across experiments, and its out-of-sample predictions are indeed reliable—it works as predicted, confirming the axiomatic approach taken in choice modeling and the axioms reflecting standard practice.

The impact of controlling for focality is striking, as already Figures 1 and 2 in the Introduction suggested. To put the results into context, Table 8 reviews Bayes Information Criteria and pseudo- R^2 for logit and FOCAL, as a function of the extent of subject heterogeneity being controlled for.¹⁷ It shows that the logit model assuming homogeneity of subjects, estimating preferences of a “representative subject”, actually has negative reliability ($R^2 = -0.091$). Given logit, controlling for subject heterogeneity allows to explain 40.1% of observed variation when using three additional parameters. Controlling for focality instead of heterogeneity yields the same reliability by using just one parameter in addition to logit, explaining 56.4% of observed variation. Thus, controlling for focality is at least as important for preference reliability as controlling for heterogeneity.

There are some implications that may be worth noting. The fact that FOCAL’s estimates are most adequate in-sample, most reliable out-of-sample, and most consistent across samples jointly suggest that the theoretically predicted (and intuitive) distinction of

¹⁷These BICs and R^2 measure counterfactual reliability of preference estimates as defined above: preferences are predicted based on estimates from other studies, while precision and choice bias are assumed to be known for the sample in question. This focuses on the reliability of the preference estimates in isolation. The other degrees of reliability yield similar results, as shown in the supplement.

Table 8: Incremental value of focal in counterfactual reliability of preference estimates

	Clairvoyance	Naivete	Homogenous		Het Prefs		Het Pref & Prec		Full Het
Logit	3558.9	9408.4	9940.9	>>	6546.9	<<	6879.4	≈	6870.5
FOCAL	3558.9	9408.4	6108.8	>>	4733.2	>>	4096.5	≈	4097.1
Logit			-0.091	<<	0.452	>>	0.399	≈	0.401
FOCAL			0.564	<<	0.738	<<	0.838	≈	0.838

Note: This table provides BIC and pseudo- R^2 for overall 4×2 models: four models of subject heterogeneity (starting with the “representative agent” model, i.e. homogeneity) and two models of choice (the status quo logit, and the generalization FOCAL). The underlying likelihood ratio tests follow Schennach and Wilhelm (2014), $>$, $<$ indicate p -values between 0.005 and 0.05, and $>>$, $<<$ indicate p -values below 0.005.

utility and focality is adequate. In turn, other models, and in particular simpler models such as logit, are significantly less robust. This is plausible, as the round-number effects are striking, but it contradicts the widespread intuition that simplicity yields robustness. In the present context, simplicity (i.e. logit) systematically produces type-1 errors in analyses of treatment effects on preferences. Thus, simplicity may yield seemingly “puzzling” results and prevent the emergence of a consensus on modeling social preferences.

Regarding simplicity and robustness, it seems intuitive that least squares methods require fewer assumptions and thus are more robust than say logit or generalizations thereof. This suggests that a generalization of least squares allowing for focality may be desirable. In principle, such generalizations are conceivable, e.g. by weighting deviations from the utility maximizer by focality, but the desirability of such generalizations is debatable. Sceptics of structural analyses consider the assumption that choice probabilities exactly take a (generalized) logit form questionable and prefer methods avoiding them, usually least squares. Least squares intuitively adheres to positivity, IIA, narrow bracketing, and absence of systematic mistakes. If one is willing to make these assumptions, however, then the formal implication is that choice probabilities are logit. Least squares requires additional assumptions, e.g. on the irrelevance of utility differences. That is, least squares requires unnecessary and untested assumptions, not logit, which suggests that a generalized least squares approach accounting for focality is unattractive.

Choice depends on preferences, precision, and “biases” due to e.g. focality, which need to be disentangled to estimate preferences. This requires multiple observations per subject, i.e. allowing for heterogeneity, FOCAL is in general not identified if there is only one observation per subject. Identifying restrictions assuming away say imprecision ($\lambda \rightarrow \infty$), efficiency concerns ($\beta = 1$), and bias ($\kappa = 0$) enable identification in scarce data sets, but these assumptions clearly are inadequate and the correspondingly identified degrees of altruism are unreliable, as logit’s inconsistency shows. Hence, if such identifying restrictions are made, the estimates cannot be used to reliably test for differences in preferences across treatments—even if the alleged effect is intuitive.¹⁸

¹⁸Examples of studies challenging social preference theory, with highly intuitive results obtained however under unreliable identifying assumption such as rationality ($\lambda \rightarrow \infty$) and zero focality ($\kappa = 0$) include the analyses of dictator games with taking options (List, 2007; Bardsley, 2008) and dictator games with sorting

The observation that multiple observations per subject are required to disentangle preferences, precision, and choice bias is not a drawback of using FOCAL, but an implication of the way subjects choose. This suggests to adopt the graphical user interface of Fisman et al. (2007). It prevents round-number effects, and thus the necessity to control for focality, while the estimates are reliable in the sense that they do not differ from the estimates of numerical DG experiments after controlling for focality (shown above). By suppressing focality effects, the graphical user interface simplifies estimation and identification, e.g. through allowing to use logit and by reducing the number of observations required. Obviously, implementing a graphical user interface similar to Fisman et al.'s is not convenient in all experiments, but it seems to be preferable when it is convenient.

Binomial and multinomial choice exhibits ordering and positioning effects (Dean, 1980; Miller and Krosnick, 1998; Feenberg et al., 2015). The first or left-most option tends to be favored by subjects, and naively analyzed, its utility is overestimated. Experimental analyses randomly reorder options across tasks to eliminate this bias. With FOCAL as model of choice, it is straightforward to check if the bias is indeed eliminated. To this end, consider the case of binomial choice and assume the left-most option has a focality bonus. Random reordering implies that in 50% of choices one option has the focality bonus, and in the other 50% the other option has it. As the supplementary material shows in detail, reordering eliminates the directional bias, but the option with the lower utility benefits more from being focal. This implies that the utility difference between the two options is underestimated, i.e. the estimated utility function is too flat and say risk aversion is overestimated. Thus, random reordering of options helps by “reducing” the bias due to focality, but it does not eliminate it.

To conclude, individual choice is systematic and predictable even in environments with many options and obvious choice biases. The choice model derived here, FOCAL, fits robustly across experiments, which I attribute to the basic idea of treating focality as choice-relevant attribute of options besides utility and the comparably strong behavioral foundation. The relative weights of utility and focality are not identified independently of the experimental design, i.e. multiple observations per subject are required if we allow for heterogeneity, and the analysis requires a “structural” model of choice. The former is an implication of the complexity of human choice and regarding the latter, FOCAL relies solely on assumptions made even in least squares analyses and has been shown to fit robustly, leaving less room than usual for scepticism. In turn, there is a tremendous upside: by using structural analyses to disentangle utility and focality, jointly with an experimental design ensuring identification, we may reliably estimate and examine preferences in non-binomial choice environments. Such reliability is required to make progress and potentially reach a consensus in (social) preference theory—similarly to the progress that has been made in understanding binomial choice and choice under risk. The method proposed here is equally applicable to a number of important problems besides preference

(Dana et al., 2006). There, stochastic choice and round-number effects clearly explain parts of the observations, and without disentangling preferences, precision, and choice bias, it is unclear whether preferences differ significantly from those in standard dictator games.

estimation. This includes analyses of numerical choice in strategic interactions, to reliably estimate depth of reasoning, and analyses of default or ordering effects, to facilitate policy recommendations in behavioral welfare economics. Analyses of focality thus offer exciting prospects in choice analysis and many opportunities for future work.

A Specification of the econometric model

Subjects' utilities exhibit constant elasticity of substitution (CES) in the incomes (π_i, π_j) ,

$$u_i(\pi_i, \pi_j) = ((1 - \alpha) \cdot (1 + \pi_i)^\beta + \alpha \cdot (1 + \pi_j)^\beta)^{1/\beta}, \quad (23)$$

where α represents the degree of altruism and β the degree of efficiency concerns. The standard assumption is that incomes π_i, π_j represent incomes measured in experimental tokens. Alternatively, incomes may represent the monetary values underlying the experimental tokens or utilities may be normalized to the range $[0, 1]$ as in $(u - u_{\min}) / (u_{\max} - u_{\min})$. The latter has been proposed by Wilcox (2011), Padoa-Schioppa (2009), and Padoa-Schioppa and Rustichini (2014) based on neuro-economic evidence that stimuli adapt to the environment. Thus, the normalization may improve fit across treatments and across experiments. The supplementary material analyzes which of these approaches captures utility best. The differences in-sample are insignificant, and out-of-sample, token-based utilities tend to fit best for logit and OGEV, while contextual utility tends to fit best for PALM and FOCAL. The differences are fairly minor overall, but for clarity, the supplementary material reports robustness checks on all results for all three approaches to utility definition. The paper reports the results for the approach favoring the status quo (logit), i.e. token-based utility.

Each subject is characterized by precision λ , altruism α , efficiency concerns β , and choice bias κ . Subjects may be heterogeneous, i.e. all parameters may be distributed randomly across subjects. Using $\mathbf{p} = (\lambda, \kappa, \alpha, \beta)$ to describe the parameter profile of a given subject and $f(\cdot|\mathbf{d})$ to describe its joint density in the population given distribution parameters \mathbf{d} , the likelihood that the model \mathbf{d} describes the choices o_s of subject $s \in S$ is

$$l(\mathbf{d}|o_s) = \int_{\mathbf{p}} f(\mathbf{p}|\mathbf{d}) \cdot \Pr(o_s|\mathbf{p}) d\mathbf{p}, \quad (24)$$

with $\Pr(o_s|\mathbf{p})$ as the probability that o_s results under parameter profile \mathbf{p} given the utility standardization and choice model being analyzed. The integral in Eq. (24) is evaluated by simulation, using quasi random numbers following standard practice (Train, 2003). The underlying distributional assumptions are as follows: altruism α is normal truncated to the interval $[-0.5, 0.5]$, efficiency β is normal without truncation, precision λ and choice bias κ are log-normal. In each case, both mean and variance are considered free parameters of the model. Overall the models thus have (up to) eight free parameters, which is conservative in relation to regression models used in experimental analyses and in relation

to the more progressive structural models (Harrison et al., 2007; Bellemare et al., 2008). Regardless, identifiability of the parameters is verified explicitly by analyzing reliability and consistency across experiments.

The supplementary material reports robustness checks investigating whether lower-dimensional models are possibly as adequate in-sample but more reliably identified and thus more robust out-of-sample. Heterogeneity of preferences is highly significant, as known from the literature and as indicated in Table 8. Heterogeneity of precision is similarly significant, both in-sample and out-of-sample. Heterogeneity of the choice bias κ is not significant in-sample but it is significant out-of-sample, i.e. allowing for heterogeneity of κ improves predictions for all “rich data sets” with at least eight observations per subject (AM02, HJ06, FKM07). In turn, identification is weak given the data from CHST07, as shown in the supplement, i.e. the two observations per subject do not allow us to disentangle preferences, precision and choice allowing for heterogeneity. For this reason, I skip predictions based on CHST07 in the main analysis, which allows me to use the most adequate model for the rich data sets allowing subjects to be heterogeneous in all four dimensions ($\alpha, \beta, \lambda, \kappa$). Aggregating over subjects, the log-likelihood of the model is

$$ll(\mathbf{d}|o) = \sum_{s \in S} \log l(\mathbf{d}|o_s) \quad (25)$$

with $o = \{o_s\}_{s \in S}$. Parameters are estimated by maximizing the log-likelihood, sequentially applying two maximization algorithms. Initially I use the robust, gradient-free NEWUOA algorithm (Powell, 2006), and subsequently I verify convergence using a Newton-Raphson algorithm. The estimates are tested by extensive cross-analysis to ensure that global maxima are found (as described in the supplementary material).

Models are discriminated by likelihood ratio tests that allow for both misspecified and potentially nested models, following Schennach and Wilhelm (2014) and as described in the supplement. Throughout the paper and the supplementary material, I indicate significance (p -values) at two levels, “weak significance” for $p = 0.05$ and “high significance” for $p = 0.005$. The former standard level has limited relevance in most cases, due to multiple testing problems resulting from testing several models on several data sets simultaneously. By analyzing multiple data sets in parallel, I avoid the reliance on a single p -value, addressing the concerns of Wasserstein and Lazar (2016) and providing the general picture, but to account for the multiple testing problem, significance at the stricter level is generally focused on.

References

- Andersen, S., Harrison, G., Lau, M., and Rutström, E. (2008). Eliciting risk and time preferences. *Econometrica*, 76(3):583–618.
- 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.

- Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, 11(2):122–133.
- Battistin, E., Miniaci, R., and Weber, G. (2003). What do we learn from recall consumption data? *Journal of Human Resources*, 38(2):354–385.
- Bellemare, C., Kröger, S., and van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76(4):815–839.
- Bernheim, B. D. and Rangel, A. (2009). Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124(1):51–104.
- Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890.
- Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Saliency theory of choice under risk. *The Quarterly Journal of Economics*, 127(3):1243–1285.
- Breitmoser, Y. (2013). Estimation of social preferences in generalized dictator games. *Economics Letters*, 121(2):192–197.
- Breitmoser, Y. (2016). The axiomatic foundation of logit and its relation to behavioral welfare. *MPRA Paper 71632*.
- Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.
- Camerer, C., Ho, T., and Chong, J. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898.
- Camerer, C. and Ho, T. H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874.
- 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.
- Covey, J. and Smith, R. (2006). How common is the ‘prominence effect’? Additional evidence to Whynes et al. *Health economics*, 15(2):205–210.
- Dana, J., Cain, D., and Dawes, R. (2006). What you don’t know won’t hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100(2):193–201.
- Dave, C., Eckel, C. C., Johnson, C. A., and Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, 41(3):219–243.

- Dean, M. L. (1980). Presentation order effects in product taste tests. *The Journal of psychology*, 105(1):107–110.
- Debreu, G. (1960). Review of 'Individual Choice Behavior' by R. Luce. *American Economic Review*, 50:186–8.
- Dinner, I., Johnson, E. J., Goldstein, D. G., and Liu, K. (2011). Partitioning default effects: why people choose not to choose. *Journal of Experimental Psychology: Applied*, 17(4):332.
- Echenique, F., Saito, K., and Tserenjigmid, G. (2014). The perception-adjusted luce model. *Working paper*.
- Engel, C. (2011). Dictator games: a meta study. *Experimental Economics*, 14:583–610.
- Feenberg, D. R., Ganguli, I., Gaule, P., Gruber, J., et al. (2015). It's good to be first: Order bias in reading and citing nber working papers. Technical report, National Bureau of Economic Research, Inc.
- Fisman, R., Kariv, S., and Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 97(5):1858–1876.
- Gul, F. and Pesendorfer, W. (2008). The case for mindless economics. *The foundations of Positive and normative Economics: A handbook*, pages 3–42.
- Harless, D. and Camerer, C. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, pages 1251–1289.
- Harrison, G. W. and Johnson, L. T. (2006). Identifying altruism in the laboratory. In Isaac, R. M. and Davis, D. D., editors, *Experiments Investigating Fundraising and Charitable Contributors*, volume 11 of *Research in experimental economics*, pages 177–223. Emerald Group Publishing Limited.
- Harrison, G. W., List, J. A., and Towe, C. (2007). Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion. *Econometrica*, 75(2):433–458.
- Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *Annals of Statistics*, pages 2244–2253.
- 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.
- Keane, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, 156(1):3–20.

- Keller, L. R. and Strazzera, E. (2002). Examining predictive accuracy among discounting models. *Journal of Risk and Uncertainty*, 24(2):143–160.
- Kőszegi, B. and Rabin, M. (2008). Choices, situations, and happiness. *Journal of Public Economics*, 92(8):1821–1832.
- Kőszegi, B. and Szeidl, A. (2013). A model of focusing in economic choice. *The Quarterly journal of economics*, 128(1):53–104.
- Lacetera, N., Pope, D. G., and Sydnor, J. R. (2012). Heuristic thinking and limited attention in the car market. *American Economic Review*, 102(5):2206–2236.
- Levine, D. K. (2012). *Is behavioral economics doomed?: The ordinary versus the extraordinary*. Open Book Publishers.
- List, J. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.
- Manski, C. F. and Molinari, F. (2010). Rounding probabilistic expectations in surveys. *Journal of Business & Economic Statistics*, 28(2):219–231.
- Manzini, P. and Mariotti, M. (2014). Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176.
- Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). Revealed attention. *American Economic Review*, 102(5):2183–2205.
- Matejka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–98.
- McFadden, D. (1974). 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.
- McKenzie, C. R., Liersch, M. J., and Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science*, 17(5):414–420.
- Miller, J. M. and Krosnick, J. A. (1998). The impact of candidate name order on election outcomes. *Public Opinion Quarterly*, 62:291–330.
- Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.
- Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience*, 29(44):14004–14014.

- Padoa-Schioppa, C. and Rustichini, A. (2014). Rational attention and adaptive coding: a puzzle and a solution. *The American economic review*, 104(5):507.
- Poltrock, S. E. and Schwartz, D. R. (1984). Comparative judgments of multidigit numbers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):32.
- Powell, M. (2006). The newuoa software for unconstrained optimization without derivatives. *Large-Scale Nonlinear Optimization*, pages 255–297.
- Read, D., Loewenstein, G., and Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, 19(1-3):171–97.
- Rust, J. (2010). Comments on: "structural vs. atheoretic approaches to econometrics" by Michael Keane. *Journal of Econometrics*, 156(1):21–24.
- Schennach, S. and Wilhelm, D. (2014). A simple parametric model selection test. Technical report, Cemmap working paper.
- 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.
- Train, K. (2003). *Discrete choice methods with simulation*. Cambridge Univ Pr.
- Wasserstein, R. L. and Lazar, N. A. (2016). The asa’s statement on p-values: context, process, and purpose. *The American Statistician* (forthcoming).
- Whynes, D., Philips, Z., and Frew, E. (2005). Think of a number... any number? *Health Economics*, 14(11):1191–1195.
- Wilcox, N. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In Cox, J. C. and Harrison, G. W., editors, *Risk aversion in experiments*, volume 12 of *Research in experimental economics*, pages 197–292. Emerald Group Publishing Limited.
- Wilcox, N. (2011). Stochastically more risk averse: A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, 162(1):89–104.