# The axiomatic foundation of logit

Breitmoser, Yves

Humboldt University Berlin

6 October 2016

# The axiomatic foundation of logit

Yves Breitmoser*
Humboldt University Berlin

October 6, 2016

**Abstract**

Multinomial logit is the canonical model of discrete choice but widely criticized for requiring specific functional assumptions as foundation. The present paper shows that logit is behaviorally founded without such assumptions. Logit's functional form obtains if relative choice probabilities are independent of irrelevant alternatives and invariant to utility translation (narrow bracketing), to relabeling options (presentation independence), and to changing utilities of third options (context independence). Least squares differs from logit only by making the additional assumption that utility is perceived to be quadratic around the utility maximizer, showing that logit is the more general model and least squares actually requires specific functional assumptions. Reviewing behavioral evidence, presentation and context independence seem to be violated in typical experiments, not IIA. Relaxing context independence yields contextual logit (Wilcox, 2011), relaxing presentation independence allows to capture "focality" of options.

*JEL–Code:* D03, C13
*Keywords:* stochastic choice, logit, axiomatic foundation, behavioral evidence, utility estimation, least squares

*I thank Friedel Bolle, Nick Netzer, Martin Pollrich, Sebastian Schweighofer-Kodritsch, Felix Weinhardt, 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

Applied theoretical analyses typically rest on preference assumptions as part of their model primitives. The necessity to understand preferences inspired a large body of work developing methods to infer preferences from choice. The main difficulty is that choice is inherently stochastic, which implies that we cannot directly infer preferences from stated choice.[1] Structural models attempt to control for stochastic mistakes in choice, but proponents of non-structural approaches argue that inference about preferences is impossible without making functional assumptions about individual choice. This renders inference on preferences unreliable. Indeed, the structural literature distinguishes three approaches of defining the locus of noise (random behavior, random preferences, and random utility),[2] for each approach a plethora of possible specifications of noise, and not a single model has been derived independently of specific functional assumptions. Thus, in response to the critique, Rust (2014, p. 820) writes that "there is an identification problem that makes it impossible to decide between competing theories without imposing ad hoc auxiliary assumptions" on say noise locus and distribution of noise.

This is troublesome, as both the assumed locus of noise and the distributional assumption are known to affect the results on identified preferences (Hey, 2005; Heckman, 2010). Further, different analysts indeed make different assumptions and thus obtain different results, which prevents the emergence of agreement on adequate representations of preferences. The plethora of approaches coexists exactly because no single approach has been founded without assuming a specific functional form at some point in the derivation. As a result, any comparison between alternative approaches boils down to judging different functional assumptions made in different places in the choice process, which appears to be impossible based solely on objective arguments (for related discussions, see e.g. Keane, 2010a,b, and Rust, 2010). For this reason, the coexistence of approaches, the diversity of contradicting results, and the general critique on structural analyses seem persistent, suggesting the literature approached a stalemate.

The present paper derives a behavioral foundation of multinomial logit,[3] solely relying on axioms on primitives of choice, thus showing that stochastic choice is founded without functional assumptions. This addresses the above critique and allows me to discuss logit and related models at a fundamental level: the assumptions underlying logit

[1] For example, choice is inconsistent across identical trials even after controlling for wealth and portfolio effects (Camerer, 1989; Starmer and Sugden, 1991), it violates the axioms of revealed preference (Andreoni and Miller, 2002; Fisman et al., 2007) and dominance relations (Birnbaum and Navarrete, 1998; Costa-Gomes et al., 2001). For further discussion of stochastic choice, see e.g. Hey (1995) and Wilcox (2008).

[2] Let $u(x|\alpha)$ denote the decision maker's utility given preference parameter $\alpha$ and $x^\star(\alpha)$ the utility maximizer. A decision maker with random behavior chooses $x^\star(\alpha)+\varepsilon$, with random preferences he chooses $x^\star(\alpha+\varepsilon)$, and with random utility he chooses $\arg\max_x\{u(x|\alpha)+\varepsilon_x\}$ for random variables $\varepsilon$ and $(\varepsilon_x)$.

[3] Multinomial logit is the most widely used model of stochastic choice. The long list of studies analyzing preferences using logit includes analyses of risk preferences (Holt and Laury, 2002; Goeree et al., 2003), social preferences (Cappelen et al., 2007; Bellemare et al., 2008), and preferences and demand functions of consumers (McFadden, 1980; Berry et al., 1995).

in relation to behavioral evidence, logit in relation to random behavior models and least squares analyses, and the intuition of how logit "averages" noise during preference estimation. This puts the subjective discussion of choice modeling on a solid basis, including the debate about parametric and nonparametric approaches, and it allows me to discuss and analyze generalizations of logit relating to behavior in standard experiments.

The main results can be summarized as follows. Choice probabilities have the specific logit form if choice satisfies independence of irrelevant alternatives (IIA), invariance to utility translation (narrow bracketing), invariance to relabeling (presentation independence), and invariance to changing utilities of third options (context independence). IIA implies that choice probabilities are functions of propensities, narrow bracketing implies a generalized logit form, presentation independence implies that solely utility is choice relevant, and context independence implies that perturbations have constant variance across choice tasks. Both presentation independence and context independence are routinely violated in economic experiments, while IIA and narrow bracketing seem to be compatible with behavior in "typical" experiments. In particular, evidence on choice violating IIA tends to resort to experiments explicitly studying similarity effects, while evidence contradicting presentation and context independence prevails across experiments.

Violations of context independence are comparably well-understood: choice is consistent across tasks if the range of potential outcomes is the same. This has been established econometrically (Wilcox, 2008, 2015) and explained neurophysiologically (Padoa-Schioppa and Rustichini, 2014; Rustichini and Padoa-Schioppa, 2015). To reflect this evidence, I also study a weak form of context independence, in conjunction with a cardinality axiom, which yields contextual logit (Wilcox, 2011). Experimental behavior appears to be largely compatible with both cardinality of utility and weak context independence, implying that contextual logit may be preferable to multinomial logit in applied work. Presentation effects are well-documented, though not formally understood. Choice has been shown to be affected by ordering, labeling, coloring, and positioning of options, including round-number and default effects. Dropping presentation independence shows that choice propensities then depend on two option characteristics, utility and focality. This finding is discussed briefly in Section 4 and extensively in Breitmoser (2016).

The results further show that both logit and contextual logit are the formal implication of assumptions tacitly made in most structural analyses. This includes random behavior and "least squares" analyses.[4] The latter equally assume IIA, narrow bracketing (even cardinality), and either context independent noise (similarly to logit) or context dependent, heteroscedastic noise (similarly to contextual logit). Further, all of these models assume that presentation effects are neutral in the sense that the utility maximizer always is the modal choice. The only difference between logit and least squares affects the way noise depends on presentation. Logit assumes that choice probabilities depend on utility

[4]Random behavior with normal trembles, i.e. least squares, has been used to estimate risk and time preferences (Choi et al., 2007; Andreoni and Sprenger, 2012), as well as utility parameters of subjects in dictator games (Fisman et al., 2007; Jakiela, 2013), public goods games (Bardsley and Moffatt, 2007), and auctions (Bajari and Hortacsu, 2005; Campo et al., 2011), to name just a few examples.

differences, while least squares assumes that choice probabilities depend on squared distances to the utility maximizer. To be clear, logit posits that the probability of choosing $x \in X$ given utility $u : X \to \mathbb{R}$ is

$$\Pr_{\text{Logit}}(x|X) = \frac{\exp\{\lambda \cdot u(x)\}}{\sum_{x' \in X} \exp\{\lambda \cdot u(x')\}},$$

for some noise parameter $\lambda \in \mathbb{R}$. Least squares is equivalent to assuming that choices are normally distributed around the utility maximizer $x^\star \in \arg\max_{x' \in X} u(x')$, assuming it is unique, with an unknown standard deviation $\sigma$. Thus, using $\phi$ to denote the standard normal density, least squares assumes that the choice probabilities are

$$\Pr_{\text{LS}}(x) = \frac{\phi\left(\frac{x-x^\star}{\sigma}\right)}{\sum_{x'} \phi\left(\frac{x'-x^\star}{\sigma}\right)} = \frac{\frac{1}{\sqrt{2\sigma^2\pi}} \cdot \exp\left\{-\frac{(x-x^\star)^2}{2\sigma^2}\right\}}{\sum_{x'} \frac{1}{\sqrt{2\sigma^2\pi}} \cdot \exp\left\{-\frac{(x'-x^\star)^2}{2\sigma^2}\right\}} = \frac{\exp\left\{-\lambda\,(x-x^\star)^2\right\}}{\sum_{x'} \exp\left\{-\lambda\,(x'-x^\star)^2\right\}}.$$

Note that this reformulation does not squeeze least squares into the logit form, but simply takes the normal density, the normalization constant $1/\sqrt{2\sigma^2\pi}$ cancels out, and the free parameters are aligned letting $\lambda = 1/2\sigma^2$. Thus, least squares obeys the logit form, i.e. logit's axioms, and additionally assumes that DM misperceives his asserted true utility $u$, for which the analyst estimates the parameters, as a quadratic function $\tilde{u}(x) = -(x-x^\star)^2$, or equivalently, $\tilde{u}(x) = u(x^\star) - (x-x^\star)^2$. This additional assumption is not supported by behavioral evidence, implying that logit uses not just theoretically weaker assumptions.[5] However, least squares provides a simple interpretation of how noise is averaged out during utility estimation and in principle requires little more than the back of an envelope to compute. Thus, least squares analyses may appear to be more transparent than logit, which may be taken informally as indication that the results are more robust. Logit has a similarly intuitive computational interpretation, derived below from its axiomatic foundation, which may help improve the perceived transparency of logit analyses. Briefly, take a parametric utility function and aggregate the utilities over all of DM's choices. Logit's estimate maximizes this aggregate utility (in a sense to be made precise), yielding the utility parameters under which DM's choices are as reasonable as possible, i.e. as close to utility maximization as possible.

Section 2 reviews the four existing foundations of logit, showing that all of them require specific functional assumptions in one place or another. Section 3 provides the behavioral foundations of multinomial logit and contextual logit avoiding such assumptions, solely using "axioms" stating invariance properties of choice. Section 4 discusses these axioms in relation to behavioral evidence and the computational intuition underlying logit. Section 5 concludes. The appendix contains all proofs.

[5] Note the difference to regression. Least squares robustly estimates the mean effect of some variable $x$ on another variable $y$. Analysts interested in utility parameters seek to understand how payoffs affect utilities and thus choice. This is not a regression, as the payoffs are not exogenous but depend on the choice made by DM, implying that least squares does not inherit the robustness from regression.

# 2 Existing foundations of logit

The notation is standard. Decision maker DM chooses option $x \in B$ from a finite budget $B \subseteq X$ with probability $\Pr(x|B)$. DM's utility $u : X \to \mathbb{R}$ is unknown, the subject of the analysis, and DM's choice exhibits stochastic noise with unknown distribution, the main obstacle of the analysis. The set of all finite subsets of $X$ is denoted as $P(X)$, and DM's choice profile Pr is a collection of probability distributions over all finite subsets of $X$, denoted as $\Pr = \{\Delta(B)\}_{B \in P(X)}$. The utility of option $x$ is denoted as $u_x$.

## 2.1 Unconditional logit

The original definition of logit, Luce (1959), states that choice is logit if a value function $v : X \to \mathbb{R}$ exists such that Pr has a logit representation. This definition is "unconditional" in that no condition about $v$'s relation to $u$ is imposed, distinguishing it from conditional logit defined by McFadden (1974) where $v = u$. Note that both conditional and unconditional models are called logit or multinomial logit in the literature.

**Definition 1** (Unconditional logit)**.** The choice profile Pr has an unconditional logit representation if there exists $v : X \to \mathbb{R}$ such that

$$\Pr(x|B) = \frac{\exp\{v(x)\}}{\sum_{x' \in B} \exp\{v(x')\}} \qquad \text{for all } x \in B \in P(X).$$

A scaling factor $\lambda$ as it is used below can be skipped without loss of generality. Since $v$ ex-post rationalizes DM's choice, I refer to it as DM's *choice utility*, thus distinguishing it from the true utility $u$. Note that $v$ is the choice utility specifically in relation to logit's functional form and defined only up to translation (addition of arbitrary constants).

Choice utility simply is a function of observed choice, for example $v(x) := \log \Pr(x|X)$ is adequate, and as such, it merely summarizes the information about utility contained in DM's choice profile. The main question will be what we can learn from it, i.e. how $v$ relates to $u$. To begin with, $v$ is defined if the choice profile Pr has an unconditional logit representation, which is the case if Pr exhibits independence of irrelevant alternatives (IIA). Assuming all choice probabilities are positive, Pr obeys IIA if

$$\frac{\Pr(x|B)}{\Pr(y|B)} = \frac{\Pr(x|B')}{\Pr(y|B')} \qquad \text{for all } x, y \in B \cap B', \tag{1}$$

for all $B, B' \in P(X)$. Following Luce (1959), the choice probabilities satisfy IIA if and only if a propensity function $V : X \to \mathbb{R}$ exists such that

$$\Pr(x|B) = \frac{V(x)}{\sum_{x' \in B} V(x')} \qquad \text{for all } x \in B \in P(X).$$

In this case, Pr is said to have a Luce representation. By positivity, Pr has a Luce repre-

sentation if and only if it has an unconditional logit representation, as $v(x) = \log V(x) = \log \Pr(x|X)$ for all $x \in X$ is then well-defined. That is, the choice probabilities satisfy IIA if and only if they have an unconditional logit representation, and in this sense, IIA and (unconditional) logit are equivalent. Fudenberg and Strzalecki (2015) establish this equivalence (amongst others) in a general model of dynamic choice.

Logit is not special in this respect, IIA is equivalent to any representation based on choice propensities. For example, fix any bijection $g : M \to \mathbb{R}_+$ for some $M \subseteq \mathbb{R}$ and say that Pr has an unconditional $g$-representation if $v : X \to \mathbb{R}$ exists such that

$$\Pr(x|B) = \frac{g(v(x))}{\sum_{x' \in B} g(v(x'))} \qquad \text{for all } x \in B \in P(X). \tag{2}$$

If choice satisfies IIA, then propensities $V(x)$ exist and Pr has a $g$-representation for any $g$, as $v(x) := g^{-1}(V(x))$ is well-defined. Thus, IIA is equivalent to any $g$-representation, rendering the equivalence of IIA and logit uninformative. As logit is only one of many possible specifications of $g$, unconditional logit thus makes a functional assumption ($g = \exp$). Unconditional logit is assumed without loss of generality only if choice utility $v$ is an affine transformation of true utility $u$. This obtains if Pr has a conditional logit representation, as defined next.

## 2.2 Conditional logit

DM's choice profile is conditional logit if the logit representation is adequate given the true utility function $u$. This follows McFadden (1974), who also analyzes the theoretical foundation of conditional logit.[6] To define the model, let us extend the notation by conditioning on $u$, i.e. given $u$, DM chooses option $x \in B$ with probability $\Pr(x|u,B) > 0$.

**Definition 2** (Conditional logit)**.** The choice profile Pr has a conditional logit representation if there exists $\lambda \in \mathbb{R}$ such that, given DM's utility $u : X \to \mathbb{R}$,

$$\Pr(x|u,B) = \frac{\exp\{\lambda \cdot u_x\}}{\sum_{x' \in B} \exp\{\lambda \cdot u_{x'}\}} \qquad \text{for all } x \in B \in P(X).$$

If Pr is conditional logit, then Pr also has an unconditional logit representation and the choice utility satisfies $v = \lambda u + r$ for some $r \in \mathbb{R}$. Then, the choice utility is an affine transformation of true utility $u$ and logit analyses indeed allow us to infer DM's utility.

Conditions for Pr to be conditional logit have been analyzed by McFadden (1974). In a first step, McFadden (1974) shows that positivity and IIA imply that DM's choice

[6] McFadden characterizes a logit model conditioning on individual attributes of DM. These individual attributes may represent free parameters in a utility representation such as CRRA. Conditional on these parameters, utility then is defined, and for the purpose of the current analysis, we may condition on the utility function itself, as is standard practice in behavioral analyses (see below).

probabilities can be represented as

$$\Pr(x|u,B) = \frac{\exp\{v(x,y|u)\}}{\sum_{x' \in B} \exp\{v(x',y|u)\}} \qquad \text{for all } x \in B, y \in X \tag{3}$$

for some function $v$, given any benchmark option $y \in X$. In contrast to the unconditional approach, which shows that choice utility simply is defined if choice is IIA, this shows that the choice utility of $x$ can be defined in relation to a single benchmark option $y$, i.e. references to other options $x' \in X$ are not required. McFadden (1974) derives Eq. (3) by defining $v(x,y|u)$ to be the log-odds of the choice between $x$ and $y$,

$$v(x,y|u) = \log\left(\frac{\Pr(x|u,\{x,y\})}{\Pr(y|u,\{x,y\})}\right). \tag{4}$$

IIA then implies Eq. (3). Since $\Pr(x|u,\{x,y\})$ and $\Pr(y|u,\{x,y\})$ may depend only on $x, y, u_x, u_y$, besides constants, this pins down the arguments of choice utility $v$. Any $y \in X$ may be chosen as benchmark option, but if $X$ is scarce, it may be impossible to express $v$ independently of a benchmark option; a richness condition resolves this issue below.

Eq. (4) does not substantially restrict $v$ and is compatible with many families of stochastic choice models, including strong utility, strict utility, and random behavior (including least squares),[7] implying that the relation of $v$ to DM's true utility $u$ is still undetermined. McFadden resolves this by Axiom 3 (page 110) assuming that the relative choice utility $v(x,y|u)$ is the difference of the utilities of $x$ and benchmark $y$.

$$v(x,y|u) = u_x - u_y \tag{5}$$

Given the exponential formulation of choice utility, the benchmark utility $u_y$ thus cancels out and choice utility $v(x)$ is implicitly assumed to equate with true utility $u_x$. Thus, Axiom 3 achieves the following: out of the vast set of potential functional forms compatible with $v(x,y,u_x,u_y)$, it selects $v(x) = u_x$, implying that the benchmark utility $u_y$ and the options $x$ and $y$ as such are choice irrelevant, but it obviously represents a specific functional assumption. The nature of the assumption becomes clearer using $v$'s definition Eq. (4), which implies that McFadden's Axiom 3 is equivalent to assuming

$$\begin{aligned}
\frac{\Pr(x|u,\{x,y\})}{\Pr(y|u,\{x,y\})} = \exp\{u_x - u_y\} \quad &\Leftrightarrow \quad \frac{\Pr(x|\cdot) + \Pr(y|\cdot)}{\Pr(y|u,\{x,y\})} = 1 + \exp\{u_x - u_y\} \\
\Leftrightarrow \quad \frac{\Pr(y|u,\{x,y\})}{\Pr(x|\cdot) + \Pr(y|\cdot)} = \frac{1}{1+\exp\{u_x - u_y\}} &= \frac{\exp\{u_y\}}{\exp\{u_x\} + \exp\{u_y\}} \\
&\Leftrightarrow \quad \Pr(x|u,\{x,y\}) = \frac{\exp\{u_x\}}{\exp\{u_x\} + \exp\{u_y\}},
\end{aligned}$$

[7] Random behavior has been defined in Footnote 2. Pr has a strong utility representation if $\Pr(x|u,B) = f(u_x - u_y)/\sum_{x' \in B} f(u_{x'} - u_y)$ for some $f : \mathbb{R} \to \mathbb{R}_+$ and $y \in X$. Pr has a strict utility representation if $\Pr(x|u,B) = (u_x)^\lambda / \sum_{x' \in B} (u_{x'})^\lambda$ for some $\lambda \in \mathbb{R}$. See also Luce and Suppes (1965).

noting that $\Pr(x|\cdot) + \Pr(y|\cdot) = 1$. The last equation is the definition of binomial logit (omitting $\lambda$), i.e. Axiom 3 is equivalent to assuming that binomial choice is logit. In turn, logit itself is not behaviorally founded; IIA merely extrapolates binomial logit to multinomial choice. This implication of Axiom 3 does not seem to have been observed in the existing literature, but it clearly shows that the existing foundation of conditional logit makes a functional assumption. Instead of assuming that binomial choice is logit, we could assume any other structure of binomial choice and then would obtain any other model compatible with IIA. For example, replacing Axiom 3 with $v(x,y|u) = g(u_x - u_y)$ for any monotone and positive $g$, we obtain any strong utility model.

## 2.3 Foundation as random utility model

Thurstone (1927) introduced the random utility model for binomial choice, focusing on utility perturbations with normal distribution. Block and Marschak (1960) introduced the multinomial random utility model allowing for arbitrary distributions of the utility perturbations. Accordingly, choice profile Pr has a random utility representation if, given utility $u$, there exists a collection of random variables $(R_x)_{x\in X}$ such that

$$\Pr(x|u,B) = P\big(u_x + R_x \geq \max_{x'\in X} u_{x'} + R_{x'}\big) \tag{6}$$

for all $x \in B$ and $B \in P(X)$. McFadden (1974) shows that conditional logit results if the utility perturbations $(R_x)$ are i.i.d. with extreme value type 1 distribution, Yellott (1977) shows that an i.i.d. random utility model satisfies IIA if and only if the utility perturbations have this particular distribution, and Strauss (1979) generalizes the result to the non-i.i.d. case. Thus, random utility models with any alternative distribution, whether or not the perturbations are i.i.d., violate independence of irrelevant alternatives.[8] In this sense, the extreme value distribution is indeed specific: it is not one of many possible choices, but the only possible choice compatible with IIA. Given IIA, in turn, the critical assumption is not that the utility perturbations have an extreme value distribution, but that the choice profile admits a random utility representation in the first place.

Considering the plethora of stochastic choice models that satisfy IIA, the assumed adequacy of the random utility representation is obviously not innocuous. Indeed, given IIA, assuming that the choice probabilities have a random utility representation is equivalent to assuming that binomial choice is logit (see also Adams and Messick, 1958)—given IIA, either assumption implies that multinomial choice is logit. This shows that an assumption equivalent to McFadden's Axiom 3 is implied by assuming adequacy of the random utility representation, although it is less obvious.

Relatedly, Thurstone's additive random utility model is not the only way of rep-

[8] Robertson and Strauss (1981) clarify the reason. Let $Y$ denote the maximum of $n$ random variables that are i.i.d. aside from location shifts and let $I$ denote the index of the variable attaining the maximum. $Y$ and $I$ are independent if and only if the random variables have the extreme value distribution. This independence ensures that the odds of choosing between two options are independent of the options otherwise available.

resenting stochastic choice by means of random variables. Alternative models include random behavior models (see e.g. Harless and Camerer, 1995) and random preference models (Falmagne, 1978; Barberà and Pattanaik, 1986), and within all model families, there are countless functional forms of incorporating perturbations. Not all of these functional forms are equally appealing, but it is clear that the functional form assumed with the additive random utility representation is just one of many possibilities.

## 2.4 Foundation in rational inattention

Matejka and McKay (2015) model choice if DM is rationally inattentive in the sense of Sims (2003). DM has limited information about the state of the world, and the state of the world defines DM's mapping of options to utilities. DM may study the state, at a cost, to reduce the uncertainty he faces. Implicitly, DM has to choose which options to study and when to stop, trading off the knowledge he gains about his utility function and his costs of studying it. After studying the state of the world, DM chooses the option with the highest expected utility. DM can buy information about the state at costs proportional to the amount of uncertainty removed by the obtained information, and here, uncertainty is measured using Shannon entropy.[9]

Matejka and McKay show that DM's choice probabilities have a generalized logit representation: given utility $u$, there exist a function $w : X \to \mathbb{R}$ and some $\lambda \in \mathbb{R}$ such that

$$\Pr(x|B) = \frac{\exp\{\lambda \cdot u_x + w(x)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u_{x'} + w(x')\}} \qquad \text{for all } x \in B \in P(X).$$

Matejka and McKay show that $w(x)$ reflects DM's prior beliefs about the optimal option, which in turn depends on the prior belief about the state and the set of possible states. By knowing the set of possible states, DM detects similar options and implicitly adapts his information strategy to similarity. Thus, $w(x)$ captures similarity effects and allows for violations of IIA as predicted by the red-bus/blue-bus example of Debreu (1960).

If DM's prior belief is flat, then $w(x) = \text{const}$ and cancel out, yielding conditional logit. Matejka and McKay (2015) work with the standard model of rational inattention and use the most widely adopted measure of entropy, but the Shannon entropy represents only an instance of a large family of entropy measures (Rényi, 1960). Its assumption is not behaviorally founded and thus it does not resolve the issue that specific functional assumptions must be made to characterize logit. For example, discussing Matejka and McKay's cost function based on Shannon entropy, Caplin and Dean (2015) "outline key behavioral properties implied by this cost function, which are significantly more restrictive than NIAS and NIAC alone" (p. 2), referring to two general conditions (NIAS and NIAC) characterizing rational information acquisition.

[9] The Shannon entropy of a random variable is defined as $H = -\sum_i P(s_i) \log P(s_i)$, with $(s_i)$ as possible realizations of the random variable and $P(s_i)$ as their respective probabilities.

## 2.5 Discussion

Parametric structural analyses are criticized for their functional assumptions, e.g. on error distributions or on choice functions. This concerns in particular also logit, as all four of logit's foundations build on functional assumptions, but is equally true for any other choice model. To refute the criticism, we may pick any of the four foundations and derive the functional assumption from say invariance assumptions. Completion of the other foundations follows by corollary. I will focus on the behavioral foundation of conditional logit. Conditional logit is generally preferred in applied work,[10] in relation to unconditional logit it mitigates the standard critique that utilities are defined post-hoc to rationalize choice,[11] and in relation to the other approaches, it avoids unobservable non-primitive entities such as utility perturbations or information purchases.

Another advantage of conditional logit is that it naturally allows to analyze choice across contexts. For example, experimenters tend to vary prizes in lotteries or transfer rates in dictator games, both as defined shortly, within subjects. In order to study the foundations of logit also in such analyses, we need to condition on the context. Different contexts induce different (unknown) utility functions across options, and thus it suffices to condition on the unknown but distinct utility function induced in a given context. This, in turn, is done in conditional logit.

# 3 The axiomatic foundation of logit

## 3.1 The model

As indicated, I extend the usual framework by allowing that the experiment spans multiple "contexts". This reflects the standard practice to analyze choice in response to varying prizes in lotteries, to varying transfer rates in Dictator games, and to varying signals in auctions. In all of these examples, the mapping from options to utilities varies across tasks, to which I refer as a variation of context. Typical assumptions in such analyses are that noise variance either is constant or varies as a function of context in a specific manner, but no such assumption has been behaviorally founded. From a technical point of view, allowing for context variation allows me to analyze "cardinal utility" rigorously, i.e. the implications of requiring choice to be robust to translation or affine transformation of utility. Such transformations induce different utility functions, i.e. different contexts,

[10] Seemingly all behavioral analyses using logit define logit according to Def. 2, i.e. take at least the functional form of utility as given or fix it entirely. The list of examples is extremely long and includes analyses of risk preferences, for overviews see Wilcox (2008) and Harrison and Rutström (2008), time preferences (Andersen et al., 2008), and social preferences (Goeree et al., 2002; Cappelen et al., 2007). Analyses of strategic behavior and learning usually fix utility uniquely and vary only precision $\lambda$, following McKelvey and Palfrey (1995) and Camerer and Ho (1999). For a review, see Camerer and Ho (2015).

[11] See e.g. Cohen and Dickens (2002). This assumes, of course, that conditional logit analyses adapt the utility function it fits behavior, which would be equivalent to the unconditional approach.

and hence their analysis requires a formal representation of context variation.

The notation is extended appropriately. A choice task is a duple $(u,B)$, where decision maker DM has to choose an option $x \in B$ from a finite set $B \subseteq X$ given his utility $u : X \to \mathbb{R}$. Given choice task $(u,B)$, the probability that DM chooses $x \in B$ is denoted as $\Pr(x|u,B)$. The set of choice tasks $(u,B)$ is $\mathcal{D} = \mathcal{U} \times P(X)$; $\mathcal{U}$ denotes the set of unknown utility functions $u : X \to \mathbb{R}$ underlying DM's choices in the various contexts, and $P(X)$ denotes the set of finite subsets of $X$. To be clear, $u_x$ captures the welfare DM derives from option $x$ in a given context—the utility function is known to exist, but its values are unknown to the analyst and the object of his analysis. To clarify the model primitives, let me discuss two examples.

**Example 1** (Choice under risk)**.** *There are four prizes,* $(\pi_1, \pi_2, \pi_3, \pi_4)$ *and each option is a lottery* $L = (\pi_i, p, \pi_j)$ *yielding* $\pi_i$ *with probability* $p$ *and* $\pi_j$ *with probability* $1-p$*. The set of options is* $X = [0,2]$ *and option* $x \in X$ *is defined as*

$$L(x) = \begin{cases} (\pi_1, x, \pi_2), & \text{if } x \leq 1 \\ (\pi_3, x-1, \pi_4), & \text{if } x > 1 \end{cases}$$

*The unknown utility* $u_x$ *is DM's (expected) utility of lottery* $L(x)$*,* $x \in X$*. Different prizes induce different contexts, i.e. different mappings from options to (expected) utility.*

It is straightforward to generalize the example to multinomial choice or choice from lotteries with more than two possible outcomes by partitioning $X$ into more subsets. Many experiments implement lists of such choice tasks following Holt and Laury (2002). These lists ask DM to choose between risky and save lotteries, $\pi_1 > \pi_3 > \pi_4 > \pi_2$, for a sequence of probabilities such as $p = 0.1, 0.2, \ldots, 1$. Using the above notation, such a list consists of the tasks $\{0.1, 1.1\}, \{0.2, 1.2\}, \ldots, \{1,2\}$, i.e. $\{\{k/10, 10 + k/10\}\}_{k=1,\ldots,10}$.

**Example 2** (Dictator game)**.** *DM is endowed with* $E$ *tokens, each token is worth* $\tau_1$ *points to DM and* $\tau_2$ *points to a second player (recipient). The set of options is* $[0,1]$*, and option* $x \in X$ *implies that DM keeps* $x \cdot E$ *tokens for himself and transfers* $(1-x) \cdot E$ *tokens to the recipient. The unknown utility function maps options (or, point distributions) to DM's welfare, and different transfer rates* $\tau_1, \tau_2$ *or endowments* $E$ *induce different contexts.*

Experimental analyses often involve variation of transfer rates and endowments (i.e. "contexts") within subjects, see e.g. Andreoni and Miller (2002), Harrison and Johnson (2006), and Fisman et al. (2007). Generalized dictator games allowing for "taking" from the recipient's endowment and incomplete information of the recipient about the number of tokens available to DM are captured straightforwardly by adapting $X$ or defining prior beliefs on the distribution of the endowment.

**Maintained assumptions** Throughout the paper, I assume that the set of choice tasks $\mathcal{D}$ is "rich" and that the choice probabilities $\Pr(\cdot|u,B)$ are positive.

**Assumption 1** (Richness)**.** The set of choice tasks $\mathcal{D} = \mathcal{U} \times P(X)$ is called rich if

1. *Transformability:* $a + b\,u \in \mathcal{U}$ for all $u \in \mathcal{U}$ and all $a, b \in \mathbb{R} : b > 0$,
2. *Convexity:* $X$ is a convex subset of $\mathbb{R}$ and $|X| > 1$,
3. *Surjectivity:* for all $u \in \mathcal{U}$, the image $u[X] = \{u_x | x \in X\}$ is a convex subset of $\mathbb{R}$ and not a singleton, and
4. *Choice variation:* there exists $u \in \mathcal{U}$ and $x, x' \in X$ such that $\Pr(x|u,X) \neq \Pr(x'|u,X)$.

Transformability ensures that we may analyze affine transformations of utility functions in the first place, by ensuring that all affine transformations are well-defined objects. Convexity and surjectivity primarily rule out scarce choice environments where the sets of options or realized utility levels (respectively) are finite or even singletons; but it will be notationally convenient to know that both domain and image of DM's utility are convex. Such assumptions are similarly made by Gul et al. (2014) and Fudenberg et al. (2015) and satisfied in choice tasks typically of interest to experimentalists (as in the examples above, using standard utility functions). Note that the utility functions may still be fairly ill-behaved, violating smoothness or even continuity for any number points. Further, "surjectivity" permits us to normalize utilities through dividing by $\sup u - \inf u$.[12] Finally, "Choice variation" rules out that choice probabilities are uniform in all contexts.

**Assumption 2** (Positivity)**.** For all choice tasks $(u, B) \in \mathcal{D}$ and all $x \in B$, $\Pr(x|u,B) > 0$.

Positivity assumes that DM does not generally manage to maximize utility and captures the widely documented phenomena that individual choice fluctuates and that dominated options have positive probability, i.e. options that fail to maximize utility for any conceivable utility function. This has been observed in choice under risk and uncertainty (Birnbaum and Navarrete, 1998), in small normal-form games (Costa-Gomes et al., 2001), and through violations of revealed preference axioms in simple distribution decisions such as dictator games (Andreoni and Miller, 2002; Fisman et al., 2007). Stochastic choice offers a simple explanation of such observations. Positivity does not imply restrictions on the locus of noise in the choice process, i.e. it is compatible with random behavior, random utility and even random preferences.[13] Positivity also is technically mild in the sense that empirically, an event occurring with zero probability is indistinguishable from one occurring with positive but small probability (McFadden, 1974).

## 3.2 Independence of Irrelevant Alternatives and Luce

IIA has been introduced in Eq. (1), but let me restate IIA for the more general choice environment analyzed now, allowing IIA to hold for each context $u \in \mathcal{U}$.

[12] Writing $\sup u$ and $\inf u$, I refer to $u$'s supremum and infimum, respectively, over its domain ($X$), i.e. $\sup u = \sup_{x \in X} u_x$ and $\inf u = \inf_{x \in X} u_x$.

[13] Random preference models (Falmagne, 1978; Barberà and Pattanaik, 1986) violate positivity in some contexts, but in general they are ruled out only by IIA. Random behavior models will be ruled out by presentation independence, as discussed below. Thus, for the purpose of interpretation, the reader may assume that DM has a well-defined utility function but a perturbed perception of it, as in the random utility model Eq. (6) or in the rational inattention model of Matejka and McKay (2015).

**Axiom 1** (Independence of Irrelevant Alternatives, IIA)**.** For all $(u,B),(u,B') \in \mathcal{D}$,

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

Since Debreu (1960), IIA has been criticized for its incompatibility with similarity effects, i.e. the intuition that similar options are not evaluated and chosen independently (further discussed below). In typical experiments, similarity effects are deliberately limited by experimental design, to enable clean inference unless the purpose is to study similarity effects. As a foundation of IIA, Gul et al. (2014) show that if choice probabilities are countably additive, IIA generally obtains if DM's (stochastic) preference ordering is complete. Given IIA, choice probabilities have a Luce representation, i.e. a propensity function $V : X \to \mathbb{R}$ exists such that $\Pr(x|B) = V(x)/\sum_{x' \in B} V(x')$ (Luce, 1959). For example, define $V(x) := \Pr(x|X) \cdot r$ for any $r > 0$. The Luce representation and the equivalence to IIA straightforwardly generalizes to multiple contexts. The following result further shows that propensities are functions solely of $x$ and $u_x$, thus tightening the result of McFadden (1974) discussed above using the richness assumption.

**Definition 3** (Luce)**.** The choice profile Pr is Luce if there exists a family of functions $\{V_u : X \times \mathbb{R} \to \mathbb{R}\}_{u \in \mathcal{U}}$ such that for all tasks $(u,B) \in \mathcal{D}$ and options $x \in B$,

$$\Pr(x|u,B) = V(x|u)/\sum_{x' \in B} V(x'|u) \qquad \text{with } V(x|u) = V_u(x,u_x). \tag{7}$$

**Lemma 1.** Pr *is Luce* $\Leftrightarrow$ Pr *satisfies Axiom 1.*

Choice propensities $V_u$ may still be context dependent, as IIA itself does not restrict choice across contexts. Even the functional forms of $V_u$ may vary across contexts, and expressed in terms of model primitives, $V$ simply is a collection of functions $\{V_u\}_{u \in \mathcal{U}}$ mapping options $x$ and utilities $u_x$ to real-valued propensities, for all $u \in \mathcal{U}$. Applied to any single context, this result is tighter than McFadden's, as it shows that the reference to a benchmark $y$ and its utility $u_y$ are not required. Still, IIA is compatible with a wide range of choice models. As illustration, consider the following family of choice models satisfying IIA with choice propensities being functions solely of $x$ and $u_x$.

$$\Pr(x|u,B) = \frac{V_u(x,u_x)}{\sum_{x' \in B} V_u(x',u_{x'})} \quad \text{with} \quad V_u(x,u_x) = c_{1|u} + f_u\big(u_x - c_{2|u}\big) + g_u\big(x - c_{3|u}\big) \tag{8}$$

with $\{f_u, g_u\}_{u \in \mathcal{U}}$ being context-specific functions ($\mathbb{R} \to \mathbb{R}$), and for the purpose of illustration, they involve context-specific constants $\{c_{1|u}, c_{2|u}, c_{3|u}\}_{u \in \mathcal{U}}$. Let for example $c_{2|u} = \sup_{x \in X} u_x$ and (if existent) $c_{3|u} = \arg\max_{x \in X} u_x$, implying that the strong utility and random behavior models are contained as special cases. This shows that the locus of noise is virtually unrestricted by IIA, only similarity effects are ruled out. Implicitly, we cannot infer any information on the relation of propensities $V$ and utilities $u$ from IIA. In relation to this family of models, McFadden's Axiom 3 assumes $V(x|u) = \exp\{u_x - u_y\}$

for some $y \in X$, i.e. specifically $f_u = \exp$, $c_{2|u} = u_y$, and $c_{1|u} = g_u = 0$.

## 3.3 Narrow bracketing and cardinality

Standard representation theorems imply that utility is defined only up to affine transformation. These theorems assume rational choice and it is not obvious why they should generalize to stochastic choice. Robustness to affine transformations may still appear desirable, however, as it justifies standard assumptions in applied work. On the one hand, invariance to translation of utilities (addition of arbitrary constants) implies that, if we assume that DM's utility is the sum of "background utility" and "experiment utility", the background utility can be factored out and the choice pattern is invariant to the level of the background utility. Then, DMs approach any single choice task independently of background utility and previous tasks, which is generally assumed in behavioral analyses. Following Read et al. (1999), I refer to it as narrow bracketing.

**Axiom 2** (Narrow bracketing)**.** $\Pr(\cdot|u,B) = \Pr(\cdot|u+r,B)$ for all $r \in \mathbb{R}$, $(u,B) \in \mathcal{D}$

On the other hand, invariance of choice to scaling utilities is robustly observed in experiments. A detailed discussion follows below, but essentially, when experimental payoffs are scaled, expected utilities of options scale proportionally under standard assumptions,[14] but observed choice probabilities are largely unaffected by such scaling. This holds both within subjects and between subject; for discussion, see e.g. Wilcox (2011) and Padoa-Schioppa and Rustichini (2014).

**Axiom 3** (Cardinality)**.** $\Pr(\cdot|u,B) = \Pr(\cdot|a+b\,u,B)$ for all $a,b \in \mathbb{R} : b > 0$, $(u,B) \in \mathcal{D}$

Narrow bracketing obtains if choice propensities are functions of utility differences, as in strong utility models (Block and Marschak, 1960), and scale invariance obtains if propensities are functions of utility ratios, as in strict utility models. While strong utility models and strict utility models in the strict sense have an empty intersection,[15] requiring robustness to affine transformation is of course not prohibitive. Amongst others,

$$\Pr(x|u,B) = f\left(\frac{u_x - \inf u}{\sup u - \inf u}\right) \Big/ \sum_{x' \in B} f\left(\frac{u_{x'} - \inf u}{\sup u - \inf u}\right)$$

for any function $f : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies cardinality (and IIA). With $f(r) = \exp(r)$ we obtain contextual logit (Wilcox, 2011), and with $f(r) = r^\lambda$ we obtain a normalized strict utility model (noting that the denominator cancels out). Similarly, all random behavior models (including least squares) are consistent with both cardinality and IIA. The next result establishes that in general, narrow bracketing merely implies a "relative Luce" representation of choice and cardinality implies a "standardized Luce" representation.

[14] This applies if the utility function is homogeneous in the payoffs, which is satisfied for utility functions used in behavioral analyses, such as CRRA, CES, inequity aversion or Prospect theoretic utilities.

[15] Recall the definition in Footnote 7 or see Luce and Suppes (1965).

**Definition 4** (Relative/Standardized Luce)**.** The choice profile Pr is relative (standardized) Luce if there exist functions $\{V_u : X \times \mathbb{R} \to \mathbb{R}_+\}_{u \in \mathcal{U}}$ such that for all choice tasks $(u,B) \in \mathcal{D}$ and all options $x \in B$, $\Pr(x|u,B) = V(x|u)/\sum_{x' \in B} V(x'|u)$ with

$$V(x|u) = V_u(x, u_x - \inf u), \qquad \text{(Relative Luce)}$$

$$V(x|u) = V_u\big(x, \tfrac{u_x - \inf u}{\sup u - \inf u}\big). \qquad \text{(Standardized Luce)}$$

**Lemma 2.**

1. *Axioms 1 and 2* $\Leftrightarrow$ Pr *is relative Luce with* $V_u = V_{u+r}$ $(\forall\, r \in \mathbb{R})$
2. *Axioms 1 and 3* $\Leftrightarrow$ Pr *is standardized Luce with* $V_u = V_{a+bu}$ $(\forall\, a,b \in \mathbb{R} : b > 0)$

This suggests that neither narrow bracketing nor cardinality are restrictive. To illustrate, the family of representations compatible with IIA and cardinality include

$$\Pr(x|u,B) = \frac{f_u\big(\frac{u_x - \inf u}{\sup u - \inf u}\big) + g_u(x - x^\star)}{\sum_{x' \in B} f_u\big(\frac{u_{x'} - \inf u}{\sup u - \inf u}\big) + g_u(x' - x^\star)} \tag{9}$$

for functions $\{f_u, g_u : \mathbb{R} \to \mathbb{R}_+\}_{u \in \mathcal{U}}$, assuming $f_u = f_{a+bu}, g_u = g_{a+bu}$ for $a,b \in \mathbb{R} : b > 0$ (reflecting the conditions in Lemma 2). Besides contextual logit and normalized strict utility as discussed above, this still allows for general random behavior models, using $f_u = 0$ and $x^\star \in \arg\max u$ (assuming it is defined), for least squares if additionally $g_u(y) = \phi(y/\sigma)$ with $\phi$ as standard normal density, and for arbitrary combinations of say strict utility and random behavior. Thus, neither IIA nor cardinality (or narrow bracketing) seem to imply any restriction of how choice propensities relate to utility $u$.

This impression is misleading. If choice is consistent across contexts, in a sense to be made precise, then narrow bracketing and cardinality allow us to infer that $f_u(r) = \exp(\lambda r)$ for all contexts $u \in U$. This will imply that choice is represented by generalized formulations of conditional logit and contextual logit, depending on whether we require narrow bracketing or cardinality. Thus, on their own, narrow bracketing and cardinality are fairly weak requirements, but they have further implications once we know more about choice across contexts.

## 3.4 Presentation independence and context independence

Fix any utility function $u$ and assume, for purpose of illustration, that $u_x = 2$ and $u_y = 0$, for some $x,y \in X$. Now consider $u' = u + 8$, which implies $u'_x = 10$ and $u'_y = 8$. By narrow bracketing, or cardinality, we know that the relative probability of choosing $x$ over $y$ is equal in both contexts $u$ and $u'$. Two seemingly related invariances are not implied. On the one hand, assume there exist $x', y' \in X$ with utilities 10 and 8 in the original context $u$, i.e. $u_{x'} = 10$ and $u_{y'} = 8$. Narrow bracketing does not imply that the

relative probability of choosing 10 ($x'$) over 8 ($y'$) in context $u$ is equal to the one of choosing 2 ($x$) over 0 ($y$) in context $u$—although we know that choosing between 2 and 0 under $u$ is equivalent to choosing between 10 and 8 in a different context $u'$. I refer to this phenomenon as "presentation effect": The probability of choosing an option with a given utility may depend on which option attains this utility. For example, presentation effects may reflect labeling or ordering of options, and are even implied in random behavior models. Random behavior assumes that choice probabilities depend on the distance to the utility maximizer, implying that options with equal utilities have different choice probabilities if utility is not symmetric around the maximizer. Formally, presentation effects are compatible with relative Luce, as choice propensities are functions $V_u(x, u_x - \inf u)$, i.e. option $x$ itself is choice relevant. Presentation independence results if choice satisfies permutation invariance: given context $u \in \mathcal{U}$ and any bijective function $f : X \to X$, permuting choice probabilities (via $f$) is equivalent to permuting utilities (via $f$),

$$\Pr(f(x) \,|\, u, f(B)) = \Pr\left(x \,|\, u \circ f, B\right) \qquad \text{for all } x \in B \in P(X). \tag{10}$$

Intuitively, given presentation independence, propensities can be expressed as functions $V_u(u_x - \inf u)$ independently of $x$ itself, but this is not formally implied, as $u \circ f$ represents a context different from $u$, i.e. we also need information on context dependence of choice.

On the other hand, assume there exists $u''$ such that $u''_x = 2$ and $u''_y = 0$, but $u \neq u''$. Hence $u''$ is neither a translation nor an affine transformation of $u$, and choice propensities under $u$ and $u''$ may be entirely unrelated given Lemma 2. This captures "context dependence": The relative probabilities of choosing options with given utilities depend on context. Strict context independence obtains if for all $u, u' \in \mathcal{U}$ and all $x, y \in X$,

$$u_x = u'_x \quad \text{and} \quad u_y = u'_y \qquad \Rightarrow \qquad \Pr(x|u, \{x,y\}) = \Pr(x|u', \{x,y\}). \tag{11}$$

By IIA, this implies that the relative probability of choosing $x$ over $y$ is equal in $u$ and $u'$ for all budget sets $B \in P(X)$. Given the behavioral evidence reviewed below, strict context independence appears to be unrealistically strict, and for this reason, let me also introduce the notion of weak context independence: Implication (11) applies only if the utility range in contexts $u$ and $u'$ is equal, i.e. if $\sup u - \inf u = \sup u' - \inf u'$. I say that choice exhibits strict/weak utility relevance if it exhibits presentation independence and strict/weak context independence, respectively.

**Axiom 4** (Strict utility relevance, SUR)**.** For all $u, u' \in \mathcal{U}$ and all $x, x', y, y' \in X$,

$$u_x = u'_{x'} \quad \text{and} \quad u_y = u'_{y'} \qquad \Rightarrow \qquad \Pr(x|u, \{x,y\}) = \Pr(x'|u', \{x',y'\}).$$

**Axiom 5** (Weak utility relevance, WUR)**.** For all $u, u' \in \mathcal{U}$ : $\sup u - \inf u = \sup u' - \inf u'$,

$$u_x = u'_{x'} \quad \text{and} \quad u_y = u'_{y'} \qquad \Rightarrow \qquad \Pr(x|u, \{x,y\}) = \Pr(x'|u', \{x',y'\}).$$

As indicated, the behavioral evidence suggests that assumptions stronger than Ax-

iom 5 may be inadequate, but before I enter this discussion, let me clarify the main result.

**Definition 5.** The choice profile Pr is **conditional logit** or **contextual logit** (respectively) if there exists $\lambda \in \mathbb{R}$ such that for all choice tasks $(u,B) \in \mathcal{D}$ and all options $x \in B$, $\Pr(x|u,B) = V(x|u)/\sum_{x' \in B} V(x'|u)$ with

$$V(x|u) = \exp\{\lambda \cdot u_x\}, \qquad \text{(Conditional logit)}$$
$$V(x|u) = \exp\{\lambda \cdot u_x/(\sup u - \inf u)\}. \qquad \text{(Contextual logit)}$$

**Theorem 1.**

1. Pr *is conditional logit* $\Leftrightarrow$ Pr *satisfies Axioms 1, 2, 4*
2. Pr *is contextual logit* $\Leftrightarrow$ Pr *satisfies Axioms 1, 3, 5*

Briefly, let me discuss the relative contributions of the three axioms per representation. By IIA, Pr has a Luce representation, and by narrow bracketing, choice propensities have the form $V_u(x, u_x - \inf u)$. Now, by WUR, options with equal utility must have equal choice propensities, i.e. $u_x = u_y$ implies $V_u(x, u_x - \inf u) = V_u(y, u_y - \inf u)$, which in turn implies $V_u(x, u_x - \inf u) = V_u(y, u_x - \inf u)$. As a result, using any $u^{-1}$ such that $u(u^{-1}(u_x)) = u_x$ for all $x$, we can define a function $\tilde{V}_u(u_x) = V_u(u^{-1}(u_x), u_x - \inf u)$ representing choice propensities solely as functions of utilities. This does not yet eliminate presentation effects, but it restricts the functional form of choice probabilities. Again, take $u \in U$ such that $u_x = 2$ and $u_y = 0$. Fix $u' = u + 8$, implying $u'_x = 10$ and $u'_y = 8$. By narrow bracketing, we know that the relative probability of choosing $x$ over $y$ is the same in both contexts. Now assume $u_{x'} = 10$ and $u_{y'} = 8$ for some $x', y' \in X$. Since $\sup u - \inf u = \sup u' - \inf u'$, WUR (first equation), transitivity (middle equation), and the simplified representation of choice propensities (last equation) yield

$$\frac{\Pr(x|u',B)}{\Pr(y|u',B)} = \frac{\Pr(x'|u,B)}{\Pr(y'|u,B)} \quad \Rightarrow \quad \frac{\Pr(x|u,B)}{\Pr(y|u,B)} = \frac{\Pr(x'|u,B)}{\Pr(y'|u,B)} \quad \Rightarrow \quad \frac{\tilde{V}_u(u_x)}{\tilde{V}_u(u_y)} = \frac{\tilde{V}_u(u_x + r)}{\tilde{V}_u(u_y + r)}$$

for all $r \in R$ (in the example, $r = 8$ was assumed). The generalization to all $B \in P(X)$ obtains by IIA, which in turn yields the implication for propensities. Thus, $\tilde{V}_u(u_x + r) = \tilde{V}_u(u_x) \cdot f(r)$, for some function $f : \mathbb{R} \to \mathbb{R}$, and differentiating with respect to $r$ implies

$$d\tilde{V}_u(u_x + r)/dr = \tilde{V}_u(u_x) \cdot f'(r) \qquad \Rightarrow \qquad d\tilde{V}_u(u_x)/du_x = \tilde{V}_u(u_x) \cdot f'(0)$$

at $r = 0$. The solution of this differential equation is $\tilde{V}(u_x) = \exp\{\lambda \cdot u_x + w_x\}$, with $\lambda = f'(0)$ and $w_x$ as an integration constant that may depend on $x$. This yields, as intermediate result, a generalized conditional logit representation of choice if we start with relative Luce and use Axiom 4; similarly we obtain a generalized contextual logit representation if we start with standardized Luce and use Axiom 5.

**Definition 6.** The choice profile Pr is generalized conditional or contextual logit if there exist $\lambda_u \in \mathbb{R}$ and $w_u : X \to \mathbb{R}$ for all $u \in \mathcal{U}$ such that for all choice tasks $(u,B) \in \mathcal{D}$ and

all options $x \in B$, $\Pr(x|u,B) = V(x|u)/\sum_{x' \in B} V(x'|u)$ with

$$V(x|u) = \exp\left\{\lambda_u \cdot u_x + w_u(x)\right\}, \qquad \text{(Generalized conditional logit)}$$
$$V(x|u) = \exp\left\{\tfrac{\lambda_u \cdot u_x}{\sup u - \inf u} + w_u(x)\right\}. \qquad \text{(Generalized contextual logit)}$$

Thus, log-propensities are linear in utility, which is the main characteristic of logit models, but choice may exhibit presentation effects ($w_u(x) \neq$ const in $x$) and context effects ($\lambda_u \neq$ const in $u$). Thus, random behavior is still contained as special case. By narrow bracketing, choice propensities can be represented such that $\lambda_u = \lambda_{u+r}$ and $w_u = w_{u+r}$ for all $r \in \mathbb{R}$. Now fix any $r < \sup u - \inf u$ and any $x, y, x', y'$ such that $u_x = u_{x'} + r$ and $u_y = u_{y'} + r$. By weak utility relevance, using $\lambda_u = \lambda_{u+r}$ and $w_u = w_{u+r}$,

$$\frac{\Pr(x|u,\{x,y\})}{\Pr(y|u,\{x,y\})} = \frac{\Pr(x'|u+r,\{x',y'\})}{\Pr(y'|u+r,\{x',y'\})} \quad \Rightarrow \quad \frac{\exp\{\lambda_u \cdot u_x + w_u(x)\}}{\exp\{\lambda_u \cdot u_y + w_u(y)\}} = \frac{\exp\{\lambda_u \cdot u_x + w_u(x')\}}{\exp\{\lambda_u \cdot u_y + w_u(y')\}}$$

we obtain $w_u(x) = w_u(x') \cdot c(r)$ and $w_u(y) = w_u(y') \cdot c(r)$ for some function $c : \mathbb{R} \to \mathbb{R}$. Applying this idea for all $x, y \in X$ and all $r < \sup u - \inf u$, we find that $c(r)$ cancels out, implying $w_u(x) =$ const in $x$ and thus cancels out. Now, presentation effects and random behavior are ruled out. It is then straightforward to rule out context effects using Axiom 4 in the case of conditional logit and Axiom 5 in the case of contextual logit.

# 4 Discussion

## 4.1 The axioms

**Independence of irrelevant alternatives** IIA had been introduced to analyses of stochastic choice by Luce (1959) and was criticized immediately (Debreu, 1960). Inspired by Debreu's red-bus/blue-bus example, logit has been generalized in many studies to reflect similarity effects, see for example nested logit (McFadden, 1976) and cross-nested logit (Vovsha, 1997; Wen and Koppelman, 2001). Such generalizations are routinely used for example in transportation research. In turn, models relaxing IIA are hardly used in industrial economics and virtually never in experimental analyses. The reasons appear to be that in demand estimation, similarity effects are not required to capture product differentiation (Nevo, 2000), though applications of nested logit in this context exist (Anderson and de Palma, 1992). Economic experiments generally avoid redundant options to enable clean inference (Davis and Holt, 1993), which limits similarity effects and thus models relaxing IIA are not considered necessary (there does not appear to be a single published paper using e.g. nested logit). Thus, IIA seems to be a reasonable assumption in applications relating to utility and demand estimation, but as all models derived here are random utility models (see below), generalizations such as nested logit are straightforward.

**Cardinality and narrow bracketing** Standard representation theorems for rational choice imply that utility is defined up to affine transformation, but representation theorems for stochastic choice do not explicitly imply this property (see e.g. Dagsvik, 2008, 2015). To further discuss cardinality, let me distinguish whether (1) an analyst can infer utility only up to affine transformation and (2) choice predictions are robust to affine transformation. The former appears to be the conventional interpretation of cardinality, while Axiom 3 requires the latter. Under rational choice, these two interpretations are equivalent, but under stochastic choice, the former does not imply the latter. When inferring utility from choice using logit, for example, $\lambda$ is unknown and a free parameter, which implies that affine transformations are indistinguishable by the analyst. Logit's predictions are robust only to utility translation, as $\lambda$ then is fixed. Thus, from a normative perspective, logit satisfies robustness to affine transformations at least in inference, and explicitly requiring the cardinality axiom is not theoretically indicated.

From a positive perspective, however, the cardinality axiom appears to be adequate. On the one hand, experimental work generally finds that after controlling for individual heterogeneity due to e.g. age, education and gender, behavior in experiments is usually independent of socio-economic background variables such as income or wealth (Gächter et al., 2004; Bellemare et al., 2008, 2011). This suggests that background utility indeed factors out and Axiom 2 (narrow bracketing) is adequate. On the other hand, across studies, experimental behavior is independent of the amounts of money at stake in experiments. This is robustly reported from meta-studies on dictator games (Engel, 2011), ultimatum games (Oosterbeek et al., 2004; Cooper and Dutcher, 2011), and trust games (Johnson and Mislin, 2011). Holt and Laury (2002) find that risk aversion increases as stakes are raised, but this may equally represent an artifact of the choice model used (Wilcox, 2008). Since the utility functions used in analysis of standard experiments are homogeneous of positive degree in the payoffs,[16] scaling of payoffs induces scaling of utilities, and these results suggest that choice behavior is robust to scaling utilities. Jointly, the existing evidence therefore suggests that the cardinality axiom indeed is adequate. Since narrow bracketing is weaker than cardinality, it is of course not inadequate in turn. Relying on the weaker assumption of narrow bracketing requires a complementary stronger assumption on context independence, however.

**Strict/Weak context independence** Axioms 4 and 5 entail assumptions on context independence and presentation independence, as discussed above. First, let me focus on context independence. The assumption of strict or weak context independence complements the assumption on transformation invariance, i.e. narrow bracketing or cardinality. Context independence clarifies in which circumstances equal utilities imply equal probabilities, while narrow bracketing and cardinality clarify in which circumstances different utilities imply equal probabilities. Due to this interrelation, these axioms cannot be chosen independently. Specifically, cardinality is not compatible with strict context indepen-

[16]This is true for CRRA utilities and Prospect theoretic utilities as used in analyses of choice under risk, for CES functions used in distribution experiments, and for inequity aversion used in ultimatum games.

dence, choice simply cannot satisfy both axioms. If it satisfies the former, it violates the latter. As discussed before, empirical evidence supports the cardinality axiom.

Specifically, the previous observation that choice is invariant to utility scaling implies that the error variance adapts to the utility range, which I call weak context dependence. Weak context dependence has been observed in a large number of studies and inspired choice models with "heteroscedastic" errors, see e.g. Hey (1995) and Buschena and Zilberman (2000). Contextual logit is a heteroscedastic model that additionally allows to define the relation "more risk averse" between decision makers (Wilcox, 2011). Wilcox (2008, 2015) shows that the notion of weak context dependence fits behavior fairly accurately, and Padoa-Schioppa and Rustichini (2014) discuss neurophysiological evidence for such "adaptive coding" in the orbitofrontal cortex. The range of neurophysiological stimuli is exogenously fixed, and to use the available resources efficiently, the best-possible outcome always induces maximal neural stimulus and the worst-possible outcome induces the minimal neural stimulus. Thus, sensitivity adapts to the outcome range, as in contextual logit, inducing weak context dependence as in Axiom 5.

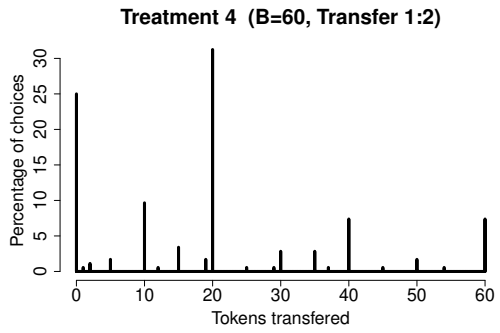
**Presentation independence** Presentation independence requires that reordering utilities is equivalent to reordering choice probabilities and implies that only option utilities are choice relevant. There exists plenty of evidence contradicting this assumption, e.g. default effects (McKenzie et al., 2006; Dinner et al., 2011; Spiegler, 2015), ordering or positioning effects (Dean, 1980; Miller and Krosnick, 1998; Feenberg et al., 2015), the left-most digit bias (Poltrock and Schwartz, 1984; Lacetera et al., 2012), and round-number effects (Heitjan and Rubin, 1991; Manski and Molinari, 2010). To illustrate the magnitude, Figure 1 provides histograms of transfers in dictator games from two experiments under mostly identical conditions. Essentially, the only difference between the experiments is the user interface: the number of tokens to be transferred is entered either manually (Figure 1a) or graphically, via mouse and slider (Figure 1b). The choice tasks are otherwise virtually equivalent, but the differences in choice patterns are drastic. The manual entry of numbers induces strong round number effects, which in turn biases parameter estimates. These observations and similar ones on default and ordering effects suggest that presentation effects are of first-order relevance, alongside context effects, which suggests that models relaxing presentation independence are critical for reliable estimation of utility. A companion paper (Breitmoser, 2016) analyzes such models.

## 4.2 Relation to random utility models

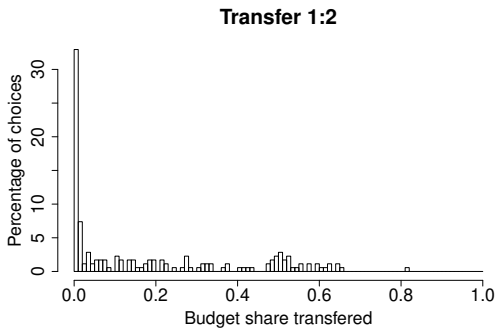
Yellott (1977) shows that if choice exhibits IIA and admits a random utility representation, then it is logit and the utility perturbations are extreme value type 1. This raised the question which behavioral assumptions are made when using the random utility representation and whether they are testable. Theorem 1 shows that in addition to IIA, narrow bracketing and strict utility relevance are equivalent to logit choice. Thus, given IIA, adopting the random utility representation is equivalent to assuming narrow bracketing

Figure 1: Dictator games where transfers are set by either manual or graphical ("slider") choice

(a) Andreoni and Miller (2002)




(b) Fisman et al. (2007)




**Note:** These treatments are representative for choice distributions in the two experiments. In these treatments, for each token given up by the dictator, two tokens are added to the recipients account.

and strict utility relevance. The random utility representation Eq. (6) clearly implies these choice properties, and given IIA, Theorem 1 shows that these assumptions also imply the random utility model, rendering it testable.

The other models studied here also have random utility representations. Contextual logit has one using the normalized utility function $\tilde{u} = u/(\sup u - \inf u)$, and for example, generalized conditional logit, has one using $\tilde{u} = u + \tilde{w}$, with $\tilde{w} = w/\lambda$, as

$$\Pr(x|u,B) = \frac{\exp\{\lambda \cdot u_x + w(x)\}}{\sum_{x' \in B} \exp\{\lambda \cdot u_{x'} + w(x')\}} = \frac{\exp\{\lambda \cdot (u_x + \tilde{w}(x))\}}{\sum_{x' \in B} \exp\{\lambda \cdot (u_{x'} + w(x'))\}}. \tag{12}$$

This implies that for all these models, relaxing IIA is straightforward—simply by assuming a generalized extreme value distribution to capture similarity effects (McFadden, 1976). For example, the corresponding "generalized nested logit" model allows for both presentation effects and similarity effects, and contextual nested logit models may capture choice based on cardinal utility functions exhibiting similarity effects.

## 4.3 The intuition underlying logit

Finally, let me discuss logit's computational intuition. To this end, let me first clarify under which assumptions choice utility is guaranteed to be an affine transformation of true utility. Generalizing Definition 1 to multiple contexts, the choice profile Pr is called "unconditional logit" if there exists a family of functions $\{v_u : X \to \mathbb{R}\}_{u \in \mathcal{U}}$ such that for all choice tasks $(u,B) \in \mathcal{D}$ and all options $x \in B$, $\Pr(x|u,B) = \exp\{v_u(x)\} / \sum_{x' \in B} \exp\{v_u(x')\}$. For later reference, define choice utility $v_u$ as follows.

**Definition 7.** Choice utility $v_u$ in context $u \in \mathcal{U}$ is $v_u(x) = \log \Pr(x|u,X)$ for all $x \in X$.

Thus, choice utility is well-defined; any other definition would be equally admissible. Now, if Pr has a conditional logit presentation in context $u$, then it also has an unconditional logit representation, and given the definition of choice utility, we know

$$\Pr(x|u,B) = \frac{\exp\{v_u(x)\}}{\sum_{x' \in B} \exp\{v_u(x')\}} = \frac{\exp\{\lambda \cdot u_x\}}{\sum_{x' \in B} \exp\{\lambda \cdot u_{x'}\}} \tag{13}$$

for all $x \in B$ and $B \in P(X)$. Hence, $v_u(x) = a + \lambda u_x$ for some $a \in \mathbb{R}$, and using $v_u(x) = \log \Pr(x|u,X)$, we obtain

$$1 = \sum_{x' \in X} \exp\{v_u(x')\} = \sum_{x' \in B} \exp\{a + \lambda \cdot u_{x'}\} \quad \Leftrightarrow \quad a = 1/\log \sum_{x' \in B} \exp\{\lambda \cdot u_{x'}\}.$$

Thus, choice utility appears to be a choice- and context-dependent affine transformation of true utility, as the additive constant $a$ depends on both $\lambda$ and $u$. This represents an obstacle to logit's computational interpretation. The next result shows that utilities may be normalized to ensure a context-independent relation of choice utility and true utility.

**Theorem 2.** *If* Pr *is conditional logit or contextual logit, then choice utility v is an affine transformation of true utility u. Specifically:*

1. *Axioms 1, 2, 4* $\Leftrightarrow$ Pr *is conditional logit* $\Leftrightarrow$ $v_u - \inf v_u = \lambda \cdot (u - \inf u)$ $\forall u \in \mathcal{U}$
2. *Axioms 1, 3, 5* $\Leftrightarrow$ Pr *is contextual logit* $\Leftrightarrow$ $v_u - \inf v_u = \lambda \cdot \frac{u - \inf u}{\sup u - \inf u}$ $\forall u \in \mathcal{U}$

*with* $\lambda$ *as obtained in the conditional/contextual logit presentation (respectively).*

The normalization of choice utility $v_u$ by subtracting the infimum reflects that $v$ is defined only up to adding arbitrary constants. Similarly, in conditional logit, utility is defined up to translation which requires subtraction of the infimum for comparability. In contextual logit, utility is defined only up to affine transformation, which requires standardization to a specific interval, here $[0,1]$, for comparability.

To discuss logit's intuition, let me begin with the known intuition underlying random behavior. Random behavior models with normal trembles implicitly determine the average choice and compute the utility parameters rationalizing this choice. This is computationally simple, but assumes that relative choice probabilities of different options are not informative with respect to their utility differences, i.e. with respect to the shape of the estimated utility function. Only the average choice is considered informative and reveals all that we may learn about parameters and shape of the utility function.

By Theorem 2, using logit, utility parameters are estimated such that choice utilities of all options are proportional to their calibrated true utilities. Since choice utilities are simply transformations of choice probabilities, logit thus assumes that relative choice probabilities are functions of utility differences, and that the shape of the probability distribution contains information about the shape of the utility function. This is supported by a number of behavioral analyses, including McKelvey and Palfrey (1998), Battalio

et al. (2001), and Weizsäcker (2003). Further, by Theorem 2, utility parameters can be estimated by regression, which illustrates the relation of choice utility and true utility transparently and avoids exposure of the analyst to logit's functional form.

The resulting utility parameters also have an independent interpretation. Let $u(x|\alpha)$ denote the utility of option $x \in X$ given parameter $\alpha \in \mathbb{R}^k$, $k \geq 1$, and consider a set of observations $O$ where all elements are observations of choices $x \in X$. Given $O$, let $(\lambda^\star, \alpha^\star)$ denote the maximum likelihood estimates to be interpreted. Define $c_\alpha = \sum_{x' \in X} \exp\{\lambda^\star \cdot u(x'|\alpha)\}$ for all $\alpha$ and based on that the normalized utility $\tilde{u}(\cdot|\alpha) = u(\cdot|\alpha) \cdot c^{\alpha^\star}/c^\alpha$. This normalization simply ensures that changing $\alpha$ does not affect the average propensity, i.e. that utility levels are comparable for all $\alpha$, or formally $\sum_{x' \in X} \exp\{\lambda^\star \cdot \tilde{u}(x'|\alpha)\}$ is constant in $\alpha$. This normalization is made without loss of generality in the sense that it does not bias the estimates,[17]

$$\arg\max_{\lambda,\alpha} \prod_{x \in O} \frac{\exp\{\lambda u(x|\alpha)\}}{\sum_{x' \in X} \exp\{\lambda u(x'|\alpha)\}} = \arg\max_{\lambda,\alpha} \prod_{x \in O} \frac{\exp\{\lambda \tilde{u}(x|\alpha)\}}{\sum_{x' \in X} \exp\{\lambda \tilde{u}(x'|\alpha)\}} = (\lambda^\star, \alpha^\star).$$

but because the utility level is now independent of $\alpha$, it allows us to interpret logit's maximum likelihood estimate of $\alpha$ given $\lambda = \lambda^\star$, which is

$$\begin{aligned} \arg\max_{\alpha} \prod_{x \in O} \frac{\exp\{\lambda \tilde{u}(x|\alpha)\}}{\sum_{x' \in X} \exp\{\lambda \tilde{u}(x'|\alpha)\}} &= \arg\max_{\alpha} \prod_{x \in O} \exp\{\lambda \tilde{u}(x|\alpha)\} \\ &= \arg\max_{\alpha} \sum_{x \in O} \lambda \tilde{u}(x|\alpha) = \arg\max_{\alpha} \sum_{x \in O} \tilde{u}(x|\alpha). \quad (14) \end{aligned}$$

That is, the logit estimate of $\alpha$ maximizes DM's total utility across choices, or in turn, logit yields the utility parameters for which DM's choices make the most sense with hindsight, portraying DM as close to utility maximization as possible. In contrast, random behavior (least squares) yields the parameters rationalizing just the average choice, considering all deviations from the mean to be plainly uninformative mistakes.

# 5 Conclusion

Multinomial logit is widely used to estimate utility and demand functions. McFadden (2001) argues that its appeal relates to its "fully consistent" axiomatic foundation linking individual characteristics (such as utilities) and choice probabilities. Yet, logit analyses are persistently criticized for making specific functional assumptions and indeed, all four existing foundations of logit require functional assumptions. The present paper resolves this critique in the sense that it provides a behavioral foundation of logit without such assumptions, building solely on invariance assumptions: independence of irrelevant alternatives and invariance to utility translation (narrow bracketing), to relabeling (presen-

[17] This obtains, as logit estimates are robust to rescaling utilities. The scaling factor $c^{\alpha^\star}/c^\alpha$ used here is a function of $\alpha$, but as $\lambda$ and $\alpha$ are independent, $\lambda$ being a free parameter comprehends this case.

tation independence), and to changing utilities of third options (context independence). These assumptions further imply the existence of a precision parameter $\lambda$ and that $\lambda$ is constant across contexts, as generally assumed in applications.

This addresses the above critique, and perhaps most notably, logit is the implication of axioms obeyed fairly widely, even in presumably robust approaches such as least squares. This suggests that these axioms may be considered consensual, rendering logit a general model of choice. In contrast, least squares additionally assumes DM perceives utility to be biased in a specific way, i.e. that DM uses a utility function that differs from the one assumed and estimated by the analyst. This is an additional, functional assumptions, implying that least squares is a more demanding, less general approach than logit.

Clarifying logit's behavioral foundation facilitates an evidence-based discussion of choice modeling. The existing behavioral evidence suggests that two of logit's assumptions are systematically violated in experiments: context independence and presentation independence. Thus, logit is less generally adequate than its relation to least squares suggests. Relaxing context independence in accordance with behavioral evidence yields contextual logit (Wilcox, 2011). Contextual logit thus promises to enable utility estimation under comparably robust assumptions, while maintaining logit's tractability. In turn, relaxing presentation independence allows to capture "focality" effects due to e.g. positioning or labeling of options, which is analyzed in detail in a companion paper (Breitmoser, 2016). The corresponding generalized logit models constitute a first generally applicable approach of capturing presentation dependence of choice. Further theoretical and behavioral analysis of presentation dependence is required to enable reliable utility estimation and to reliably predict nudging effects.

Finally, recent studies of individual choice have led to a surge of models capturing behavioral biases. These studies extend the toolkit available in applied work tremendously, allowing to account for a wide range of biases including limited consideration sets (Masatlioglu et al., 2012; Manzini and Mariotti, 2014), salience (Bordalo et al., 2012), focusing (Kőszegi and Szeidl, 2013), choice aversion (Fudenberg and Strzalecki, 2015), certainty effects (Cerreia-Vioglio et al., 2015), and satisficing (Tyson, 2008; Papi, 2012). The respective models are founded either in rational choice or as generalized Luce models, i.e. their application on data requires further assumptions on error distributions. The results in the present paper complement these studies in this respect, and are complemented by them in turn. These recent studies show how to relax say IIA in the above analysis to account for a number of prominent choice biases. In turn, the above results show that logit and contextual logit allow to extend these novel choice models in an axiomatically consistent manner, rendering them directly applicable in empirical work.

# References

Adams, E. and Messick, S. (1958). An axiomatic formulation and generalization of successive intervals scaling. *Psychometrika*, 23(4):355–368.

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

Anderson, S. and de Palma, A. (1992). Multiple product firms: a nested logit approach. *The Journal of Industrial Economics*, 40(92):261–276.

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

Andreoni, J. and Sprenger, C. (2012). Estimating time preferences from convex budgets. *American Economic Review*, 102(7):3333–3356.

Bajari, P. and Hortacsu, A. (2005). Are structural estimates of auction models reasonable? evidence from experimental data. *Journal of Political Economy*, 113(4):703–741.

Barberà, S. and Pattanaik, P. K. (1986). Falmagne and the rationalizability of stochastic choices in terms of random orderings. *Econometrica*, 54(3):707–15.

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

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

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.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890.

Birnbaum, M. H. and Navarrete, J. B. (1998). Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty*, 17(1):49–79.

Block, H. D. and Marschak, J. (1960). Random orderings and stochastic theories of responses. *Contributions to probability and statistics*, 2:97–132.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. *Quarterly Journal of Economics*, 127(3):1243–1285.

Breitmoser, Y. (2016). Stochastic choice, systematic mistakes and preference estimation. *MPRA Paper 72779*.

Buschena, D. and Zilberman, D. (2000). Generalized expected utility, heteroscedastic error, and path dependence in risky choice. *Journal of Risk and Uncertainty*, 20(1):67–88.

Camerer, C. (1989). An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty*, 2(1):61–104.

Camerer, C. and Ho, T. H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874.

Camerer, C. F. and Ho, T.-H. (2015). Behavioral game theory experiments and modeling. In Young, H. P. and Zamir, S., editors, *Handbook of Game Theory with Economic Applications*, volume 4, pages 517–573. Elsevier.

Campo, S., Guerre, E., Perrigne, I., and Vuong, Q. (2011). Semiparametric estimation of first-price auctions with risk-averse bidders. *Review of Economic Studies*, 78(1):112–147.

Caplin, A. and Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203.

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.

Cerreia-Vioglio, S., Dillenberger, D., and Ortoleva, P. (2015). Cautious expected utility and the certainty effect. *Econometrica*, 83(2):693–728.

Choi, S., Fisman, R., Gale, D., and Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, 97(5):1921–1938.

Cohen, J. L. and Dickens, W. T. (2002). A foundation for behavioral economics. *American Economic Review*, 92(2):335–338.

Cooper, D. J. and Dutcher, E. G. (2011). The dynamics of responder behavior in ultimatum games: a meta-study. *Experimental Economics*, 14(4):519–546.

Costa-Gomes, M., Crawford, V., and Broseta, B. (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235.

Dagsvik, J. K. (2008). Axiomatization of stochastic models for choice under uncertainty. *Mathematical Social Sciences*, 55(3):341–370.

Dagsvik, J. K. (2015). Stochastic models for risky choices: A comparison of different axiomatizations. *Journal of Mathematical Economics*, 60:81–88.

Davis, D. and Holt, C. (1993). *Experimental Economics*. Princeton University Press.

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.

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

Falmagne, J.-C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology*, 18(1):52–72.

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.

Fudenberg, D., Iijima, R., and Strzalecki, T. (2015). Stochastic choice and revealed perturbed utility. *Econometrica*, 83(6):2371–2409.

Fudenberg, D. and Strzalecki, T. (2015). Dynamic logit with choice aversion. *Econometrica*, 83(2):651–691.

Gächter, S., Herrmann, B., and Thöni, C. (2004). Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence. *Journal of Economic Behavior and Organization*, 55(4):505–531.

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

Goeree, J. K., Holt, C. A., and Palfrey, T. R. (2003). Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, 45(1):97–113.

Gul, F., Natenzon, P., and Pesendorfer, W. (2014). Random choice as behavioral optimization. *Econometrica*, 82(5):1873–1912.

Harless, D. and Camerer, C. (1995). An error rate analysis of experimental data testing nash refinements. *European Economic Review*, 39(3):649–660.

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. and Rutström, E. E. (2008). Risk aversion in the laboratory. In Cox, J. C. and Harrison, G. W., editors, *Risk aversion in experiments*, volume 12 of *Research in experimental economics*, pages 41–196. Emerald Group Publishing Limited.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic literature*, 48(2):356–398.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *Annals of Statistics*, pages 2244–2253.

Hey, J. (1995). Experimental investigations of errors in decision making under risk. *European Economic Review*, 39(3):633–640.

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

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655.

Jakiela, P. (2013). Equity vs. efficiency vs. self-interest: on the use of dictator games to measure distributional preferences. *Experimental Economics*, 16(2):208–221.

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889.

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

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

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.

Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. Wiley New York.

Luce, R. D. and Suppes, P. (1965). Preference, utility, and subjective probability. In Luce, R. D., Bush, R. R., and Galanter, E., editors, *Handbook of Mathematical Psychology, Vol. III*, pages 252–410. Wiley, New York.

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.

McFadden, D. (1980). Econometric models for probabilistic choice among products. *The Journal of Business*, 53(3):13–29.

McFadden, D. (2001). Economic choices. *American Economic Review*, 91(3):351–378.

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

McKelvey, R. D. and Palfrey, T. R. (1998). Quantal response equilibria for extensive form games. *Experimental Economics*, 1(1):9–41.

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.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy*, 9(4):513–548.

Oosterbeek, H., Sloof, R., and Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188.

Padoa-Schioppa, C. and Rustichini, A. (2014). Rational attention and adaptive coding: a puzzle and a solution. *American Economic Review*, 104(5):507.

Papi, M. (2012). Satisficing choice procedures. *Journal of Economic Behavior & Organization*, 84(1):451–462.

Poltrock, S. E. and Schwartz, D. R. (1984). Comparative judgments of multidigit numbers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):32.

Read, D., Loewenstein, G., and Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, 19(1-3):171–97.

Rényi, A. (1960). On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561.

Robertson, C. and Strauss, D. (1981). A characterization theorem for random utility variables. *Journal of Mathematical Psychology*, 23(2):184–189.

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

Rust, J. (2014). The limits of inference with theory: A review of wolpin (2013). *Journal of Economic Literature*, 52(3):820–850.

Rustichini, A. and Padoa-Schioppa, C. (2015). A neuro-computational model of economic decisions. *Journal of Neurophysiology*, 114(3):1382–1398.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Spiegler, R. (2015). Choice complexity and market competition. *Annual Review of Economics*.

Starmer, C. and Sugden, R. (1991). Does the random-lottery incentive system elicit true preferences? an experimental investigation. *American Economic Review*, 81(4):971–978.

Strauss, D. (1979). Some results on random utility models. *Journal of Mathematical Psychology*, 20(1):35–52.

Thurstone, L. (1927). A law of comparative judgment. *Psychological review*, 34(4):273–286.

Tyson, C. J. (2008). Cognitive constraints, contraction consistency, and the satisficing criterion. *Journal of Economic Theory*, 138(1):51–70.

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

Weizsäcker, G. (2003). Ignoring the rationality of others: evidence from experimental normal-form games. *Games and Economic Behavior*, 44(1):145–171.

Wen, C. and Koppelman, F. (2001). The generalized nested logit model. *Transportation Research Part B*, 35(7):627–641.

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.

Wilcox, N. T. (2015). Error and generalization in discrete choice under risk. *Working paper*.

Yellott, J. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144.

# Appendix

# A Proofs of Lemmas 1 and 2

**Proof of Lemma 1** Fix $u \in \mathcal{U}$ and define $u_x = u_x$ for all $x \in X$. The claimed value function $V(x|u)$ is independent of $B$, which implies that the resulting choice representation satisfies IIA (establishing $\Rightarrow$). To prove that IIA implies Luce ($\Leftarrow$), note first that $\Pr(x|u,\{x,y\})$ is in general a function of $x, y, u_x, u_y$. By positivity, it is possible to define $V(x,y,u_x,u_y) := \Pr(x|u,\{x,y\})/\Pr(y|u,\{x,y\})$, and thus by IIA (see McFadden, 1974, p. 109, for details),

$$\Pr(x|u,B) = \frac{V(x,y,u_x,u_y)}{\sum_{x' \in B} V(x',y,u_{x'},u_y)} \qquad \text{for all } x,y \in B \text{ and all } B \in P(X). \tag{15}$$

Since this holds true for all $x, y \in B$ and all $B \in P(X)$, and it does so for all $y \in X$. Hence, the odds of choosing $x$ over $x'$ are constant for any pair of benchmark options $y, y' \in X$,

$$\frac{\Pr(x|u,B)}{\Pr(x'|u,B)} = \frac{V(x,y,u_x,u_y)}{V(x',y,u_{x'},u_y)} = \frac{V(x,y',u_x,u_{y'})}{V(x',y',u_{x'},u_{y'})} \quad \text{for all } x,x',y,y' \in B \text{ and all } B \in P(X).$$

and by convexity of $X$ in $\mathbb{R}$ (richness), this can be expressed as

$$\frac{d}{dy}\frac{V(x,y,u_x,u_y)}{V(x',y,u_{x'},u_y)} = 0.$$

As a result, functions $f(y,u_y)$ and $V_1(x,u_x)$ exist such that $V(x,y,u_x,u_y) = V_1(x,u_x) \cdot f(y,u_y)$ for all $x, y \in X$, and we can write, for all $B \in P(X)$, $x \in B$ and $y \in X$,

$$\Pr(x|u,B) = \frac{V(x,y,u_x,u_y)}{\sum_{x' \in B} V(x',y,u_{x'},u_y)} = \frac{V_1(x,u_x)}{\sum_{x' \in B} V_1(x',u_{x'})}.$$

Thus, the Luce representation obtains for any $u \in \mathcal{U}$, establishing $\Leftarrow$. □

**Proof of Lemma 2** If Pr is relative Luce with $(\lambda_u, w_u) = (\lambda_{\tilde{u}}, w_{\tilde{u}})$ for all $u, \tilde{u} \in \mathcal{U}$ with $\tilde{u} = u + r$ ($r \in \mathbb{R}$), it satisfies Axioms 1 and 2, establishing $\Leftarrow$ in point 1. If Pr is standardized Luce with $(\lambda_u, w_u) = (\lambda_{\tilde{u}}, w_{\tilde{u}})$ for all affine $u, \tilde{u} \in \mathcal{U}$ satisfies Axioms 1 and 3, establishing $\Leftarrow$ in point 2. In turn, by Lemma 1, Pr satisfies IIA (if and) only if there exists $V$ such that $\Pr(x|u,B) = V(x|u)/\sum_{x' \in B} V(x'|u)$ for all $x \in B$ and all $(u,B) \in \mathcal{D}$. That is, there exists a collection of functions $(V_u)_{u \in \mathcal{U}}$ such that $\Pr(x|u,B) = V_u(x,u_x)/\sum_{x' \in B} V_u(x',u_{x'})$. Now fix $u \in \mathcal{U}$ and note that, given this representation of Pr, by both Axiom 2 and Axiom

3 we obtain

$$\frac{V_u(x,u_x)}{\sum_{x'\in B} V_u(x',u_{x'})} = \frac{V_{u+r}(x,u_x+r)}{\sum_{x'\in B} V_{u+r}(x',u_{x'}+r)} \qquad \text{for all } r\in\mathbb{R} \text{ and } (u,B)\in\mathcal{D}. \quad (16)$$

Next define the auxiliary functions $(\tilde{V}_u)_{u\in\mathcal{U}}$ such that $\tilde{V}_u(x,u_x-\inf u) = V_u(x,u_x)$ for all $x\in X$ and all $u\in\mathcal{U}$. Hence, $\Pr(x|u,B) = \tilde{V}_u(x,u_x-\inf u)/\sum_{x'\in B}\tilde{V}_u(x',u_{x'}-\inf u)$, and given Eq. (16), this implies

$$\frac{\tilde{V}_u(x,u_x-\inf u)}{\sum_{x'\in B}\tilde{V}_u(x',u_{x'}-\inf u)} = \frac{\tilde{V}_{u+r}(x,u_x+r-\inf(u+r))}{\sum_{x'\in B}\tilde{V}_{u+r}(x',u_{x'}+r-\inf(u+r))} = \frac{\tilde{V}_{u+r}(x,u_x-\inf u)}{\sum_{x'\in B}\tilde{V}_{u+r}(x',u_{x'}-\inf u)}$$

for all $r\in\mathbb{R}$, $(u,B)\in\mathcal{D}$, $x\in B$. Hence, Pr has a relative Luce representation with $\tilde{V}_u = \tilde{V}_{u+r}$ for all $u\in\mathcal{U}$ and all $r\in\mathbb{R}$, establishing $\Rightarrow$ in point 1.

Based on that, fix $u\in\mathcal{U}$ such that $\sup u - \inf u = 1$ and note that by Axiom 3, $\Pr(x|u,B) = \Pr(x|u\cdot r,B)$ for all $r>0$, i.e.

$$\Pr(x|u\cdot r,B) = \Pr(x|u,B) = \frac{\tilde{V}_u(x,u_x-\inf u)}{\sum_{x'\in B}\tilde{V}_u(x',u_{x'}-\inf u)} = \frac{\tilde{V}_u(x,\frac{ru_x-\inf ru}{\sup ru-\inf ru})}{\sum_{x'\in B}\tilde{V}_u(x',\frac{ru_{x'}-\inf ru}{\sup ru-\inf ru})}$$

for all $r>0$, $B\in P(X)$, $x\in B$; note that $\sup ru - \inf ru = r$, since $\sup u - \inf u = 1$. Hence, $\Pr(x|u\cdot r,B)$ has a standardized Luce representation with $\tilde{V}_{ru} = \tilde{V}_u$ for all $r>0$. By above, we already know $\tilde{V}_{r+u} = \tilde{V}_u$ for all $r\in\mathbb{R}$, implying $\tilde{V}_u = \tilde{V}_{a+bu}$ for all $a,b\in\mathbb{R}: b>0$ and all $u\in\mathcal{U}$, establishing $\Rightarrow$ in point 2. □

# B Proof of Theorem 1

First, let me extend the domain the utility functions to budget sets, with corresponding utility sets as value.

**Definition 8.** Pick any $u\in\mathcal{U}$ and $B\in P(X)$. Define $n := |B|$ and $b_i$, $i=1,\dots,n$, such that $B = \{b_i\}_{i=1,\dots,n}$. Then, $u(B) := \{u(b_i)\}_{i=1,\dots,n}$.

By IIA, Axiom 4 implies that for all $u,\tilde{u}\in\mathcal{U}$, all $B,\tilde{B}\in P(X)$, and all $x\in B, y\in\tilde{B}$,

$$u(B) = \tilde{u}(\tilde{B}) \quad \text{and} \quad u_x = \tilde{u}_y \qquad \Leftrightarrow \qquad \Pr(x\,|\,u,B) = \Pr(y\,|\,\tilde{u},\tilde{B}) \quad (17)$$

Correspondingly, Axiom 5 implies that (17) holds if $\sup u - \inf u = \sup u' - \inf u'$.

**Proof of Point 1, $\Rightarrow$:** By Lemma 2, Pr satisfies Axioms 1 and 2 if and only if it has a relative Luce representation. Logit satisfies Axiom 4, establishing $\Rightarrow$. □

**Proof of Point 1, $\Leftarrow$:** We have to show that, given Axioms 1 and 2, Axiom 4 implies logit.

*Step 1 (Representation independently of x):*
Pick any $u \in \mathcal{U}$ and $x, y \in X$. By Axiom 4, if $u_x = u_y$, then $\Pr(x|u,B) = \Pr(y|u,B)$ for any $B \in P(X)$ such that $x, y \in B$, and thus

$$u_x = u_y \qquad \Rightarrow \qquad V_u(x, u_x - \inf u) = V_u(y, u_x - \inf u). \tag{18}$$

Thus, choice propensities in any given context $u \in U$ solely depend on utilities. For any $u \in U$, fix an inverse $u^{-1}$ such that $u(u^{-1}(r)) = r$ for all $r$ in the image of $u$. Note that this inverse is not generally unique, but by the previous observation, the propensities $V_u(u^{-1}(u_x), u_x - \inf u)$ are independent of which inverse is chosen. Hence, we can define a function $\tilde{V}_u : \mathbb{R} \to \mathbb{R}_+$ by $\tilde{V}_u(u_x) = V_u(u^{-1}(u_x), u_x - \inf u)$, such that

$$\Pr(x|u,B) = \frac{\tilde{V}_u(u_x)}{\sum_{x' \in B} \tilde{V}_u(u_{x'})} \qquad \text{for all } x \in B, (u,B) \in \mathcal{D}, \tag{19}$$

representing propensities solely as functions of utilities $u_x$. Note that this does not rule out presentation effects; $\tilde{V}_u$ depends on context $u \in \mathcal{U}$, and the result merely states that $u_x$ contains the information required to implicitly represent presentation effects for any $u$.

*Step 2 (Generalized logit representation):*
Define $x, y \in X$ and $x', y' \in X$ such that (1) $u_y - u_x = r$, (2) $u_{y'} - u_{x'} = r$, and (3) $u_{x'} - u_x = r$, for some $r \in \mathbb{R}$. Hence, $u'_x = u_{x'}$ and $u'_y = u_{y'}$. Thus, by Axiom 4 (first equality, note that Axiom 5 actually suffices) and Axiom 2 (second equality)

$$\frac{\Pr(x'|u,\{x',y'\})}{\Pr(y'|u,\{x',y'\})} = \frac{\Pr(x|u',\{x,y\})}{\Pr(y|u',\{x,y\})} = \frac{\Pr(x|u,\{x,y\})}{\Pr(y|u,\{x,y\})}. \tag{20}$$

Using the representation from Eq. (19), for all $r < (\sup u - \inf u)/2$ and all $B \in P(X)$,

$$\frac{\tilde{V}_u(u_x)}{\sum_{x' \in B} \tilde{V}_u(u_{x'})} = \frac{\tilde{V}_u(u_x + r)}{\sum_{x' \in B} \tilde{V}_u(u_{x'} + r)} \qquad \text{for all } X \in B \text{ and } (u,B) \in \mathcal{D}. \tag{21}$$

Hence, $\tilde{V}_u(u_x + r) = \tilde{V}_u(u_x) \cdot h(r)$ for $r \approx 0$ (and some function $h : \mathbb{R} \to \mathbb{R}$), implying $\tilde{V}_u(u_x + r)/\tilde{V}_u(u_x) = h(r)$, i.e. it is independent of $u_x$ and hence it is differentiable in $u_x$, hence $\log \tilde{V}_u(u_x + r) - \log \tilde{V}_u(u_x)$ is differentiable in $u_x$, and thus $\tilde{V}_u(u_x + r)$ and $\tilde{V}_u(u_x)$ are differentiable in $u_x$. Differentiating $\tilde{V}_u(u_x + r) = \tilde{V}_u(u_x) \cdot h(r)$ at $r = 0$, we obtain

$$d\tilde{V}_u(u_x)/du_x = \tilde{V}_u(u_x) \cdot h'(0) \qquad \Rightarrow \qquad \tilde{V}_u(u_x) = \exp\{\lambda \cdot u_x + c(x)\}$$

as the solution of this differential equation, for some integration constant $c(x)$. Hence, $V_u(x, u_x) = \exp\{\lambda \cdot u_x + w(x)\}$ with $w(x) := c(x)$ for all $x \in X$. As this holds separately

for all $u \in \mathcal{U}$, $V(x|u) = \exp\{\lambda_u \cdot u_x + w_u(x)\}$ obtains, i.e.

$$\Pr(x|u,B) = \frac{\exp\{\lambda_u \cdot u_x + w_u(x)\}}{\sum_{x' \in B} \exp\{\lambda_u \cdot u_{x'} + w_u(x')\}}. \tag{22}$$

Finally, by narrow bracketing, this implies that we can represent Pr using $\lambda_u = \lambda_{u+r}$ as well as $w_u = w_{u+r}$ for all $r \in \mathcal{R}$, as then

$$\Pr(x|u+r,B) = \frac{\exp\{\lambda_u \cdot (u_x + r) + w_u(x)\}}{\sum_{x' \in B} \exp\{\lambda_u \cdot (u_{x'} + r) + w_u(x')\}}. = \frac{\exp\{\lambda_u \cdot u_x + w_u(x)\}}{\sum_{x' \in B} \exp\{\lambda_u \cdot u_{x'} + w_u(x')\}} = \Pr(x|u,B).$$

*Step 3:*
Now, pick any $u \in \mathcal{U}$ and $x, y \in X$ such that $u_x = u_y$. By Axiom 4, $\Pr(x|u,B) = \Pr(y|u,B)$ for any $B \in P(X)$ such that $x, y \in B$. Given that Pr satisfies Eq. (22), we thus obtain that $u_x = u_y$ implies $w_u(x) = w_u(y)$. Hence, it is possible to represent $w_u$ alternatively as a function of $u_x$, instead of $x$, showing that the representation Eq. (22) does not violate the result of Step 1 (that propensities may be represented solely as a function of utilities).

*Step 4 (Presentation independence):*
Next, take any $u \in \mathcal{U}$, any $\tilde{u} \in \mathcal{U}$, and define $u' = a + b\,u$ ($a, b \in \mathbb{R} : b > 0$) such that $\inf u' \leq \inf \tilde{u}$ and $\sup u' > \sup \tilde{u}$; such $u' \in \mathcal{U}$ exists by richness (transformability). Define $X' \subseteq X$ such that for all $x \in X$, there is exactly one $x' \in X' : u'_x = u'_{x'}$. Define $\tilde{X}$ such that for each $x \in X$, there is exactly one $\tilde{x} \in \tilde{X} : u_x = \tilde{u}_{\tilde{x}}$.

Define the function $f : X' \to [\inf u', \sup u']$ as $f(x') = u_{x'}$ for all $x' \in X'$. Note that $f$ is a bijection and thus invertible. Extend $f$ and $f^{-1}$ to be set functions as in Definition 8. Pick any finite $\tilde{B} \subset \tilde{X}$ and define $B' = f^{-1}\big(\tilde{u}(\tilde{B})\big)$. Thus, $|B'| = |\tilde{B}|$ and $\tilde{u}(\tilde{B}) = f(B') = u'(B')$.

For any $y \in \tilde{B}$, if $x = f^{-1}\big(\tilde{u}_y\big)$, then $\tilde{u}_y = f(x) = u'_x$, and by Axiom 4,

$$\Pr(y \,|\, \tilde{u}, \tilde{B}) = \Pr(x \,|\, u', B') = \frac{\exp\{\lambda_{u'} \cdot u'_x + w_{u'}(x)\}}{\sum_{x' \in B'} \exp\{\lambda_{u'} \cdot u'_{x'} + w_{u'}(x')\}}.$$

As stated, this obtains for all $y \in \tilde{B}$ and all $\tilde{B} \subset \tilde{X}$ (with corresponding $x$ and $B'$). Using the above result that for all $x, y \in X$, $\tilde{u}_x = \tilde{u}_y$ implies $w_{\tilde{u}}(x) = w_{\tilde{u}}(y)$, we thus obtain

$$\Pr(x \,|\, \tilde{u}, B) = \frac{\exp\{\lambda_{u'} \cdot \tilde{u}_x + w_{u'}(f^{-1}(\tilde{u}_x))\}}{\sum_{x' \in B} \exp\{\lambda_{u'} \cdot \tilde{u}_{x'} + w_{u'}(f^{-1}(\tilde{u}_{x'}))\}}$$

for all $x \in B$ and all $B \in P(X)$. Defining $\hat{\lambda} = \lambda_{u'}$ and $\hat{w} : [\inf u', \sup u'] \to \mathbb{R}$ such that

$\hat{w}(u'_x) = w_{u'}(x)$ for all $x \in X'$, this implies

$$\Pr(x \,|\, \tilde{u}, B) = \frac{\exp\{\hat{\lambda} \cdot \tilde{u}_x + \hat{w}(\tilde{u}_x)\}}{\sum_{x' \in B} \exp\{\hat{\lambda} \cdot \tilde{u}_{x'} + \hat{w}(\tilde{u}_{x'})\}}. \tag{23}$$

Since this holds true for all $\tilde{u}$ such that $\inf u' \leq \inf \tilde{u}$ and $\sup u' \geq \sup \tilde{u}$, it also holds true for $\tilde{u}_\varepsilon = \tilde{u} + \varepsilon$ if $0 < \varepsilon \leq \sup u' - \sup \tilde{u}$, implying

$$\Pr(x \,|\, \tilde{u}_\varepsilon, B) = \frac{\exp\{\hat{\lambda} \cdot [\tilde{u}_x + \varepsilon] + \hat{w}(\tilde{u}_x + \varepsilon)\}}{\sum_{x' \in B} \exp\{\hat{\lambda} \cdot [\tilde{u}_{x'} + \varepsilon] + \hat{w}(\tilde{u}_{x'} + \varepsilon)\}} = \frac{\exp\{\hat{\lambda} \cdot \tilde{u}_x + \hat{w}(\tilde{u}_x + \varepsilon)\}}{\sum_{x' \in B} \exp\{\hat{\lambda} \cdot \tilde{u}_{x'} + \hat{w}(\tilde{u}_{x'} + \varepsilon)\}}.$$

By Axiom 2, $\Pr(x \,|\, \tilde{u}, B) = \Pr(x \,|\, \tilde{u}_\varepsilon, B)$, and thus there exists a function $h : \mathbb{R} \to \mathbb{R}$ such that $\hat{w}(\tilde{u}_x + \varepsilon) = \hat{w}(\tilde{u}_x) + h(\varepsilon)$, i.e. $\varepsilon$ cancels out. Hence, we can represent propensities given $\tilde{u}_\varepsilon$ equivalently as $\hat{w}(\tilde{u}_x + \varepsilon) = \hat{w}(\tilde{u}_x)$ for all $\varepsilon \leq \sup u' - \sup \tilde{u}$ and all $x \in X$. By surjectivity of $\tilde{u}$ (richness), it follows that $\hat{w}$ is constant, which implies that $w_{u'}$ and $w_{\tilde{u}}$ are constant and cancel out. Hence, for any $\tilde{u} \in \mathcal{U}$, $\Pr(x \,|\, \tilde{u}, B)$ has a logit representation with $\lambda = \lambda_{\tilde{u}} = \lambda_{u'}$.

*Step 5 (Context independence):*
Pick any two $\tilde{u}_1, \tilde{u}_2 \in \mathcal{U}$, and any $u' \in \mathcal{U}$ such that $u' = a + b\,u$ $(a, b \in \mathbb{R} : b > 0)$ such that $\inf u' \leq \inf\{\tilde{u}_1, \tilde{u}_2\}$ and $\sup u' \leq \inf\{\tilde{u}_1, \tilde{u}_2\}$. By the previous results, both $\Pr(x \,|\, \tilde{u}_1, B)$ and $\Pr(x \,|\, \tilde{u}_1, B)$ have logit representations with $\lambda_{\tilde{u}_1} = \lambda_{\tilde{u}_2} = \lambda_{u'}$, establishing Point 1, $\Leftarrow$. □

**Proof of Point 2, $\Rightarrow$:** By Lemma 2, Pr satisfies Axioms 1 and 3 if and only if it has a standardized Luce representation. Contextual logit satisfies Axiom 5, establishing $\Rightarrow$. □

**Proof of Point 2, $\Leftarrow$:** We have to show that, given Axioms 1 and 3, Axiom 5 implies contextual logit.

*Steps 1–2 (Generalized contextual logit):*
First, fix $u \in \mathcal{U}$ such that $\sup u - \inf u = 1$. Hence,

$$\Pr(x|u, B) = \frac{V_u\big(x, \frac{u_x - \inf u}{\sup u - \inf u}\big)}{\sum_{x' \in B} V_u\big(x', \frac{u_{x'} - \inf u}{\sup u - \inf u}\big)} = \frac{V_u(x, u_x - \inf u)}{\sum_{x' \in B} V_u(x', u_{x'} - \inf u)},$$

i.e. conditional on context $u$, Pr also a relative Luce representation. Thus we may follow the arguments in the proof of Point 1 ($\Leftarrow$), up to Eq. (22), and obtain

$$\Pr(x|u, B) = \frac{\exp\{\lambda_u \cdot u_x + w_u(x)\}}{\sum_{x' \in B} \exp\{\lambda_u \cdot u_{x'} + w_u(x')\}} = \frac{\exp\big\{\frac{\lambda_u \cdot u_x}{\sup u - \inf u} + w_u(x)\big\}}{\sum_{x' \in B} \exp\big\{\frac{\lambda_u \cdot u_{x'}}{\sup u - \inf u} + w_u(x')\big\}},$$

with $\lambda_{u+r} = \lambda_u$ and $w_{u+r} = w_u$ for all $r \in \mathbb{R}$. By Axiom 3, $\Pr(x|u,B) = \Pr(x|u \cdot r,B)$ for all $r > 0$, i.e.

$$\Pr(x|u\cdot r,B) = \Pr(x|u,B) = \frac{\exp\{\lambda_u \cdot u_x + w_u(x)\}}{\sum_{x'\in B}\exp\{\lambda_u \cdot u_{x'} + w_u(x')\}} = \frac{\exp\left\{\frac{\lambda_u \cdot ru_x}{\sup ru - \inf ru} + w_u(x)\right\}}{\sum_{x'\in B}\exp\left\{\frac{\lambda_u \cdot ru_{x'}}{\sup ru - \inf ru} + w_u(x')\right\}}$$

for all $r > 0$, $B \in P(X)$, $x \in B$; note that $\sup ru - \inf ru = r$, since $\sup u - \inf u = 1$. Hence, using $u' = ru$,

$$\Pr(x|u',B) = \frac{\exp\left\{\frac{\lambda_{u'} \cdot u'_x}{\sup u' - \inf u'} + w_{u'}(x)\right\}}{\sum_{x'\in B}\exp\left\{\frac{\lambda_{u'} \cdot u'_{x'}}{\sup u' - \inf u'} + w_{u'}(x')\right\}},$$

with $w_{u'} = w_u$ and $\lambda_{u'} = \lambda_u$. By above, we already know $w_{r+u} = w_u$ and $\lambda_{r+u} = \lambda_u$ for all $r \in \mathbb{R}$, implying $\lambda_u = \lambda_{a+bu}$ and $w_u = w_{a+bu}$ for all $a,b \in \mathbb{R} : b > 0$ and all $u \in \mathcal{U}$.

*Step 3:* Next, pick any $u \in \mathcal{U}$ and any $x,y \in X$ such that $u_x = u_y$. By Axiom 5, this implies $w_u(x) = w_u(y)$, i.e. $u_x = u_y$ implies $w_u(x) = w_u(y)$.

*Step 4 (Presentation independence):*
Now, pick any $u',\tilde{u} \in \mathcal{U}$ such that $\inf u' = \inf \tilde{u} = 0$ and $\sup u' = \sup \tilde{u} = 1$. Note that $\sup u' - \inf u' = \sup \tilde{u} - \inf \tilde{u} = 1$ initially allows me to drop the normalization by $\sup u - \inf u$ in the choice propensities. Given this restriction of the images of $u'$ and $\tilde{u}$, Axiom 5 implies, simply following the proof above, up to Eq. (23),

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

for all $x \in B$ and all $B \in P(X)$, with $\hat{\lambda} = \lambda_{u'} = \lambda_{u'}/(\sup u' - \inf u')$ and $\hat{w} : [\inf u', \sup u'] \to \mathbb{R}$ such that $\hat{w}(u'_x) = w_{u'}(x)$ for all $x \in X'$. Again, define $\tilde{u}_\varepsilon = \tilde{u} + \varepsilon$, with $\varepsilon > 0$. Noting that the image of $\tilde{u}_\varepsilon$ is not contained in the image of $u'$, Axiom 5 applies only to options $x : \tilde{u}_\varepsilon(x) \leq 1$, but given this restriction, the arguments made in the proof of above, following Eq. (23) imply

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

for all $x \in B$ and all $B \in P(X)$ such that $\max \tilde{u}_\varepsilon(B) \leq 1$. By Axiom 3, $\Pr(x\,|\,\tilde{u},B) = \Pr(x\,|\,\tilde{u}_\varepsilon,B)$, which similarly to above implies $\hat{w}(\tilde{u}_x+\varepsilon) = \hat{w}(\tilde{u}_x)$, now only for all $x \in X : u_x + \varepsilon \leq 1$, but for all $\varepsilon \in (0,1)$, including all $\varepsilon \approx 0$. Hence, $\hat{w}$ is constant, implying that $w_{u'}$ and $w_{\tilde{u}}$ are constant and that given $u'$ or $\tilde{u}$, Pr has a contextual logit representation with $\lambda = \lambda_{\tilde{u}} = \lambda_{u'}$, recalling that $\sup u' - \inf u' = 1$ and $\sup \tilde{u} - \inf \tilde{u} = 1$.

*Step 5 (Weak context independence):* Finally, pick any two $u_1, u_2 \in \mathcal{U}$. Define $u' = (u_1 - \inf u_1)/(\sup u_1 - \inf u_1)$ and $\tilde{u} = (u_2 - \inf u_2)/(\sup u_2 - \inf u_2)$. By step 2, $\lambda_{u_1} = \lambda_{u'}$ and $w_{u_1} = w_{u'}$ as well as $\lambda_{u_2} = \lambda_{\tilde{u}}$ and $w_{u_2} = w_{\tilde{u}}$. By step 4, $\lambda_{u'} = \lambda_{\tilde{u}}$ and $w_{u'} = w_{\tilde{u}} = const$, and by transitivity, $\lambda_{u_1} = \lambda_{u_2}$ and $w_{u_1} = w_{u_2} = const$, implying the latter cancel out and that given $u_1$ or $u_2$, Pr has a contextual logit representation with the $\lambda_{u_1} = \lambda_{u_2} = \lambda$. Since this obtains for all $u_1, u_2 \in \mathcal{U}$, Point 2, $\Leftarrow$ is established. □

# C Proof of Theorem 2

**Proof of Point 1, $\Rightarrow$:** If Pr is conditional logit, then it also has an unconditional logit representation, and we know by the definition of choice utility $v_u$ that, for all $u \in \mathcal{U}$ and all $x \in X$,

$$\begin{aligned} & \Pr(x|u,X) = \frac{\exp\{v_u(x)\}}{\sum_{x' \in X} \exp\{v_u(x')\}} = \frac{\exp\{\lambda u_x\}}{\sum_{x' \in X} \exp\{\lambda u_{x'}\}}, \\ \Leftrightarrow \qquad & \Pr(x|u,X) = \frac{\exp\{v_u(x) - \inf v_u\}}{\sum_{x'} \exp\{v_u(x') - \inf v_u\}} = \frac{\exp\{\lambda(u_x - \inf u)\}}{\sum_{x'} \exp\{\lambda(u_{x'} - \inf u)\}} \end{aligned}$$

Now define a sequence $(x_\varepsilon)$ such that $\lim_{\varepsilon \to 0} v_u(x_\varepsilon) = \inf v_u$, which implies $\lim_{\varepsilon \to 0} u(x_\varepsilon) = \inf u$ as $v_u = \lambda u + r$ for some $r \in R$, and by positivity

$$\lim_{\varepsilon \to 0} \frac{\Pr(x_\varepsilon|u,X)}{\Pr(x|u,X)} = \frac{\exp\{0\}}{\exp\{v_u(x) - \inf v_u\}} = \frac{\exp\{\lambda \cdot 0\}}{\exp\{\lambda(u_x - \inf u)\}}$$

for all $x \in X$. Hence, $v_u(x) - \inf v_u = \lambda(u_x - \inf u)$ with $\lambda > 0$ by richness (choice variation) for all $x \in X$ and $u \in \mathcal{U}$. □

**Proof of Point 1, $\Leftarrow$:** Fix $u \in \mathcal{U}$. If point 3 holds true, then $v_u = a + \lambda u$ with $a = \inf v_u - \lambda \inf u$, and by the definition of unconditional logit,

$$\Pr(x|u,B) = \frac{\exp\{v_u(x)\}}{\sum_{x' \in B} \exp\{v_u(x')\}} = \frac{\exp\{a + \lambda u_x\}}{\sum_{x' \in B} \exp\{a + \lambda u_{x'}\}} = \frac{\exp\{\lambda u_x\}}{\sum_{x' \in B} \exp\{\lambda u_{x'}\}}$$

for all $(u,B) \in \mathcal{D}$ and all $x \in B$, i.e. Pr is has a conditional logit representation for $\lambda$. □

**Proof of Point 2, $\Rightarrow$:** If Pr is contextual logit, then it also has an unconditional logit representation, and we know by the definition of choice utility $v_u$ that, for all $u \in \mathcal{U}$ and

all $x \in X$,

$$\Pr(x|u,X) = \frac{\exp\{v_u(x)\}}{\sum_{x'} \exp\{v_u(x')\}} = \frac{\exp\left\{\lambda \cdot \frac{u_x}{\sup u - \inf u}\right\}}{\sum_{x'} \exp\left\{\lambda \cdot \frac{u_{x'}}{\sup u - \inf u}\right\}}$$

$$\Leftrightarrow \qquad \Pr(x|u,X) = \frac{\exp\{v_u(x) - \inf v_u\}}{\sum_{x'} \exp\{v_u(x') - \inf v_u\}} = \frac{\exp\left\{\lambda \cdot \frac{u_x - \inf u}{\sup u - \inf u}\right\}}{\sum_{x'} \exp\left\{\lambda \cdot \frac{u_{x'} - \inf u}{\sup u - \inf u}\right\}}$$

Now define a sequence $(x_\varepsilon)$ such that $\lim_{\varepsilon \to 0} v_u(x_\varepsilon) = \inf v_u$, which implies $\lim_{\varepsilon \to 0} u(x_\varepsilon) = \inf u$ as $v_u = a\,u + r$, with $a = \lambda/(\sup u - \inf u) > 0$ by richness (choice variation) and some $r \in R$, and by positivity

$$\lim_{\varepsilon \to 0} \frac{\Pr(x_\varepsilon|u,X)}{\Pr(x|u,X)} = \frac{\exp\{0\}}{\exp\{v_u(x) - \inf v_u\}} = \frac{\exp\left\{\lambda \cdot 0\right\}}{\exp\left\{\lambda \cdot \frac{u_x - \inf u}{\sup u - \inf u}\right\}}$$

for all $x \in X$. Hence, $v_u(x) - \inf v_u = \lambda \cdot \frac{u_x - \inf u}{\sup u - \inf u}$ for all $x \in X$ and $u \in \mathcal{U}$. □

**Proof of Point 2, $\Leftarrow$:** Fix $u \in \mathcal{U}$. If point 3 holds true, then $v_u = a + \lambda \cdot \frac{u}{\sup u - \inf u}$ with $a = \inf v_u - \lambda \cdot \frac{\inf u}{\sup u - \inf u}$, and by the definition of unconditional logit,

$$\Pr(x|u,B) = \frac{\exp\{v_u(x)\}}{\sum_{x' \in B} \exp\{v_u(x')\}} = \frac{\exp\{a + \lambda \cdot \frac{u_x}{\sup u - \inf u}\}}{\sum_{x' \in B} \exp\{a + \lambda \cdot \frac{u_{x'}}{\sup u - \inf u}\}} = \frac{\exp\{\lambda \cdot \frac{u_x}{\sup u - \inf u}\}}{\sum_{x' \in B} \exp\{\lambda \cdot \frac{u_{x'}}{\sup u - \inf u}\}}$$

for all $(u,B) \in \mathcal{D}$ and all $x \in B$, i.e. Pr is has a contextual logit representation for $\lambda$. □