The Axiomatic Foundation of Logit

Yves Breitmoser (HU Berlin)

Discussion Paper No. 78

March 7, 2018

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

The axiomatic foundation of logit

Yves Breitmoser*
Humboldt University Berlin

March 7, 2018

Abstract

Multinomial logit is the canonical model of discrete choice but widely criticized for requiring functional form 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, to relabeling options (presentation independence), and to changing utilities of third options (context independence). Reviewing behavioral evidence, presentation and context independence seem to be violated in typical experiments, though not IIA and translation invariance. 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

*I thank Friedel Bolle, Nick Netzer, Martin Pollrich, Sebastian Schweighofer-Kodritsch, Felix Weinhardt, Georg Weizsäcker and audiences in Heidelberg, 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 and CRC TRR 190) 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 work typically rests on preference assumptions as part of their model primitives. The resulting 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, implying that we cannot directly infer preferences from stated choice.[1] Structural modeling attempts to control for stochastic mistakes in choice, but proponents of non-structural approaches argue that structural modeling is impossible without making specific functional form assumptions to fix the distribution of noise, which 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 founded without specific functional form 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 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 form 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 establishes a behavioral foundation of multinomial logit[3] solely relying on invariance assumptions about choice, thus showing that even specific models of stochastic choice may be founded without functional form assumptions. This addresses the above critique and allows me to discuss logit and related models at a more

[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 α 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 ε and (ε_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).

fundamental level: the assumptions underlying logit 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 analysis attempts to put the subjective discussion of choice modeling in applied work on a more solid basis, by clarifying the assumptions implicitly made, and to provide an axiomatic foundation to enable generalizations of logit that better align with observed behavior.

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, invariance to relabeling (presentation independence), and invariance to changing utilities of third options (context independence). IIA implies that choice probabilities are functions of propensities, translation invariance implies that log-propensities are linear in utilities (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. Combined, they yield multinomial logit. Presentation independence and context independence are routinely violated in typical economic experiments, while IIA and translation invariance seem to be compatible with behavior. 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 neuro-physiologically (Padoa-Schioppa and Rustichini, 2014; Rustichini and Padoa-Schioppa, 2015). To reflect this evidence, I weaken the assumption of context independence and show that this yields, in conjunction with a cardinality axiom, the contextual logit model of 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 very briefly in Section 3 and extensively in Breitmoser (2017).

Section 2 reviews the four existing foundations of logit, showing that all of them require specific functional form 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.

2 Existing foundations of logit

This section briefly reviews the existing foundations of logit to clarify how they rest on functional form assumptions linking unobserved utility and observed choice probabilities. The next section will rebuild logit based on a weaker invariance assumption (translation invariance). 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, and DM's choice exhibits stochastic noise with unknown distribution. The utility of option x is denoted as u_x. 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 $\Pr = \{\Delta(B)\}_{B \in P(X)}$.

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 λ as it is used below can be skipped without loss of generality. Since v post-hoc rationalizes DM's choice, I refer to it as DM's *choice utility*, thus distinguishing it from the true utility u. For example, $v(x) := \log \Pr(x|X)$ is adequate. Note that v is the choice utility specifically in relation to logit's functional form and it is defined only up to translation (addition of arbitrary constants).

Luce (1959) showed that Pr has an unconditional logit representation if and only if Pr satisfies positivity and independence of irrelevant alternatives (IIA), the latter requiring

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

for all $B, B' \in P(X)$. As a result, 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 representation 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, given positivity, 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 equally 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 unconditional logit uninformative. As logit represents only one of many possible specifications of g, unconditional logit obtains only if we make a specific functional form assumption ($g = \exp$).

2.2 Conditional logit

DM's choice profile is conditional logit if the logit representation obtains for the true utility function u (McFadden, 1974).[4] To define the conditional logit model, we therefore need to extend the notation of choice probability by conditioning on u: 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, choice utility is an affine transformation of true utility u and logit analyses allow us to infer DM's utility.

In his derivation, McFadden (1974) first demonstrates that positivity and IIA imply that DM's choice 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 $y \in X$. That is, given IIA, all choice

[4] 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 are well-defined once we know the choice utility of any option x in a binary choice problem against an arbitrary benchmark option y. Specifically, 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, but it does not impose a substantial restriction on how v relates to u. Eq. (4) is compatible with many families of stochastic choice models, including strong utility, strict utility, and random behavior (including least squares),[5] 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.

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

This is McFadden's functional form assumption relating observed choice probabilities, via Eq. (4), to unobserved utilities. Given the exponential formulation of choice utility, the benchmark utility u_y now 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 $v(x) = u_x$. Indeed, using v's definition Eq. (4), McFadden's Axiom 3 is equivalent to assuming

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

noting that $\Pr(x|\cdot)+\Pr(y|\cdot) = 1$. The last equation is the definition of binomial logit (omitting λ), 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 form 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.

[5]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).

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.[6] 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. 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 these model families, there are countless functional forms of incorporating perturbations. Indeed, given IIA, assuming that the choice probabilities have a random utility representation is equivalent to assuming that binomial choice is logit (see 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 the specific random utility model defined above, although it is less obvious.

2.4 Foundation in rational inattention

Matejka and McKay (2015) model choice assuming 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

[6] 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.

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.[7]

Matejka and McKay show that DM's choice then has 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 $w_x =$ 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 functional form 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.

3 The axiomatic foundation of logit

3.1 Formal framework

As above, X denotes the set of options and decision maker DM has to choose an option $x \in B$ from a finite set $B \subseteq X$ given utility $u : X \to \mathbb{R}$. The probability that DM chooses $x \in B$ is denoted as $\Pr(x|u, B)$.

The pair (u, B) is called *choice task* and from the perspective of the analyst, every choice task is defined by a *context*, which defines the utility function u, and an *option set B*. Since the context uniquely defines the utility function, I will treat these terms synonymously when it does not cause confusion, speaking of the context u when adopting the perspective of the analyst and of the utility u when adopting the perspective of DM. Let me provide an example.

Example 1 (Choice under risk)**.** *There are four prizes,* $(\pi_1, \pi_2, \pi_3, \pi_4)$ *and each option is*

[7] 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.

a lottery $L = (\pi_i, p, \pi_j)$ yielding π_i with probability p and π_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 prize profiles $(\pi_1, \pi_2, \pi_3, \pi_4)$ represent different contexts and induce different utility mappings $u : X \to \mathbb{R}$. In conjunction with the option set $B \subset X$, any prize profile (or, context) defines a specific choice task (u, B).

The set of choice tasks available to the analyst is $\mathcal{D} = \mathcal{U} \times P(X)$. $\mathcal{U}$ denotes the set of utility functions $u : X \to \mathbb{R}$ underlying DM's choices in the contexts that may be constructed by the analyst, 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. I assume that the set of choice tasks $\mathcal{D}$ is "rich" in the following sense.[8]

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

1. *Transformability:* $a + bu \in \mathcal{U}$ 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 exist (u, B) and $x, x' \in B$ such that $\Pr(x|u, B) \neq \Pr(x'|u, B)$.

Transformability ensures that we may analyze affine transformations of utility functions in the first place, ensuring that all affine transformations are well-defined objects. Convexity and surjectivity 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 richness is required for uniqueness and satisfied in choice tasks typically of interest to experimentalists (such as choice under risk, using standard utility functions). Note that the utility functions may still be fairly ill-behaved, violating smoothness or even continuity for any number points. Finally, "Choice variation" rules out the trivial case that choice probabilities are uniform in all choice tasks. In addition, I assume positivity.

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

[8] With slight abuse of notation, I identify all real numbers as constant functions such that addition and multiplication of a function with a real are well-defined. Thus, for any $u : X \to \mathbb{R}$ and any $a, b \in \mathbb{R}$, $u' = a + bu$ is equivalent to $u'_x = a + bu_x$ for all $x \in X$. Further, 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 $\inf u = \inf_{x \in X} u_x$.

Positivity allows that DM fails to maximize utility, however rarely, and captures the widely documented phenomenon that individual choice fluctuates and involves dominated options. This has been observed in many different contexts, including choice under risk (Birnbaum and Navarrete, 1998), small normal-form games (Costa-Gomes et al., 2001), and dictator games (Andreoni and Miller, 2002; Fisman et al., 2007), to name just a few. 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.[9] 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

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

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

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

Gul et al. (2014) show that if choice probabilities are countably additive, IIA obtains if DM's (stochastic) preference ordering is complete. As discussed above, IIA implies that 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). 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 exploiting our 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 be context dependent, as IIA itself does not restrict choice across contexts u. Even the functional forms of V_u may vary across contexts, and

[9] 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).

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 if the environment is rich.

IIA is compatible with a wide range of choice models. As a running example, consider the following family of choice models satisfying IIA with choice propensities being functions of x and u_x. Below, I illustrate how subsequent axioms refine this family.

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

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 Translation invariance and cardinality

Invariance to translation of utilities 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, DM approaches any single choice task independently of background utility and previous tasks, which is generally assumed in behavioral analyses.

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

The range of behavioral analyses supporting the assumption of translation invariance is reviewed below. At this point, let me simply note that it is an assumption relating observed choice probabilities and unobserved utilities, but contrary to the related functional form assumptions made in the existing foundations reviewed above, it is a mere invariance assumption which is testable and has been tested extensively.

Invariance of choice to scaling utilities is also 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,[10] but observed choice probabilities are largely unaffected by such scaling. In conjunction with translation invariance, this implies cardinality of utility.

[10]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.

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

Translation invariance 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,[11] requiring robustness to affine transformation is of course not prohibitive. Amongst others, all models satisfying

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

for any function $f : \mathbb{R}_+ \to \mathbb{R}_+$ satisfy 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, translation invariance implies what I call a "relative Luce" representation of choice and cardinality implies a "standardized Luce" representation.

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

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

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

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 translation invariance nor cardinality are very restrictive. To illustrate, the family of representations compatible with IIA and cardinality include

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

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

[11] Recall the definition in Footnote 5 or see Luce and Suppes (1965).

$g_u(y) = \phi(y/\sigma)$ with ϕ as standard normal density, and for arbitrary combinations of say strict utility and random behavior. Thus, neither IIA nor cardinality (or translation invariance) seem to substantially restrict how choice propensities relate to utility u.

This impression is somewhat misleading. If choice is consistent across contexts, in a sense to be made precise, then translation invariance 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 translation invariance or cardinality. Thus, on their own, translation invariance 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 translation invariance, 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 $u_{y'} = 8$. translation invariance 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 unrealistic, and for this reason, I introduce a 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 or 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 Axiom 5 may be inadequate, but before I enter this discussion, let me state 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* ⇔ Pr *satisfies Axioms 1, 2, 4*
2. Pr *is contextual logit* ⇔ 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 translation invariance, 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 translation invariance, 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.

Thus, log-propensities are linear in utility, which is the main characteristic of logit models, but choice may exhibit presentation effects ($w_x \neq$ const) and context effects (both λ and w may depend on context u). To formally capture the context effects, let $\lambda_u \in \mathbb{R}$ and $w_u : X \to \mathbb{R}$ be the respective terms in context u. By translation invariance, 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 to 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 Behavioral evidence for the axioms

Cardinality and translation invariance Behavioral evidence seems to support both scale invariance and translation invariance surprisingly clearly, suggesting the cardinality axiom is adequate. The most direct evidence comes from a comparably unusual source, neuro-economics, and lies in the so-called “adaptive coding”: The neuronal representation of subjective values (“utilities”) adapts to the range of values in the context of the choice. Specifically, the baseline activity of the cell encoding the value of a given object generally represents the minimum of the utility range in a given context, and its peak activity adapts the maximum of the utility range. Such adaptation to choice environments is efficient considering the physical limitations in neuronal firing rates and builds on a wealth of evidence starting with Tremblay and Schultz (1999) and Padoa-Schioppa (2009), which is reviewed in detail in Louie and De Martino (2014) and Camerer et al. (2017). As a result, if the subjective values of options (i.e. their utilities) are translated by some constant $a \in \mathbb{R}$ or scaled by some constant $b \in \mathbb{R}$, the nervous cells adapt to these manipulations and factor them out as described in the cardinality axiom.

There is additional evidence from choice data. 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 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 thus supports (translation invariance) is adequate. Read et al. (1999) refer to this observation as narrow bracketing and provide further evidence. 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,[12] scaling of payoffs induces scaling of utilities, and these results suggest that choice behavior also is robust to scaling utilities.

Jointly, the existing evidence therefore suggests that the cardinality axiom indeed is adequate. Since translation invariance is weaker than cardinality, it is of course not inadequate in turn. Relying on the weaker assumption of translation invariance requires a complementary stronger assumption on context independence, however.

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 differen-

[12]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.

tiation (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.

Strict/Weak context independence Context independence clarifies in which circumstances equal utilities imply equal probabilities, while translation invariance and cardinality clarify in which circumstances different utilities imply equal probabilities. Due to this interrelation, these axioms are not independent. Specifically, cardinality is not compatible with strict context independence—if choice satisfies the former, it violates the latter. As discussed before, empirical evidence supports the cardinality axiom and thus leads us to reject strict context independence. Specifically, the previous observation that choice is invariant to utility scaling implies that the "error variance" in choice adapts to the utility range, which was labeled weak context independence above. Aside from the neuro-economic and choice evidence on scale invariance discussed above, weak context independence 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 independence fits behavior impressively across contexts.

Presentation independence There exists plenty of evidence suggesting that presentation independence in the strict form assumed above is inadequate. Most directly, reordering options does affect choice probabilities (Dean, 1980; Miller and Krosnick, 1998; Feenberg et al., 2017). Similarly, relabeling options also affects choice probabilities, as suggested by 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). Such presentation effects are substantial in magnitude and analyzed in detail in Breitmoser (2017).

4.2 Interpreting logit

Parameter estimates obtained using the so-called random behavior models have a straightforward interpretation, most obviously those obtained by least squares: loosely speaking, one determines the average choice and computes the utility parameters rationalizing this choice. This is computationally simple and transparent, which could be taken as suggestion that the random behavior approach is natural and adequate. In contrast, estimates obtained from random utility models such as logit do not seem to lend themselves to a simple interpretation, suggesting that logit modeling is a bit of a black box. The following seeks to correct that impression by establishing a simple interpretation.

As defined above, 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, let us define the choice utility v_u in this definition as follows.

Definition 6. Choice utility v_u in context $u \in \mathcal{U}$ is $v_u(x) = \log \Pr(x|u,X)$ for all $x \in X$.

Any translation of v_u will be equally admissible for our purpose. 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

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

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 is an affine transformation of true utility, and the additive constant a depends on both λ and u (i.e. on DM's precision and the context). The next result establishes that choice utilities can be normalized, without using knowledge about u, to guarantee that they are simply a scaled transformation of (normalized) true utilities.

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 λ as obtained in the conditional/contextual logit presentation (respectively).

Theorem 2 provides a first interpretation: Using logit, utility parameters are estimated such that the observed choice utilities of all options are proportional to their calibrated true utilities. The normalization of choice utility v_u by subtracting its infimum reflects that v is defined only up to adding arbitrary constants. In contextual logit, utility is defined only up to affine transformation, which requires standardization to a specific interval, here $[0,1]$, for comparability. This interpretation highlights that logit captures the empirical observation that the shape of the choice distribution contains information about the shape of the utility function (see e.g. McKelvey and Palfrey, 1998, Battalio et al., 2001, and Weizsäcker, 2003). This stands in contrast to, and implicitly falsifies, random behavior models where only the average choice is considered informative.

Given the above results, we can provide a second simple interpretation. Let $u(x|\alpha)$ denote the utility of option $x \in X$ given parameter α, 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 logit's maximum likelihood estimates to be interpreted. Define $c_\alpha = \sum_{x' \in X} \exp\{\lambda^\star \cdot u(x'|\alpha)\}$ for all α and based on that the normalized utility $\tilde{u}(\cdot|\alpha) = u(\cdot|\alpha) \cdot c^{\alpha^\star}/c^{\alpha}$. Clearly, utility parameters are not affected by normalizing utility in this way,

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

but the normalization ensures that changing α does not affect the average log-propensity. As a result, the "utility levels" are invariant to changing α in the sense that $\sum_{x' \in X} \exp\{\lambda^\star \cdot \tilde{u}(x'|\alpha)\}$ is constant in α. Thus, logit's maximum likelihood estimate of α, taking $\lambda = \lambda^\star$ as given, satisfies

$$\begin{aligned} \arg\max_{\alpha} \prod_{x \in O} \frac{\exp\{\lambda \tilde{u}(x|\alpha)\}}{\sum_{x' \in X} \exp\{\lambda \tilde{u}(x'|\alpha)\}} &= \arg\max_{\alpha} \prod_{x \in O} \exp\{\lambda \tilde{u}(x|\alpha)\} \\ &= \arg\max_{\alpha} \sum_{x \in O} \lambda \tilde{u}(x|\alpha) = \arg\max_{\alpha} \sum_{x \in O} \tilde{u}(x|\alpha). \end{aligned} \tag{13}$$

That is, the logit estimate of α maximizes DM's total utility across choices, or in turn, logit yields the utility parameters for which DM's choices make the most sense in hindsight, portraying DM as close to utility maximization as possible.

5 Conclusion

Multinomial logit is widely used to capture stochastic choice when estimating 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 functional form assumptions and indeed, all existing foundations of logit require functional form assumptions. The present paper resolves this critique by providing a behavioral foundation of logit without relying on functional form assumptions, building solely on invariance assumptions: independence of irrelevant alternatives and invariance to utility translation, to relabeling (presentation independence), and to changing utilities of third options (context independence). Our representation result further establishes the existence of a precision parameter λ and that λ is constant across contexts, as generally assumed in applications.

In addition to addressing the critique that structural modeling requires functional form assumptions, establishing a behavioral foundation avoiding functional form assumptions also facilitates an evidence-based discussion of choice modeling. For, assumptions about the functional form linking unobserved utilities and observed choice probabilities, as in the existing foundations of logit, are (currently) not testable, while assumptions about scale and translation invariance are testable and have been tested. Perhaps most notably, replacing the functional form assumptions by translation invariance reveals that two assumptions were implicitly made and these assumptions tend to be violated in experimental studies: context independence and presentation independence. Relaxing context independence yields the contextual logit model of 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 presentation effects due to e.g. positioning or labeling of options. This is analyzed in detail in Breitmoser (2017).

References

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

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.

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

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.

Breitmoser, Y. (2017). Discrete choice with representation effects. CRC TRR 190 Working Paper.

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., Cohen, J., Fehr, E., Glimcher, P., and Laibson, D. (2017). *Neuroeconomics*, volume 2, pages 153–217. Princeton University Press, editors: john kagel and alvin roth edition.

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.

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.

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.

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., Ganguli, I., Gaule, P., and Gruber, J. (2017). ItâĂŹs good to be first: Order bias in reading and citing nber working papers. *Review of Economics and Statistics*, 99(1):32–39.

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

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 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.

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.

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.

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.

Louie, K. and De Martino, B. (2014). The neurobiology of context-dependent valuation and choice. In *Neuroeconomics (Second Edition)*, pages 455–476. Elsevier.

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.

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. (1998). Quantal response equilibria for extensive form games. *Experimental Economics*, 1(1):9–41.

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. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience*, 29(44):14004–14014.

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

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

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.

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.

Tremblay, L. and Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708.

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),

$$\Pr(x|u,B) = \frac{V(x,y,u_x,u_y)}{\sum_{x' \in B} V(x',y,u_{x'},u_y)} \qquad \text{for all } x,y \in B \text{ and all } B \in P(X). \tag{14}$$

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 $\Leftarrow$. $\square$

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

$$\frac{V_u\big(x,u_x\big)}{\sum_{x'\in B}V_u\big(x',u_{x'}\big)}=\frac{V_{u+r}\big(x,u_x+r\big)}{\sum_{x'\in B}V_{u+r}\big(x',u_{x'}+r\big)}\qquad\text{for all } r\in\mathbb{R}\text{ and }(u,B)\in\mathcal{D}.\tag{15}$$

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. (15), this implies

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

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.

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

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 7. 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 $x\in B,y\in\tilde{B}$,

$$u(B)=\tilde{u}(\tilde{B})\quad\text{and}\quad u_x=\tilde{u}_y\qquad\Leftrightarrow\qquad\Pr(x\,|\,u,B)=\Pr(y\,|\,\tilde{u},\tilde{B})\tag{16}$$

Correspondingly, Axiom 5 implies that (16) 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

$$u_x = u_y \qquad \Rightarrow \qquad V_u(x, u_x - \inf u) = V_u(y, u_x - \inf u). \tag{17}$$

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

$$\Pr(x|u,B) = \frac{\tilde{V}_u(u_x)}{\sum_{x' \in B} \tilde{V}_u(u_{x'})} \qquad \text{for all } x \in B, (u,B) \in \mathcal{D}, \tag{18}$$

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)

$$\frac{\Pr(x'|u,\{x',y'\})}{\Pr(y'|u,\{x',y'\})} = \frac{\Pr(x|u',\{x,y\})}{\Pr(y|u',\{x,y\})} = \frac{\Pr(x|u,\{x,y\})}{\Pr(y|u,\{x,y\})}. \tag{19}$$

Using the representation from Eq. (18), for all $r < (\sup u - \inf u)/2$ and all $B \in P(X)$,

$$\frac{\tilde{V}_u(u_x)}{\sum_{x' \in B} \tilde{V}_u(u_{x'})} = \frac{\tilde{V}_u(u_x + r)}{\sum_{x' \in B} \tilde{V}_u(u_{x'} + r)} \qquad \text{for all } X \in B \text{ and } (u,B) \in \mathcal{D}. \tag{20}$$

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.

$$\Pr(x|u,B) = \frac{\exp\{\lambda_u \cdot u_x + w_u(x)\}}{\sum_{x' \in B} \exp\{\lambda_u \cdot u_{x'} + w_u(x')\}}. \tag{21}$$

Finally, by translation invariance, 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. (21), 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. (21) 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 $u' = a + bu$ $(a, b \in \mathbb{R} : b > 0)$ 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 7. 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

$$\Pr(x \,|\, \tilde{u}, B) = \frac{\exp\{\hat{\lambda} \cdot \tilde{u}_x + \hat{w}(\tilde{u}_x)\}}{\sum_{x' \in B} \exp\{\hat{\lambda} \cdot \tilde{u}_{x'} + \hat{w}(\tilde{u}_{x'})\}}. \tag{22}$$

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. ε 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, $\Leftarrow$. $\square$

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 $\Rightarrow$. $\square$

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,

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

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. (21), and obtain

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

with $\lambda_{u+r} = \lambda_u$ 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.

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

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. (22),

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

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. (22) 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 $x \in X$,

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

Now define a sequence (x_ε) 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 λ. □

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_ε) 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 $u \in \mathcal{U}$. $\square$

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 λ. $\square$