The seeming unreliability of rank-ordered data as a consequence of model misspecification

Yan, Jin and Yoo, Hong Il

May 2014

The seeming unreliability of rank-ordered data as a consequence of model misspecification

Jin Yan[a], Hong Il Yoo[b]

a. Department of Economics, The Chinese University of Hong Kong

Shatin, Hong Kong. jyan@cuhk.edu.hk

b. Durham University Business School, Durham University

Durham, United Kingdom. h.i.yoo@durham.ac.uk

May 28, 2014

Abstract

The rank-ordered logit model's coefficients often vary significantly with the depth of rankings used in the estimation process. The common interpretation of the unstable coefficients across ranks is that survey respondents state their more and less preferred alternatives in an incoherent manner. We point out another source of the same empirical regularity: stochastic misspecification of the random utility function. An example is provided to show how the well-known symptoms of incoherent ranking behavior can result from stochastic misspecification, followed by Monte Carlo evidence. Our finding implies that the empirical regularity can be addressed by the development of robust estimation methods.

JEL classification: C25, C52, C81

Keywords: rank-ordered logit, exploded logit, ranking, qualitative response, stated preference

1 Introduction

The use of stated preference data has become commonplace in the discrete choice modeling literature.[1] The stated preference surveys provide practical means to collect data for analyzing consumer preferences for both non-market goods and potential market goods, allowing estimation of choice models when revealed preference data do not exist (Vossler *et al.*, 2012). This explains the popularity of applying stated preference data in research areas characterized by the scarcity of revealed preference data, including environmental economics, health economics and transportation economics.

In relation to multinomial choice data which record the chosen alternative in each choice set, rank-ordered data record the ranking of all available alternatives from best to worst. A stated preference survey can collect both types of data as easily, by eliciting either a choice or rank ordering over the same set of alternatives (Caparros *et al.*, 2008).[2] Econometric models for both types of data can be derived from an identical random utility maximization model, the most popular among them being the multinomial logit (MNL) model (McFadden, 1974) and the rank-ordered logit (ROL) model (Beggs *et al.*, 1981). The extra information that rank orderings provide can then be exploited to estimate the utility coefficients of interest more precisely.

A long standing issue in rank-ordered data analysis is that the estimated ROL coefficients often vary significantly with the depth of rankings incorporated in the estimation process (Chapman and Staelin, 1982). In particular, the estimates often become attenuated monotonically with successive incorporation of each worse-ranked alternative, as if the residual variance increases because respondents are less certain about their less preferred alternatives (Hausman and Ruud, 1987).

As Hanley *et al.* (2001) summarize, the common interpretation of this empirical regularity is that rank-ordered data are unreliable due to the cognitive burden of ranking several objects, which induces behavioral inconsistencies in how the respondents arrive at their better and worse alternatives. Under such interpretation, several studies have explored the implications of different survey designs for consistency in respondent behavior (Boyle *et al.*, 2001; Foster and Mourato, 2002; Caparros *et al.*, 2008; Scarpa *et al.*, 2011). Likewise, several econometric models have been developed to accommodate the relative cognitive difficulties of identifying better and worse alternatives

[1] See references in popular econometrics textbooks of Greene (2008) and Train (2009).

[2] Outside stated preference settings, rank orderings are often harder to observe than multinomial choices, but can still be observed in, for example, a recall survey (Berry *et al.*, 2004; Train and Winston, 2007) asking consumers to name both actually purchased and another closely considered products.

(Hausman and Ruud, 1987; Fok *et al.*, 2012; Yoo and Doiron, 2013), as well as more generally changing decision protocols across ranks (Ben-Akiva *et al.*, 1992).

This paper advances an alternative explanation for the instability of the ROL coefficients across ranks: stochastic misspecification of the random utility function. We present analytic examples and Monte Carlo evidence, pointing out that even a minor departure from the postulated error distribution can induce the ROL estimates to exhibit the very sort of variation which has been read as symptoms of inconsistent ranking behavior. Since the ROL model relies on the independence of irrelevant alternatives property to express a rank-ordering probability as a product of marginal choice probabilities, its susceptibility to stochastic misspecification has been suspected previously (Hausman and Ruud, 1987; Layton, 2000). But the actual consequences of stochastic misspecification, in particular that they include the empirical regularity in question, have not been explored and demonstrated to date.

In empirical applications, the true distribution of the error terms in the utility function is very rarely known. Our findings suggest that a new estimation method robust to stochastic misspecification is needed to separate the effects of stochastic misspecification from the true inconsistency in ranking behavior.

The remainder of this paper is organized as follows. Section 2 reviews the rank-ordered logit model and the issue of unstable coefficients. Section 3 presents analytic examples showing that this empirical regularity may arise from stochastic misspecification. Section 4 presents Monte Carlo evidence on the consequences of estimating ROL when stochastic misspecification is present. Section 5 concludes with recommendations for future research.

2 Unstable ROL coefficients across ranks

We use the following notations to describe the usual cross-sectional setting of rank-ordered data. Agent $n \in \{1, 2, \cdots, N\}$ faces a choice set of $J_n > 2$ alternatives. The alternatives are assumed to be labeled numerically, and for simplicity J_n is assumed to equal J for all N agents.[3] Each agent, thus, faces the choice set $\mathbf{J} = \{1, 2, \ldots, J\}$. Each agent states which M out of J alternatives she likes best, where $1 \leq M \leq J - 1$,

[3]That $J_n = J$ is also true in most of empirical applications. Our subsequent discussion, nevertheless, can be easily adapted to cases where the number of alternatives varies across the agents by making notations related to the choice set size agent-specific.

and ordinally ranks these M alternatives from best to worst without a tie.[4] We use $r_n \equiv (r_{n1}, \cdots, r_{nM})$ to denote agent n's rank ordering of the M alternatives, where $r_{nm} \in \mathbf{J}$ indexes the alternative ranked m^{th} best. Finally, the collection of her m best alternatives is denoted by $\mathbf{J}_{n,m} = \{r_{n1}, \ldots, r_{nm}\}$.

Following McFadden's (1974) random utility framework, assume that agent n obtains utility U_{nj} from alternative $j \in \mathbf{J}$

$$U_{nj} = x'_{nj}\beta + \epsilon_{nj}, \tag{1}$$

where x_{nj} is an observed K-vector that contains the characteristics of agent n and alternative j, β is a K-vector of taste coefficients, and ϵ_{nj} is the random utility part that is unobservable to econometricians.[5]

When $M = 1$, a multinomial discrete choice model with the following choice probability can be derived from the random utility maximization hypothesis

$$P(r_n, x_n, \beta) = Pr\left(U_{r_{nM}} > max_{j \in \mathbf{J} \setminus \mathbf{J}_{n,M}} U_{nj}\right), \tag{2}$$

where $x_n = (x_{n1}, \cdots, x_{nJ})$ and $\setminus$ denotes the set difference operator. When $M > 1$, the probability of observing agent n's rank ordering can be similarly derived as that of observing a preference relation

$$P(r_n, x_n, \beta) = Pr\left(U_{r_{n1}} > \cdots > U_{r_{nM}} > max_{j \in \mathbf{J} \setminus \mathbf{J}_{n,M}} U_{nj}\right). \tag{3}$$

The maximum likelihood estimation (MLE) is often applied to estimate β when the distribution of the error terms, ϵ_{nj}, is assumed to be known. When the error terms are i.i.d. type I extreme value, closed-form expressions for formulas (2) and (3) exist: they are the multinomial logit (MNL) model (McFadden, 1974) and the rank-ordered logit (ROL) model (Beggs *et al.*, 1981) respectively. The ROL probability of observing agent n's rank ordering is

$$P(r_n, x_n, \beta) = \prod_{m=1}^{M} \left[\frac{e^{x'_{nr_{nm}}\beta}}{\sum_{j \in \mathbf{J} \setminus \mathbf{J}_{n,m-1}} e^{x'_{nj}\beta}} \right], \tag{4}$$

where $\mathbf{J}_{n,0}$ is an empty set.[6]

[4] In the special case when $M = J - 1$, all J alternatives are ranked from the best to worst.
[5] All the vectors are column vectors.
[6] When $M = 1$, formula (4) is the MNL probability of observing agent n's choice.

The ROL formula (4) is a product of MNL formulas. A single observation on agent n's rank ordering is exploded into M pseudo-observations on choices, in Train's (2009, p.157) parlance. The m^{th} pseudo-observation is constructed as an independent observation on a choice among a set of alternatives excluding $\mathbf{J}_{n,m-1}$. The sample size effectively increases with M; when $M \geq 2$, β can be more precisely estimated than when each agent's best alternative in $\mathbf{J}$ is observed alone (*i.e.* $M = 1$).

Note that each agent's rank ordering can be recoded as if M had been smaller than it actually is, say integer Q such that $1 \leq Q \leq M$. The ROL formula (4) implies that, when the model is correctly specified, β can be consistently estimated via MLE using any of potential response variables detailing the top Q ranks; discarding the bottom $(M - Q)$ pseudo-observations results only in efficiency loss.[7]

Starting from Chapman and Staelin (1982), however, several empirical studies have found the sensitivity of the ROL coefficients to the depth of pseudo-observations that MLE exploits. Specifically, the estimates tend to vary significantly as Q is successively increased from 1 through M, that is as MLE incorporates information on each worse-ranked alternative incrementally. As Hausman and Ruud (1987) observed, the estimates also often become attenuated monotonically as Q is increased, as if the coefficients are normalized with respect to an increasingly larger error variance.

Over years, this empirical regularity has been interpreted as a data problem, symptomizing inconsistencies in how respondents state their more and less preferred alternatives (Hanley *et al.*, 2001). The cognitive burden of rank-ordering several alternatives has been postulated as the underlying cause of inconsistent ranking behavior, with the pattern of attenuation being taken as an indication that respondents find it easier to tell what they like better. Under such interpretation, several studies have explored the implications of different survey designs for consistency in respondent behavior (Boyle *et al.*, 2001; Foster and Mourato, 2002; Caparros *et al.*, 2008; Scarpa *et al.*, 2011), and the use of econometric models describing the cognitive process of completing a rank-ordering task (Hausman and Ruud, 1987; Ben-Akiva *et al.*, 1992; Fok *et al.*, 2012; Yoo and Doiron, 2013).

Within such a paradigm, the ROL formula (4) is viewed as a special case of a general specification

$$P(r_n,\, x_n,\, \{\alpha_m\}_{m=1}^{M}) = \prod_{m=1}^{M} \left[\frac{e^{x'_{nr_{nm}}\alpha_m}}{\sum_{j \in \mathbf{J} \setminus \mathbf{J}_{n,m-1}} e^{x'_{nj}\alpha_m}} \right], \qquad (5)$$

[7]The ROL model reduces to the MNL model when the rank orderings are recoded as choices ($Q = 1$).

which involves a distinct K-vector of coefficients α_m influencing the choice of the m^{th} best alternative. This specification, however, is incompatible with the microeconomic approach of modeling a rank ordering as a realized preference relation, encapsulated in equations (1) and (3), because no distributional assumption on ϵ_{nj} leads to equation (5). As Ben-Akiva *et al.* (1992, p.153) make explicit, a different behavioral framework is needed to conceptualize a rank ordering as a constructed sequence of choices.[8]

Below, we point out another potential source of the same empirical regularity: namely, that the error terms ϵ_{nj} are not i.i.d. extreme value. This explanation is compatible with the microeconomic approach using which Beggs *et al.* (1981) introduced the ROL model. It is also relevant to most of empirical works wherein the postulated error distribution serves only as an approximation to unknown true distributions. More importantly, it opens doors to the development of more general an econometric solution than the *ad hoc* modeling of cognitive processes through restrictions on $\{\alpha_m\}_{m=1}^M$, the validity of which are data-specific.

3 The impact of stochastic misspecification

In this section, we use an analytic example to demonstrate how stochastic misspecification can induce the ROL coefficients to be unstable across ranks. The overall impact of coefficient attenuation following incorporation of worse-ranked alternatives is a decrease in the magnitude of the systematic component of utility. To focus on this, consider a random sample wherein each agent n ranks three alternatives according to the random utility function

$$U_{nj} = \beta_j + \epsilon_{nj}, \tag{6}$$

where β_j is the systematic utility and ϵ_{nj} is the unobserved utility that agent n obtains from alternative $j \in \{1, 2, 3\}$. We can normalize $\beta_3 = 0$ because only differences in utility matter to the observed behavior (Train, 2009, Ch 2). For brevity, subscript n will be omitted except when specifying a sample log-likelihood function.

[8]Specifically, assume now that agent n constructs her response by solving M independent random utility maximization problems in sequence. The choice set at the m^{th} problem is $\mathbf{J} \setminus \mathbf{J}_{n,m-1}$, and the utility-maximizing alternative in this choice set is ranked m^{th} best in her response. Equation (5) results when agent n derives utility $U_{nj,m}$ from each alternative $j \in \mathbf{J} \setminus \mathbf{J}_{n,m-1}$:

$$U_{nj,m} = x'_{nj}\alpha_m + \epsilon_{nj,m},$$

where the error terms $\epsilon_{nj,m}$ are independent across m, and i.i.d. type I extreme value.

Suppose that ϵ_j are independent and identically distributed over j, with $F(\cdot)$ and $f(\cdot)$ as the true distribution and density functions respectively. Then, the true choice probability of alternative j is $P_j \equiv P_j(\beta_1, \beta_2)$ where $P_j(\cdot)$ is defined as

$$\begin{aligned} P_j(b_1, b_2) &= Pr(b_j + \epsilon_j > b_k + \epsilon_k \, for \, k \in \mathbf{J} \setminus \{j\}) \\ &= \int \left[\prod_{k \in \mathbf{J} \setminus \{j\}} F(b_j - b_k + \epsilon_j) \, f(\epsilon_j) \right] d\epsilon_j \end{aligned} \tag{7}$$

with $\mathbf{J} = \{1, 2, 3\}$ and $b_3 = 0$ by normalization. Assume that $P_j \in (0, 1)$ for $j \in \mathbf{J}$.

When the distribution function $F(\cdot)$ is known, MLE can be applied to consistently estimate the parameter vector $\beta \equiv (\beta_1, \beta_2)$ and the choice probabilities (P_1, P_2, P_3). In practice, the true distribution and density functions are rarely known. In most cases, MLE is operationalized by assuming that the distribution and density functions of the error terms are $G(\cdot)$ and $g(\cdot)$ respectively. Define

$$G_j(b_1, b_2) = \int \left[\prod_{k \in \mathbf{J} \setminus \{j\}} G(b_j - b_k + \epsilon_j) \, g(\epsilon_j) \right] d\epsilon_j \ . \tag{8}$$

Then, the log-likelihood function of a random sample of N agents

$$\frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{3} 1(U_{nj} > U_{nk} \, for \, k \in \mathbf{J} \setminus \{j\}) \cdot ln \left[G_j(b_1, b_2) \right]$$

converges to its probability limit

$$\begin{aligned} l(b_1, b_2) &= E\left\{ \sum_{j=1}^{3} 1(U_j > U_k \, for \, k \in \mathbf{J} \setminus \{j\}) \cdot ln \left[G_j(b_1, b_2) \right] \right\} \\ &= \sum_{j=1}^{3} P_j \cdot ln \left[G_j(b_1, b_2) \right] \end{aligned} \tag{9}$$

as N goes to infinity. We can show that there is a unique $b_o \equiv (b_{o,1}, b_{o,2}) \in \mathbb{R}^2$ that maximizes $l(b_1, b_2)$ at which

$$G_1(b_{o,1}, b_{o,2}) = P_1, \tag{10}$$

$$G_2(b_{o,1}, b_{o,2}) = P_2, \tag{11}$$

when $G(\cdot)$ is an increasing continuous distribution function. Equations (10) and (11) imply that even when the pseudo-true vector b_o, to which the ML estimator converges, is different from the true parameter vector β because the assumed distribution $G(\cdot)$ is different from the true distribution $F(\cdot)$, the choice probabilities are still consistently estimated.

The ROL model is equivalent to the MNL model when the estimation process uses information on the best alternative only. For the MNL model, $G(\cdot)$ is the type I extreme value, $EV(0,1,0)$, distribution function. The MNL estimator of β converges to $b_{MNL} \equiv (b_{MNL,1},\, b_{MNL,2})$ that solves

$$G_1(b_{MNL}) = \frac{e^{b_{MNL,1}}}{e^{b_{MNL,1}} + e^{b_{MNL,2}} + 1} = P_1, \tag{12}$$

$$G_2(b_{MNL}) = \frac{e^{b_{MNL,2}}}{e^{b_{MNL,1}} + e^{b_{MNL,2}} + 1} = P_2. \tag{13}$$

In other words,

$$b_{MNL,1} = ln\left(\frac{P_1}{1 - P_1 - P_2}\right),\ b_{MNL,2} = ln\left(\frac{P_2}{1 - P_1 - P_2}\right).$$

For a further discussion, consider the case where $(\beta_1, \beta_2) = (\frac{\pi^2}{\sqrt{2}}, \frac{\pi^2}{\sqrt{8}}) = (2.221, 1.111)$ and $\epsilon_j \sim Unif[-\frac{\pi^2}{2}, \frac{\pi^2}{2}]$.[9] We can calculate the choice probabilities $(P_1,\, P_2,\, P_3) = (\frac{64}{96},\, \frac{25}{96},\, \frac{7}{96})$ analytically, and find the probability limit of the MNL estimator as $b_{MNL} = (2.213,\, 1.273)$. When $\epsilon_j \sim N(0, \frac{\pi^2}{6})$ instead, we can compute the choice probabilities $(P_1,\, P_2,\, P_3) = (0.686, 0.247, 0.067)$ numerically and obtain $b_{MNL} = (2.326, 1.304)$.[10] In the uniform case, the MNL estimator is biased downward for β_1 and upward for β_2, while in the normal case, it is biased upward for both. This asymptotic bias does not result from re-normalizing the overall scale of utility because both $Unif[-\frac{\pi^2}{2}, \frac{\pi^2}{2}]$ and $N(0, \frac{\pi^2}{6})$ have the same variance as $EV(0,1,0)$. Despite the biased coefficients, the choice probabilities are consistently estimated by the MNL model as noted earlier.

We do not have this luck when we estimate the ROL model using information on the full rank ordering, *i.e.* best and second-best alternatives. Let $P_{jkl} = Pr(U_j > U_k > U_l)$ be the true probability of a specific rank ordering $(j,\, k,\, l)$, and $\mathbf{T}$ be the set of all possible rank orderings *i.e.* permutations of the choice set $\mathbf{J} = \{1, 2, 3\}$. Table 1 provides the true probability of each $(j,\, k,\, l) \in \mathbf{T}$ for the above two cases.

[9] Since the overall scale of utility is irrelevant to the observed behavior (Train, 2009, Ch 2), this case is observationally equivalent to the case where $\beta = (\frac{1}{2}, \frac{1}{4})$ and $\epsilon_j \sim Unif[-\frac{1}{2}, \frac{1}{2}]$, a configuration which facilitates the analytic derivation of true choice and rank-ordering probabilities.

[10] It is generally difficult to evaluate the true choice and rank-ordering probabilities given a specific distributional choice for ϵ_j. The normal case is an exception due to the availability of multinomial probit and rank-ordered probit likelihood evaluators in many software packages.

Table 1: True probabilities of rank-orderings

$\epsilon_j \sim Unif[-\frac{\pi}{\sqrt{2}}, \frac{\pi}{\sqrt{2}}]$		$\epsilon_j \sim N(0, \frac{\pi^2}{6})$	
$P_{123} = 44/96$	$P_{132} = 20/96$	$P_{123} = 0.483$	$P_{132} = 0.203$
$P_{213} = 20/96$	$P_{231} = 5/96$	$P_{213} = 0.203$	$P_{231} = 0.043$
$P_{312} = 5/96$	$P_{321} = 2/96$	$P_{312} = 0.043$	$P_{321} = 0.023$

P_{jkl} is the true probability that options j,k and l are most, second-most and least preferred respectively.

According to the ROL model, the probability of the rank ordering $(j,\, k,\, l)$, given systematic utility vector $b \equiv (b_1,\, b_2)$, is

$$G_{jkl}(b_1,\, b_2) = \frac{e^{b_j}}{e^{b_j} + e^{b_k} + e^{b_l}} \cdot \frac{e^{b_k}}{e^{b_k} + e^{b_l}}.$$

The ROL log-likelihood function of a random sample of N agents

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{(j,\,k,\,l)\in\mathbf{T}} 1(U_{nj} > U_{nk} > U_{nl}) \cdot ln\left[G_{jkl}(b_1,\, b_2)\right]$$

converges, as N approaches infinity, to its probability limit

$$\begin{aligned} l_{ROL}(b_1,\, b_2) &= E\left\{\textstyle\sum_{(j,\,k,\,l)\in\mathbf{T}} 1(U_j > U_k > U_l) \cdot ln\left[G_{jkl}(b_1,\, b_2)\right]\right\} \\ &= \textstyle\sum_{(j,\,k,\,l)\in\mathbf{T}} P_{jkl} \cdot ln\left[G_{jkl}(b_1,\, b_2)\right]. \end{aligned} \tag{14}$$

The ROL estimator of β converges to the unique pseudo-true vector that maximizes $l_{ROL}(b_1,\, b_2)$. This pseudo-true vector, $b_{ROL} \equiv (b_{ROL,1},\, b_{ROL,2})$, solves the following first-order conditions:

$$\frac{e^{b_{ROL,1}}}{e^{b_{ROL,1}} + e^{b_{ROL,2}} + 1} + P_2\frac{e^{b_{ROL,1}}}{e^{b_{ROL,1}} + 1} + P_3\frac{e^{b_{ROL,1}}}{e^{b_{ROL,1}} + e^{b_{ROL,2}}} = P_1 + P_{213} + P_{312}, \tag{15}$$

$$\frac{e^{b_{ROL,2}}}{e^{b_{ROL,1}} + e^{b_{ROL,2}} + 1} + P_1\frac{e^{b_{ROL,2}}}{e^{b_{ROL,2}} + 1} + P_3\frac{e^{b_{ROL,2}}}{e^{b_{ROL,1}} + e^{b_{ROL,2}}} = P_2 + P_{123} + P_{321}. \tag{16}$$

It is difficult to obtain the closed-form solution for equations (15) and (16). But we can numerically solve for $b_{ROL} = (1.845,\, 0.882)$ in the uniform case, and $b_{ROL} = (1.974,\, 0.952)$ in the normal case. In comparison with their MNL counterparts $b_{MNL} = (2.213,\, 1.273)$ in the uniform case and $b_{MNL} = (2.326, 1.304)$ in the normal case, the

ROL estimator's probability limit b_{ROL} has a smaller magnitude in each argument. When we estimate the ROL model with a random sample, we will observe that the coefficient estimates vary with the depth of rankings because the probability limit of those estimates *per se* varies with whether information on only the best (b_{MNL}) or the best and second-best alternatives (b_{ROL}) are used in the estimation process.[11]

The choice probabilities are no longer consistently estimated by the ROL model in this example, as summarized in Table 2.

Table 2: ROL vs true choice probabilities

$\epsilon_j \sim Unif[-\frac{\pi}{\sqrt{2}}, \frac{\pi}{\sqrt{2}}]$	$\epsilon_j \sim N(0, \frac{\pi^2}{6})$
$G_1(b_{ROL}) = 0.649 < 0.667 = P_1$	$G_1(b_{ROL}) = 0.667 < 0.686 = P_1$
$G_2(b_{ROL}) = 0.248 < 0.260 = P_2$	$G_2(b_{ROL}) = 0.240 < 0.247 = P_2$
$G_3(b_{ROL}) = 0.103 > 0.073 = P_3$	$G_3(b_{ROL}) = 0.093 > 0.067 = P_3$

P_j is the true probability that option j is the most preferred option. $G_j(b_{ROL})$ is the corresponding asymptotic ROL prediction, obtained by evaluating the MNL formulas in (12) and (13) at b_{ROL}.

The Shannon entropy of the ROL predictions are 0.86 in the uniform case and 0.83 in the normal case, larger than their true counterparts 0.81 and 0.79. In other words, ROL squeezes the three choice probabilities closer to one another because of the attenuation in the estimated systematic utility.

By comparing equations (12) and (13) that determine the pseudo-true vector for the MNL estimator with equations (15) and (16) that determine the pseudo-true vector for the ROL estimator, we can clearly see why the MNL does a better job in estimating the choice probabilities under stochastic misspecification. Equations (12) and (13) imply that b_{MNL} is determined in a way such that the MNL choices probabilities match with the true choice probabilities.[12] In comparison, b_{ROL} is determined to match other kinds

[11]Our presentation focuses on the issue of attenuation but our example also can shed light on why some studies (*e.g.* Ben-Akiva *et al.*, 1992; Layton, 2000) find that alternative-specific constants are either magnified or attenuated while other coefficients are attenuated across ranks. Suppose that alternative 2 is used as the base alternative to estimate ($\alpha_1 = \beta_1 - \beta_2$, $\alpha_3 = -\beta_2$). Then, in the uniform case, the probability limit of the MNL estimator is $(0.940, -1.273)$ and that of the ROL estimator is $(0.963, -0.882)$: the estimator of α_1 is magnified while that of α_2 is attenuated.

[12]In our 3-alternative example involving two alternative-specific intercepts, the MNL choice probabilities exactly equal to the true choice probabilities. In more complicated cases, *e.g.* involving generic coefficients on continuous attributes, the MNL choice probabilities are usually different from the true choice probabilities. However, the pseudo-true vector is the one that makes the the MNL choice probabilities as close as possible to the true choice probabilities. This explains why, in some empirical studies, the MNL model is a good approximation for other parametric models such as the multinomial

of probabilities. The left-hand side of equation (15) is the probability that alternative 1 is ranked as a top-two (*i.e.* either best or second-best) alternative under the assumed extreme value distribution of the error terms, while the right-hand side of equation (15) is the true probability that alternative 1 is a top-two alternative.[13] Therefore, the pseudo-true vector b_{ROL} is determined in a way such that the probability of observing each alternative as a top-two alternative under the assumed distribution is as close as possible to its counterpart under the true distribution.

From equation (15), we can see that $b_{ROL} = b_{MNL}$ if

$$P_2 \frac{e^{b_{ROL,1}}}{e^{b_{ROL,1}} + 1} = P_{213}, \tag{17}$$

and

$$P_3 \frac{e^{b_{ROL,1}}}{e^{b_{ROL,1}} + e^{b_{ROL,2}}} = P_{312}, \tag{18}$$

which requires the independence of irrelevant alternatives (IIA) property. The IIA property only holds when the error terms are i.i.d. type I extreme value or can be normalized as so (Anderson *et al.*, 1992, Ch 2). Except in such special cases, the sensitivity of the estimated coefficients to the depth of rankings used in the estimation process needs not symptomize data unreliability resulting from incoherent ranking behavior.

4 Evidence from simulation experiments

In this section, we provide simulated examples to further illustrate that the instability of ROL coefficients across ranks needs not symptomize a data problem, because it can arise when the ROL model is estimated in the presence of stochastic misspecification. We have conducted 3 sets of simulation experiments on the finite sample behavior of the ROL estimates under different configurations of the systematic component of utility and choice set size. Experiment A uses the identical configuration as our analytic examples involving 3 alternatives and 2 identified intercepts. Experiment B considers 5 alternatives and follows a more typical setup for a simulation study by incorporating generic attributes. Experiment C applies synthetic data generated by combining actual stated preference data with simulated errors.

The common setup for all 3 experiments can be summarized as follows, using the

probit model (MNP): see for example Dow and Endersby (2004).

[13] Similar analysis applies for equation (16).

same notations as in Section 2. We generate random samples of N agents who rank J alternatives according to the random utility function in equation (1). The ranking behavior of each agent is consistent with random utility maximization as described by Anderson *et al.* (1992): given the realizations of the random component of utility or error terms, ϵ_{nj}, each agent has a deterministic preference relation over all alternatives which enables her to rank them from best to worst in an unambiguous and consistent manner. Her rank ordering, r_n, coincides with her realized preference relation. Because her rank ordering is not constructed as a sequence of choices, the issue of whether she chooses the $(q+1)^{\text{th}}$ best alternative in the same manner as the q^{th} best (Ben-Akiva *et al.*, 1992), and that of whether she indeed chooses the q^{th} and $(q+1)^{\text{th}}$ in sequence (Scarpa *et al.*, 2011) are irrelevant in our simulated data.

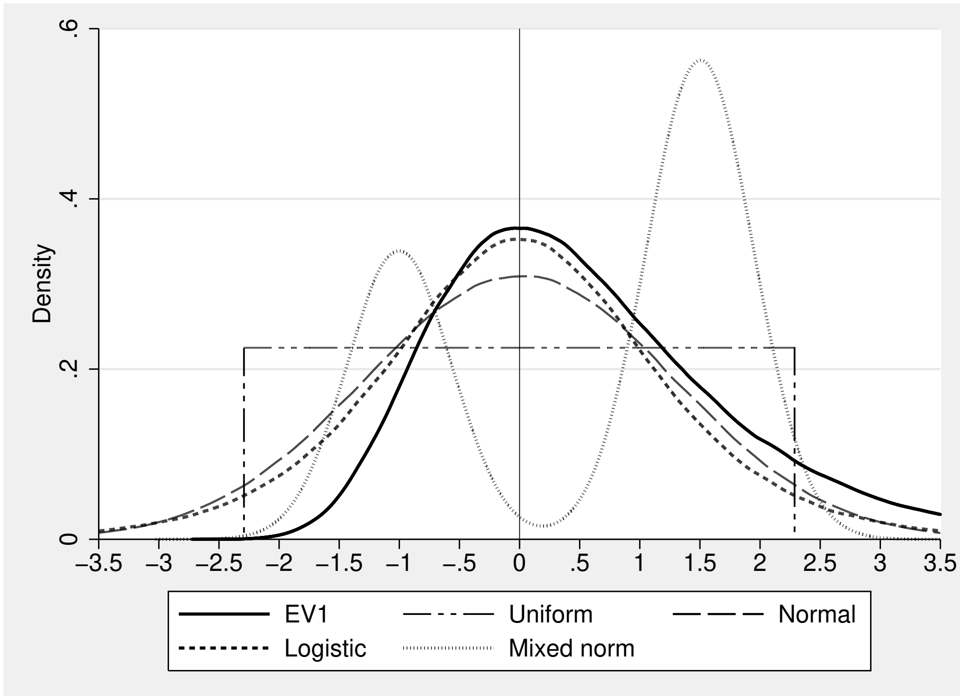
When the error terms are i.i.d. type I extreme value, $EV(0,1,0)$, the correctly specified model is ROL. In practice, the error terms may depart from this distribution in several aspects, for example due to heteroscedasticity and non-zero correlations, the extent of which is likely to be context-specific.[14] To emphasize the generality of our findings, we have selected for presentation cases where the true error terms are i.i.d. but not necessarily as $EV(0,1,0)$. Specifically, each experiment simulates the error terms ϵ_{nj} as i.i.d. random variables with the variance of $\pi^2/6$ from one of the following 5 distributions in turn:[15]

- $EV(0,1,0)$, to obtain benchmark results in the absence of misspecification.
- $Unif[-\frac{\pi}{\sqrt{2}}, \frac{\pi}{\sqrt{2}}]$ and $N(0, \frac{\pi^2}{6})$: Our analytic examples have employed these uniform and normal distributions.
- $Logistic(0, \frac{1}{\sqrt{2}})$: The logistic distribution is closely related to the extreme value distribution in that differencing two independent $EV(0,1,0)$ random variables results in a $Logistic(0,1)$ random variable.
- $0.369 \cdot N(-1, 0.184) + 0.631 \cdot N(1.5, 0.193)$: Fox (2007, p.1014) has constructed this mixed normal distribution to compare the performance of the MNL estimator

[14]For instance, an omitted attribute shared by some alternatives can induce ϵ_{nj} to be positively correlated over them (Train, 2009, Ch 4) and omitted random heterogeneity in β can induce ϵ_{nj} to be heteroscedastic and correlated over j (Train, 2009, Ch 6).

[15]While the $EV(0,1,0)$ and mixed normal distributions have the Euler's constant ($\approx$0.5775) as their mean whereas the other distributions have zero mean, this difference is inconsequential. Because only differences in utility matter for the observed behavior (Train, 2009, Ch 2), drawing the error terms from a distribution with non-zero mean, say $N(0.5775, \pi^2/6)$, is equivalent to drawing them from the same distribution with zero mean, say $N(0, \pi^2/6)$.

Figure 1: Density functions used in simulating error terms



EV1 errors are drawn from $EV(0,1,0)$; Uniform from $Unif[-\frac{\pi}{\sqrt{2}}, \frac{\pi}{\sqrt{2}}]$; Normal from $N(0, \frac{\pi^2}{6})$; Logistic from $Logistic(0, \frac{1}{\sqrt{2}})$; and Mixed norm(al) from $0.369 \cdot N(-1, 0.184) + 0.631 \cdot N(1.5, 0.193)$.

and his semi-parametric procedure.

Figure 1 plots the corresponding density functions.

In every experiment, 10,000 random samples of N agents are generated using each distribution. For each sample, we obtain three sets of estimation results.

First, we estimate ROL $J-1$ times, each time using the top Q ranks of observed responses where integer Q varies from 1 through $J-1$. In each estimation, the likelihood of agent n's response is specified as in equation (4), replacing M with an appropriate Q. Recall that with $Q=1$, this is the MNL likelihood of her most preferred alternative and with $Q=J-1$, the ROL likelihood of her full rank ordering. We use $b^Q = (b_1^Q, \cdots, b_K^Q)$ to denote the resulting estimates of β, and report their means and empirical standard deviations (SD) over the 10,000 repetitions.

Second, we conduct the likelihood ratio tests of consistency in ranking behavior as in Chapman and Staelin (1982) and Ben-Akiva *et al.* (1992). These tests are motivated by viewing an observed rank ordering as a sequence of choices. They are performed at each Q by comparing the sample likelihood of the restricted model in equation (4) with that of the unrestricted model in (5), again replacing M with an appropriate Q. Rejection of $H_0 : \alpha_m = \beta$ for all $m = 1, 2, \cdots, Q$ is taken as evidence that the most and less preferred alternatives are not chosen in a coherent manner, and the 2^{nd} through Q^{th} ranks of the observed responses are unreliable for use in estimation of preferences (Hanley *et al.*, 2001). We conduct each test LR_Q at the nominal size of 5%, and report the empirical rejection frequencies in the 10,000 repetitions.

Finally, we estimate the HROL model of Hausman and Ruud (1987). The likelihood of agent n's response is now specified as a special case of equation (5) with $M = J - 1$, $\alpha_1 = \beta$ and $\alpha_m = \sigma_m \beta$ where β and scalars σ_m for $2 \leq m \leq J - 1$ are parameters to be estimated. The fact that σ_m is often decreasing in m has been interpreted as evidence that respondents are less certain about less preferred alternatives. This interpretation has inspired modern modeling approaches that generalize the use of such scale parameters to capture the respondent's ranking capabilities (Scarpa *et al.*, 2011; Fok *et al.*, 2012; Yoo and Doiron, 2013). We report the mean and empirical standard deviation of each σ_m's estimates, denoted $\widehat{\sigma}_m$, over the 10,000 repetitions.

We now turn to a more specific discussion of each experiment. In Experiment A, agent n ranks $J = 3$ alternatives in order of utility U_{nj}:

$$\begin{gathered} U_{nj} = \beta_j + \epsilon_{nj} \\ n = 1, 2, \cdots, N \quad j = 1, 2, 3 \end{gathered} \tag{19}$$

where $\beta_1 = \frac{\pi}{\sqrt{2}} = 2.221, \beta_2 = \frac{\pi}{\sqrt{8}} = 1.111$ and β_3 is normalized to 0 as in our analytic example. We repeat this experiment for 3 different sample sizes: $N = 100, 300, 500$.

Table 3 summarizes the selected estimation results. When ϵ_{nj} are drawn from EV(0,1,0), the results are as expected because ROL is the correctly specified model. Recall that vectors $b^1 = (b^1_1,\, b^1_2)$ and $b^2 = (b^2_1,\, b^2_2)$ exploit the first rank and full (here, best and second-best) ranks, respectively. Efficiency gains from using all ranks are evident. Arguments of b^2 are much less dispersed than those of b^1 under the same sample size configuration. The empirical size of the test of consistency in ranking behavior, LR_2, is always close to its nominal size of 5%, with the largest deviation of 5.8% occurring when $N = 100$. The mean scale estimate $\widehat{\sigma}_2$ from HROL is almost 1, indicating no attenuation.

When ROL is a misspecified model, efficiency gains still remain but the rest of the results change dramatically. Consider the cases of ϵ_{nj} drawn from the uniform and normal distributions, which have been analyzed in Section 3. In both cases, subject to sampling error which is smaller in larger samples, the mean of b^1 and b^2 are in line with their probability limits (*i.e.* b_{MNL} and b_{ROL}), mimicking the empirical regularity of coefficient attenuation across ranks. The LR_2 test rejects the null much more frequently, already in 20% of 10,000 simulated samples when $N = 100$; the frequency increases substantially to 55% and 78% when $N = 300$ and 500. These are false rejections if the null is interpreted as consistency in ranking behavior, and correct rejections if the null is interpreted as independence of irrelevant alternatives (IIA) or equivalently the assumption that ϵ_{nj} are i.i.d $EV(0, 1, 0)$. Our results suggest that this type of test could be problematic as a test of consistency in ranking behavior in empirical works when ROL is employed as an approximation to an unknown true model. It appears more appropriate when viewed as a test of IIA, showing more desirable size and power properties than the tests of IIA which can be implemented using multinomial choice data.[16] The mean $\widehat{\sigma}_2$ lies below 0.7, seemingly suggestive of increased noise in the second rank. In summary, b^1 and b^2 appear misleadingly consistent with the interpretation that our simulated agents have chosen their best and second-best alternatives in sequence, and experienced greater cognitive difficulties in identifying the latter.

The preceding qualitative conclusions remain unchanged when ϵ_{nj} are drawn from the logistic and the mixed normal distributions. The logistic case yields very similar results as the previous two cases. The mixed normal case, presumably because the density of this distribution deviates arguably the most from the $EV(0, 1, 0)$ density, yields more quantitatively striking results. The average difference between b^1 and b^2, both in level and as a proportion of b^1, becomes much larger, as also indicated by much smaller the mean of $\widehat{\sigma}_2$ (now, below 0.48) than in the three preceding cases. In consequence, the LR_2 test rejects the null much more frequently, already 64% of times when $N = 100$ and more than 97% of times in larger samples.

Experiment B incorporates a larger number of alternatives, $J = 5$, and considers an environment where even MNL choice probabilities are inconsistent in the presence of stochastic misspecification.[17] Agent n ranks $J = 5$ alternatives in order of utility U_{nj}:

$$U_{nj} = \beta_1 x_{nj1} + \beta_2 x_{nj2} + \beta_3 x_{nj3} + \epsilon_{nj} \qquad (20)$$
$$n = 1, 2, \cdots, N \quad j = 1, 2, 3, 4, 5$$

[16] Fry and Harris (1996) investigate the finite sample properties of these tests in detail.

[17] We use the same systematic component of utility as what Lee (1995, p.397) has specified to compare the performance of the MNL estimator and his semi-parametric procedures.

Table 3: Experiment A on hypothetical DGP in equation (19)

$\beta = (\beta_1 = 2.221, \beta_2 = 1.111)$

N=100	b^1		b^2			
	b_1	b_2	b_1	b_2	LR_2	$\widehat{\sigma}_2$
EV1	2.281	1.157	2.245	1.126	5.80%	1.060
	(0.422)	(0.458)	(0.261)	(0.226)		(0.351)
Uniform	2.283	1.330	1.866	0.892	21.71%	0.680
	(0.431)	(0.459)	(0.241)	(0.213)		(0.251)
Normal	2.396	1.360	1.996	0.963	19.88%	0.692
	(0.449)	(0.479)	(0.251)	(0.214)		(0.239)
Logistic	2.497	1.395	2.104	1.024	20.87%	0.696
	(0.467)	(0.497)	(0.261)	(0.215)		(0.226)
Mixed norm	2.478	1.410	1.837	0.881	63.84%	0.477
	(0.465)	(0.495)	(0.266)	(0.194)		(0.186)

N=300	b^1		b^2			
	b_1	b_2	b_1	b_2	LR_2	$\widehat{\sigma}_2$
EV1	2.239	1.125	2.228	1.114	5.35%	1.017
	(0.231)	(0.252)	(0.149)	(0.132)		(0.179)
Uniform	2.236	1.289	1.854	0.885	55.36%	0.657
	(0.234)	(0.249)	(0.138)	(0.120)		(0.133)
Normal	2.353	1.327	1.982	0.959	54.48%	0.666
	(0.241)	(0.260)	(0.143)	(0.121)		(0.123)
Logistic	2.446	1.353	2.088	1.014	54.65%	0.674
	(0.247)	(0.266)	(0.147)	(0.122)		(0.122)
Mixed norm	2.426	1.367	1.825	0.876	97.57%	0.463
	(0.253)	(0.272)	(0.150)	(0.111)		(0.101)

N=500	b^1		b^2			
	b_1	b_2	b_1	b_2	LR_2	$\widehat{\sigma}_2$
EV1	2.233	1.122	2.226	1.115	5.00%	1.009
	(0.175)	(0.191)	(0.114)	(0.099)		(0.136)
Uniform	2.226	1.282	1.849	0.883	78.63%	0.651
	(0.178)	(0.190)	(0.108)	(0.092)		(0.100)
Normal	2.341	1.314	1.981	0.955	77.67%	0.665
	(0.183)	(0.197)	(0.108)	(0.095)		(0.095)
Logistic	2.432	1.340	2.085	1.012	77.38%	0.674
	(0.191)	(0.206)	(0.115)	(0.096)		(0.093)
Mixed norm	2.417	1.358	1.824	0.875	99.93%	0.461
	(0.192)	(0.206)	(0.117)	(0.086)		(0.077)

Each row summarizes the results over 10,000 random samples of N agents generated from the specified density; we use the same abbreviation for each density as defined in notes to Figure 1. $b^Q = (b_1, b_2)$ are the ROL estimates of true coefficients $\beta = (\beta_1, \beta_2)$ using the first Q ranks of each agent's response; we report their mean and empirical standard deviation (in parentheses). LR_Q is the likelihood ratio test for consistency in ranking behavior across the first Q ranks; we report its rejection frequency at the nominal size of 5%. $\widehat{\sigma}_Q$ is the scale for the Qth best pseudo-choice, estimated by HROL that uses all available ranks; we report its mean and empirical standard deviation (in parentheses).

where $(\beta_1, \beta_2, \beta_3) = (1, -1, 1)$. Each observed attribute x_{nji} is obtained by drawing from an i.i.d. random variable. x_{nj1} is generated as a uniform random variable with support on $[-1, 1]$; x_{nj2} as a Poisson random variable with mean 2 and truncated with support on $[0, 5]$; and finally x_{nj3} as a truncated standard normal random variable with support on $[-1.8, 1.8]$. We also repeat this experiment for 3 sample sizes: $N = 100, 300, 500$.

Table 4 summarizes the selected estimation results of Experiment B. When ϵ_{nj} are drawn from $EV(0, 1, 0)$, all ROL coefficient estimates b^1, b^2, b^3 and b^4 closely resemble the true parameter values on average. Efficiency gains from exploiting deeper ranks are also evident. For example, using all ranks (b^4) in $N = 100$ yields similarly dispersed estimates as using the top-rank (b^1) in $N = 300$. All tests (LR_2, LR_3,LR_4) of consistency in ranking behavior across the subscripted ranks (or more appropriately, that of IIA in light of our earlier discussion) have sizes close to the nominal 5% level. The HROL scale estimates $\widehat{\sigma}_2$, $\widehat{\sigma}_3$ and $\widehat{\sigma}_4$ are close to 1 on average as expected.

When ϵ_{nj} is drawn from other distributions, the results except efficiency gains change again. Since the present experiment involves more than 3 alternatives, we can now examine how the coefficient estimates change as the ROL estimator exploit deeper ranks successively: the mean estimates b^Q continue to decline in magnitude as we increase Q. All LR_Q tests reject the null much more often than 5%, once again calling into question their use as tests for consistency in ranking behavior. Even when only the top-two ranks (LR_2) are used in estimation, the null is rejected from 9% to 38% of samples when $N = 100$, and from 20% to 92% of samples when $N = 300$. These rejection frequencies increase further as deeper ranks are incorporated into estimation. Similarly, $\widehat{\sigma}_Q$ also decreases in the depth of ranking Q, resulting in $1 > \widehat{\sigma}_2 > \widehat{\sigma}_3 > \widehat{\sigma}_4$ on average. Without knowing the true DGPs, we might have taken both sets of results as evidence that our simulated agents feel less certain about their less preferred alternatives.

The impact of stochastic misspecification tends to be the most striking when ϵ_{nj} are drawn from the mixed normal distribution. One more pattern concerning other distributions is evident in the present experiment, which is to be seen in Experiment C again: the impact tends to be the greatest for the uniform, normal and logistic cases in order. This ordering agrees with the impression Figure 1 conveys regarding how much the density of each distribution overlaps with the $EV(0, 1, 0)$ density.

In further experimentation, we have found that the qualitative conclusions remain unchanged when the above experiments are repeated after scaling up or down all true

Table 4: Experiment B on hypothetical DGP in equation (20)

$\beta = (\beta_1 = 1, \beta_2 = -1, \beta_3 = 1)$

N=100	b^1			b^2			b^3		
	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
EV1	1.024	-1.023	1.024	1.013	-1.013	1.014	1.010	-1.010	1.012
	(0.261)	(0.155)	(0.195)	(0.184)	(0.111)	(0.139)	(0.156)	(0.093)	(0.119)
Uniform	1.102	-1.117	1.106	0.963	-0.969	0.965	0.900	-0.898	0.902
	(0.266)	(0.156)	(0.198)	(0.182)	(0.103)	(0.132)	(0.149)	(0.081)	(0.110)
Normal	1.122	-1.128	1.116	1.012	-1.018	1.009	0.949	-0.948	0.945
	(0.269)	(0.162)	(0.205)	(0.189)	(0.109)	(0.142)	(0.156)	(0.089)	(0.118)
Logistic	1.138	-1.148	1.138	1.046	-1.052	1.045	0.982	-0.984	0.981
	(0.275)	(0.168)	(0.205)	(0.192)	(0.115)	(0.144)	(0.160)	(0.095)	(0.121)
Mixed norm	1.229	-1.249	1.231	1.007	-1.015	1.006	0.895	-0.894	0.893
	(0.284)	(0.178)	(0.217)	(0.191)	(0.111)	(0.143)	(0.156)	(0.086)	(0.114)
	b^4								
	b_1	b_2	b_3	LR_2	LR_3	LR_4	$\widehat{\sigma}_2$	$\widehat{\sigma}_3$	$\widehat{\sigma}_4$
EV1	1.009	-1.009	1.011	5.17%	5.39%	5.73%	1.020	1.029	1.049
	(0.143)	(0.086)	(0.109)				(0.201)	(0.220)	(0.274)
Uniform	0.872	-0.869	0.874	18.13%	23.92%	24.99%	0.769	0.691	0.679
	(0.135)	(0.074)	(0.100)				(0.138)	(0.144)	(0.174)
Normal	0.910	-0.905	0.906	11.11%	18.14%	23.58%	0.833	0.729	0.652
	(0.143)	(0.081)	(0.108)				(0.150)	(0.157)	(0.173)
Logistic	0.936	0.936	0.936	8.94%	15.20%	24.40%	0.863	0.749	0.643
	(0.146)	(0.086)	(0.111)				(0.170)	(0.158)	(0.176)
Mixed norm	0.846	-0.838	0.844	37.57%	61.68%	68.81%	0.675	0.540	0.499
	(0.140)	(0.075)	(0.103)				(0.116)	(0.115)	(0.137)

N=300	b^1			b^2			b^3		
	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
EV1	1.007	-1.008	1.009	1.005	-1.005	-1.005	1.004	-1.004	1.005
	(0.146)	(0.087)	(0.110)	(0.110)	(0.062)	(0.080)	(0.089)	(0.052)	(0.068)
Uniform	1.088	-1.103	1.088	0.958	-0.964	0.955	0.895	-0.895	0.894
	(0.151)	(0.088)	(0.112)	(0.106)	(0.058)	(0.077)	(0.087)	(0.047)	(0.063)
Normal	1.101	-1.112	1.102	1.002	-1.008	1.002	0.940	-0.941	0.940
	(0.151)	(0.089)	(0.113)	(0.107)	(0.062)	(0.080)	(0.080)	(0.051)	(0.067)
Logistic	1.121	-1.127	1.120	1.037	-1.042	1.038	0.975	-0.977	0.975
	(0.153)	(0.092)	(0.116)	(0.109)	(0.064)	(0.082)	(0.092)	(0.054)	(0.069)
Mixed norm	1.209	-1.231	1.210	0.995	-1.007	0.995	0.886	-0.888	0.886
	(0.161)	(0.099)	(0.120)	(0.109)	(0.063)	(0.081)	(0.089)	(0.049)	(0.066)

(Continued on the next page.)

Table 4: Experiment B (*ctd.*)

(Continued from the previous page.)

	b^4								
	b_1	b_2	b_3	LR_2	LR_3	LR_4	$\widehat{\sigma}_2$	$\widehat{\sigma}_3$	$\widehat{\sigma}_4$
EV1	1.003	-1.004	1.004	4.95%	4.88%	4.88%	1.007	1.010	1.014
	(0.082	(0.048	(0.062)				(0.112)	(0.121)	(0.147)
Uniform	0.870	-0.866	0.869	53.08%	73.85%	76.98%	0.763	0.680	0.666
	(0.079	(0.043)	(0.058)				(0.077)	(0.081)	(0.096)
Normal	0.902	-0.899	0.903	29.17%	54.73%	72.63%	0.822	0.717	0.637
	(0.081)	(0.046)	(0.061)				(0.084)	(0.085)	(0.094)
Logistic	0.930	-0.928	0.930	19.58%	44.25%	72.24%	0.856	0.739	0.626
	(0.083)	(0.049)	(0.064)				(0.089)	(0.088)	(0.095)
Mixed norm	0.838	-0.834	0.838	91.73%	99.61%	99.85%	0.669	0.534	0.488
	(0.080)	(0.043)	(0.059)				(0.065)	(0.064)	(0.075)

N=500	b^1			b^2			b^3		
	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
EV1	1.003	-1.004	1.004	1.002	-1.002	1.003	1.002	-1.002	1.003
	(0.112)	(0.067)	(0.084)	(0.082)	(0.048)	(0.061)	(0.069)	(0.041)	(0.052)
Uniform	1.084	-1.099	1.084	0.954	-0.960	0.953	0.893	-0.893	0.892
	(0.117)	(0.067)	(0.086)	(0.081)	(0.045)	(0.059)	(0.067)	(0.036)	(0.049)
Normal	1.097	-1.108	1.099	1.000	-1.006	1.000	0.938	-0.939	0.938
	(0.116)	(0.069)	(0.086)	(0.081)	(0.048)	(0.062)	(0.068)	(0.039)	(0.052)
Logistic	1.114	-1.123	1.115	1.033	-1.039	1.034	0.972	-0.974	0.973
	(0.117)	(0.071)	(0.089)	(0.083)	(0.050)	(0.064)	(0.070)	(0.041)	(0.054)
Mixed norm	1.209	-1.225	1.207	0.994	-1.004	0.992	-0.886	-0.886	0.884
	(0.125)	(0.076)	(0.094)	(0.086)	(0.048)	(0.063)	(0.070)	(0.037)	(0.050)

	b^4								
	b_1	b_2	b_3	LR_2	LR_3	LR_4	$\widehat{\sigma}_2$	$\widehat{\sigma}_3$	$\widehat{\sigma}_4$
EV1	1.002	-1.002	1.002	4.94%	5.03%	4.91%	1.005	1.007	1.010
	(0.064)	(0.038)	(0.048)				(0.086)	(0.093)	(0.111)
Uniform	0.868	-0.864	0.867	80.35%	94.62%	96.23%	0.760	0.680	0.662
	(0.061)	(0.033)	(0.044)				(0.059)	(0.062)	(0.074)
Normal	0.901	-0.898	0.901	48.64%	82.00%	94.52%	0.820	0.714	0.634
	(0.062)	(0.036)	(0.047)				(0.066)	(0.065)	(0.073)
Logistic	0.927	-0.926	0.928	32.35%	70.62%	93.58%	0.853	0.737	0.623
	(0.064)	(0.038)	(0.049)				(0.069)	(0.068)	(0.073)
Mixed norm	0.838	-0.832	0.837	99.52%	100.00%	100.00%	0.668	0.532	0.487
	(0.063)	(0.033)	(0.046)				(0.050)	(0.049)	(0.057)

See notes to Table 3.

parameters by the same proportion.[18] But both variations in b^Q across Q, and the empirical sizes of the LR_Q tests under misspecification, have become smaller (larger) when the true parameters have been scaled up (down). That is, the impact of misspecification expectedly varies with how much the observed response is influenced by the systematic component of utility relative to the error terms.

Experiment C explores whether our previous results hold under an empirically plausible configuration of the systematic component of utility. Specifically, we generate synthetic data sets by combining the systematic component of utility estimated using an actual stated preference data set with simulated error terms. The actual data set is a subset of the electricity supplier choice data analyzed in Huber and Train (2001), which accompanies Stata module -mixlogit- (Hole, 2007).[19] We estimate MNL using this data set, and then generate rank-ordered responses from the following process:

$$U_{nj} = x_{nj}^{'}\beta + \epsilon_{nj} \tag{21}$$
$$n = 1, 2, \cdots, 1195 \quad j = 1, 2, 3, 4$$

where x_{nj} is a vector of 6 observed generic attributes distinguishing $J = 4$ alternatives in the actual data, and β is a vector of the empirical MNL estimates: see Hole (2007) for the full data description and estimates. To save space, we summarize the ROL estimates of 3 parameters out of 6: β_1 or the coefficient on the contract length that can be 0, 1 or 5 years; β_2 or the coefficient on the dummy indicator of whether the supplier is well-known (=1) or not (=0); and β_3 or the coefficient on the contract price in cents per kWH which can be 0, 7 or 9.

Table 5 summarizes the selected estimation results, which are qualitatively similar to what we have found in Experiment B. When ϵ_{nj} are drawn from $EV(0, 1, 0)$, all coefficient estimates b^1, b^2 are b^3 are close to the true parameters on average, and both LR_2 and LR_3 result in empirical rejection frequencies close to the nominal 5% level.

When ϵ_{nj} are drawn from other distributions, on average, b^1 are magnified relative to the true parameters, and the usual pattern of attenuation is observed across b^2 and b^3. This pattern is the most striking for the mixed normal, uniform, normal and logistic cases in order. The inequality $1 > \widehat{\sigma}_2 > \widehat{\sigma}_3$ also holds on average. The quantitative

[18] This is equivalent to maintaining the same true parameter values as before, and then specifying a smaller or larger variance of ϵ_{nj} than $\pi^2/6$.

[19] We have chosen this data set on the basis of its ready accessibility. Most of stated preference data sets are not as easily accessible, either because they are proprietary or have been collected for exclusive use by a team of researchers.

Table 5: Experiment C on synthetic DGP in equation (21)

$\beta = (\beta_1 = -0.140, \beta_2 = 1.055, \beta_3 = -0.635)$

N=1195	b^1			b^2			b^3		
	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
EV1	-0.140	1.060	-0.637	-0.140	1.057	-0.636	-0.140	1.057	-0.636
	(0.016)	(0.086)	(0.044)	(0.012)	(0.061)	(0.034)	(0.010)	(0.052)	(0.031)
Uniform	-0.151	1.135	-0.638	-0.122	0.931	-0.552	-0.114	0.860	-0.532
	(0.016)	(0.086)	(0.044)	(0.011)	(0.059)	(0.034)	(0.010)	(0.049)	(0.030)
Normal	-0.152	1.151	-0.664	-0.133	1.014	-0.599	-0.121	0.916	-0.563
	(0.017)	(0.087)	(0.045)	(0.012)	(0.061)	(0.035)	(0.010)	(0.050)	(0.031)
Logistic	-0.157	1.181	-0.690	-0.141	1.071	-0.632	-0.127	0.960	-0.589
	(0.017)	(0.089)	(0.045)	(0.012)	(0.062)	(0.035)	(0.010)	(0.051)	(0.032)
Mixed norm	-0.187	1.303	-0.755	-0.135	0.989	-0.580	-0.113	0.832	-0.527
	(0.018)	(0.090)	(0.047)	(0.012)	(0.061)	(0.036)	(0.010)	(0.051)	(0.032)
	LR_2	LR_3	$\widehat{\sigma}_2$	$\widehat{\sigma}_3$					
EV1	4.85%	5.05%	1.003	1.005					
			(0.074)	(0.089)					
Uniform	99.25%	99.50%	0.648	0.641					
			(0.055)	(0.067)					
Normal	83.96%	98.86%	0.747	0.594					
			(0.057)	(0.063)					
Logistic	67.58%	99.05%	0.787	0.583					
			(0.058)	(0.061)					
Mixed norm	100.00%	100.00%	0.484	0.372					
			(0.041)	(0.048)					

See notes to Table 3.

results are even more dramatic in the present experiment than in Experiment B. The equality of coefficients across first two ranks (LR_2) is rejected 68% of times in the logistic case, 84% in the normal case, and almost always in the uniform and mixed normal cases. The equality across all ranks (LR_3) is almost always rejected in every case. These larger rejection frequencies do not stem from a much larger sample and the presence of more parameters alone. The coefficient estimates *per se* exhibit greater divergence across ranks. On average, both b^2 and b^3 show much larger attenuation relative to b^1 than their counterparts in Experiment B, which can also be inferred from the visibly smaller means of $\widehat{\sigma}_2$ and $\widehat{\sigma}_3$, particularly in the mixed normal case.

5 Concluding remarks

In empirical works, the rank-ordered logit (ROL) estimates often vary significantly with the depth of rankings exploited in the estimation process. The common interpretation of this regularity is that rank-ordered data are unreliable because the cognitive burden of ranking several alternatives induces respondents to choose their most preferred alternatives differently from less preferred ones (Hanley *et al.*, 2001).

We advance an alternative explanation: stochastic misspecification of the random utility function. Our analytic and simulated examples illustrate the consequences of estimating a misspecified ROL model as follows. Even when all assumptions relevant to the estimated model are true except that the i.i.d. unobservables are not extreme value distributed, the estimates exhibit the pattern of variation which misleadingly agrees with the increasing cognitive burden of identifying less preferred alternatives. The tests of coefficient equality across ranks, when viewed as tests for the null of consistency in ranking behavior, falsely reject the null much more often than the nominal size of 5%; their empirical sizes even exceed 90% in some cases.

In practice, some amount of stochastic misspecification is inevitable because any model is but an approximation to reality. Our findings suggest that the sensitivity of the ROL estimates to the depth of rankings is to be naturally expected and needs not symptomize a data quality problem. Viewing this sensitivity as a consequence of stochastic misspecification, instead of the cognitive burden, has distinct methodological implications. On one hand, it vindicates the microeconomic approach of modeling rank orderings as realizations of random preference relations, as practiced by Calfee *et al.* (2001), Train and Winston (2007) and Dagsvik and Liu (2009). On the other hand, it provides a platform for developing more general an econometric procedure to address this sensitivity than the *ad hoc* modeling of cognitive processes. We conclude with a further remark on the latter implication.

In our view, a fruitful direction for future research lies in developing a semi-parametric procedure for rank-ordered data that allows consistent estimation of random utility model coefficients without specifying a particular error distribution. Such a procedure is a direct solution to the sensitivity of the ROL estimates to the depth of rankings, when its underlying cause is stochastic misspecification. It can also be used for a more clearcut test of consistency in ranking behavior, because the semi-parametric estimates must remain robust to the depth of rankings in the absence of inconsistent ranking behavior. Hausman and Ruud (1987) apply a precursor to this type of procedure which allows consistent estimation of coefficients on continuous attributes, and find some evidence of robustness. Since most of attributes in stated preference surveys are discrete, as in the Huber and Train (2001) data that we have used in our simulation, a more general procedure needs to be developed for wider empirical applicability.

Acknowledgment

We wish to thank Denise Doiron and Denzil G. Fiebig for helpful discussion. This paper is partly based on Yoo's previous research presented at the Australian Conference of Health Economists in 2012 and the International Choice Modelling Conference in 2013.

References

Anderson S, Palma A, Thisse J. 1992. *Discrete choice theory of product differentiation*. MIT Press: Cambridge, MA.

Beggs S, Cardell S, Hausman J. 1981. Assessing the potential demand for electric cars. *Journal of Econometrics* **16**: 1-19. [Link]

Ben-Akiva M, Morikawa T, Shiroishi F. 1992. Analysis of the reliability of preference ranking data. *Journal of Business Research* **24**: 149-164. [Link]

Berry S, Levinsohn J, Pakes A. 2004. Differentiated products demand system from a combination of micro and macro data. *Journal of Political Economy* **112**: 68-105. [Link]

Boyle K, Holmes T, Teisl M, Roe B. 2001. A Comparison of Conjoint Analysis Response Formats. *American Journal of Agricultural Economics* **83**: 441-54. [Link]

Calfee J, Winston C, Stempski R. 2001. Econometric issues in estimating consumer preferences from stated preference data: a case study of the value of automobile travel time. *Review of Economics and Statistics* **83**: 699-707. [Link]

Caparros A, Oviedo J, Campos P. 2008. Would you choose your preferred alternative? Comparing choice and recoded ranking experiments. *American Journal of Agricultural Economics* **90**: 843-855. [Link]

Chapman R, Staelin R. 1982. Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research* **19**: 288-301. [Link]

Dagsvik J, Liu G. 2009. A framework for analyzing rank-ordered data with application to automobile demand. *Transportation Research Part A* **43**: 1-12. [Link]

Dow J, Endersby J. 2004. Multinomial probit and multinomial logit: a comparison of choice models for voting research. *Electoral studies* **23**: 107-122. [Link]

Fok D, Paap R, Van Dijk B. 2012. A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. *Journal of Applied Econometrics* **27**: 831-846. [Link]

Foster V, Mourato G. 2002. Testing for consistency in contingent ranking experiments. *Journal of Environmental Economics and Management* **44**: 309-328. [Link]

Fry T, Harris M. 1996. A Monte Carlo study of tests for the independence of irrelevant alternatives property. *Transportation Research Part B* **30**: pp.19-30. [Link]

Greene W. 2008. *Econometric Analysis*, 6th ed. Pearson Prentice Hall: New Jersey.

Hanley N, Mourato S, Wright R. 2001. Choice modelling approaches: a superior alternative for environmental valuation? *Journal of Economic Surveys* **15**: 435-462. [Link]

Hausman J, Ruud P. 1987. Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics* **34**: 83-104. [Link]

Hole A. 2007. Fitting mixed logit models by using maximum simulated likelihood. *Stata Journal* **7**: 388-401. [Link]

Huber J, Train K. 2001. On the similarity of classical and Bayesian estimates of individual mean partworths. *Marketing Letters* **12**: 259-269. [Link]

Layton D. 2000. Random coefficient models for stated preference surveys. *Journal of Environmental Economics and Management* **40**: 21-36. [Link]

Lee L-F. 1995. Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics* **65**: 381-428. [Link]

McFadden D. 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka P. (Ed.), *Frontiers in Econometrics*. Academic Press: New York, pp.105-142. [Link]

Scarpa R, Notaro S, Louviere J, Raffaeli R. 2011. Exploring scale effects of best/worst rank ordered choice data to estimate benefits of tourism in Alpine Grazing Commons. *American Journal of Agricultural Economics* **93**: 813-828. [Link]

Train K. 2009. *Discrete choice methods with simulation*, 2nd ed. Cambridge University Press: New York. [Link]

Train K, Winston C. 2007. Vehicle choice behavior and the declining market share of U.S. automakers. *International Economic Review* **48**: 1469-1496. [Link]

Vossler C, Doyon M, Rondeau D. 2012. Truth in consequentiality: theory and field evidence on discrete choice experiments. *American Economic Journal: Microeconomics* **4**: 145-171. [Link]

Yoo H, Doiron D. 2013. The use of alternative preference elicitation methods in complex discrete choice experiments. *Journal of Health Economics* **32**: 1166-1179. [Link]