Preference reversal or limited sampling? Maybe túngara frogs are rational after all.

Paulo Natenzon*

October 2016

Abstract

In this paper, we demonstrate that revealed preference analysis and standard welfare analysis can be applied to context-dependent choice data. We apply the Bayesian probit model (Natenzon, 2016) to the experimental dataset on frog mating selection from Lea and Ryan (2015). We identify stable preferences from context-dependent data, and offer a new perspective in the debate about the rationality of context-dependent choice. We show that the Bayesian probit outperforms any random utility specification in goodness of fit and out of sample prediction. We conclude that our model presents a useful alternative to random utility —the current workhorse of discrete choice estimation— for applications where decision makers systematically exhibit context effects, including attraction, compromise and phantom alternative effects.

*PRELIMINARY DRAFT. COMMENTS WELCOME. Department of Economics, Washington University in St. Louis, St. Louis, MO 63130, USA. E-mail: pnatenzon@wustl.edu. I am indebted to John Yiran Zhu (The Wharton School, University of Pennsylvania) for bringing the dataset to my attention, and to Amanda M. Lea for useful comments.

1 Introduction

Consider the following type of choice reversal. Alternative B is chosen in more than 50% of the choice trials in binary comparisons between A and B,

$$P(B, \{A, B\}) > 1/2$$

while alternative A is chosen in more than 50% of the choice trials in ternary comparisons among A, B and C,

$$P(A, \{A, B, C\}) > 1/2.$$

Every time experimental researchers find a new recipe to systematically generate this type of choice reversal, it becomes a new puzzle. Famous examples include decoy effects —such as the attraction and the compromise effects— and phantom alternative effects (Huber et al. (1982), Huber and Puto (1983), Simonson (1989), Soltani, De Martino and Camerer (2012)). The choice frequency reversal above is puzzling because it is incompatible with random utility models —the de facto current workhorse in discrete choice estimation— including logit, probit, nested logit, mixed logit etc.

The random utility framework maintains the assumption that decision makers maximize utility, while allowing the utility of each option to be a random variable. This framework can generate a rich variety of choice patterns, accommodating heterogeneity of tastes in a population, stochastic taste shocks, and hand trembling mistakes. But it cannot accommodate the type of context dependence that is observed with decoy and phantom alternative effects. Often this leads analysts to conclude that the dataset is altogether incompatible with utility maximization. For example, Lea and Ryan (2015) interpret choice frequency reversals as arising from the reversal of the underlying preferences:

> *Female túngara frogs reversed their preferences in the presence of an irrelevant alternative in two separate experiments and thus violate a key assumption of mate choice models derived from decision theory. (Lea and Ryan, Science, Aug 2015)*

Our contribution in this paper is to demonstrate that revealed preference and standard welfare analysis can be applied to context-dependent choice data. We apply the Bayesian probit model (Natenzon, 2015) to the experimental dataset on frog mating selection from Lea and Ryan (2015). We identify stable, underlying preferences from context-dependent data, and offer a new perspective in the debate about the rationality of context-dependent choice.

The Bayesian probit maintains the assumption that preferences are represented by stable utility values, but relaxes the assumption that decision makers have perfect information about the value of each alternative when making a choice. The incredibly rich dataset on frog mating selection from Lea and Ryan (2015) offers a unique opportunity to show the advantages of this approach.

Why frogs? Decoy and phantom alternative effects were originally discovered in consumer choice experiments in marketing. They have been robustly replicated in medical decision making, voting, hiring and many other settings. But this type of context-dependent behavior is not unique to humans. Versions of the attraction and compromise effects have been documented for monkeys, birds, bees and even a slime mold (see Section 9 for the related literature). Due to the very careful experimental design (described in Section 2), this particular dataset on mating choices by túngara frogs presents an ideal setting to illustrate the advantages of our approach, allowing us to abstract from many distracting complications. Our model can transparently be shown to outperform any random utility specification in goodness-of-fit and out of sample prediction. We conclude that our model presents a useful alternative to random utility in applications where decision makers systematically exhibit context effects, including attraction, compromise and phantom alternative effects.

We describe the experimental data in Section 2, and the Bayesian probit model in Section 4. We show the advantages of the Bayesian probit over any random utility model in fitting the data in Section 6 and in out-of-sample prediction in Section 7. We show how classic welfare analysis can be extended to context-dependent choice , and discuss the implications for the rationality of context-dependent choice data in Section 8. Section 9 discusses our contribution in the

context of the existing literature and Section 10 concludes.

2 Choice reversals in the frog data

Female túngara frogs choose mating partners based on the sound of their call. Lea and Ryan (2015) simulate three different male frog calls in the lab, which they label as target (A), competitor (B) and decoy (C). The first three rows of Table 1 show how often female frogs chose each alternative when every pairwise combination was offered: A versus B, B versus C, and A versus C. Binary comparisons are statistically significant and (stochastically) transitive: $B \succ A$, $A \succ C$, and $B \succ C$. Hence the binary choice data reveals a complete and transitive ranking of the three options: $B \succ A \succ C$.

Presented alternatives	n	A	B	C
A and B	118	.37	.63	–
B and C	90	–	.69	.31
A and C	90	.84	–	.16
A, B, and C	40	.55	.28	.17
A, B, and $\cancel{C}$	79	.61	.39	–

Table 1: Choice frequencies by female túngara frogs in the dataset of Lea and Ryan (2015). The first three rows correspond to binary choice data and support $B \succ A \succ C$ as a rational benchmark. The fourth row shows choice frequencies when all three options are available. The fifth row shows frequencies for A and B when option C was located on the ceiling, so that it was presented but unreachable. While B is more likely to be chosen than A in binary choice (first row), the opposite happens in the presence of C (last two rows).

First choice reversal. The fourth row in Table 1 show that frogs were more likely to choose A over B when all three alternatives were offered. This contradicts the ranking $B \succ A \succ C$ obtained in binary choices.

Second choice reversal. The last row of Table 1 shows the frequencies of choice for options A and B when frogs could hear A, B and C but C was a *phantom alternative*. While the three options were equidistant from the frog in the experimental chamber, A and B were placed on the floor, while option C was

placed on the ceiling. Hence all three male calls could be heard, but only A and B were choosable. Comparing the first row to the last row of Table 1, we find that the presence of phantom alternative C significantly reversed the propensity of choosing A over B.

Simulated call attributes. The sound of male calls in túngara frogs are complex, with dozens of different measurable attributes. The experiment differentiated the three simulated options across two dimensions: the rate of calls per second, and a measure of static attractiveness. Both dimensions are desirable — previous studies have shown that increases in each of these dimensions lead, in general, to females choosing the option more often (see the supplemental online appendix in Lea and Ryan (2015) for details).

Following the marketing experimental literature, the three options were labeled target (A), competitor (B) and decoy (C). This labeling reflects their relative position in line with many experiments in the marketing literature on decoy effects (cf. Figure 1.D in Lea and Ryan (2015)). In particular:

(i) The target (A) and the competitor (B) are comparable alternatives. Neither options dominates the other;

(ii) The decoy (C) is more similar to the target (A) than to the competitor (B): C is closer to A than to B in every attribute.

(iii) The decoy (C) is more extreme than the other options and likely inferior. While C has a slightly higher value than A in one attribute, it has, by far, the lowest value of the other attribute among the three options.

3 Choice reversals versus random utility

The types of choice reversal observed in the frog data have been found in many settings. Every time a new experimental setup that is capable of systematically generate this type of choice reversal is discovered, it immediately becomes a puzzle. Famous examples are the attraction and the compromise effects. These puzzles arise because the data cannot be rationalized by any random utility

model, the current workhorse of discrete choice estimation. Commonly used random utility models include multinomial probit, logit, and generalizations of logit such as nested logit, mixed logit, and so on.

Random utility models maintain the assumption of utility maximization, while incorporating a stochastic additive shock to utility. In other words, the utility of each choice alternative is written as

$$U_i = \mu_i + \varepsilon_i, \qquad i = A, B, C$$

where μ_i is the deterministic component of utility and ε_i is a random term that captures taste variation, hand trembling mistakes, and other uncontrolled factors that appear random to the econometrician. For every state of nature in which $[U_A > U_B \text{ and } U_A > U_C]$, we obviously also have, in particular, that $[U_A > U_B]$. Therefore, the probability of these events must satisfy:

$$\mathbb{P}[U_A > U_B] \geq \mathbb{P}[U_A > U_B \text{ and } U_A > U_C] \tag{1}$$

Hence, when agents maximize random utility, the probability of A being chosen can only decrease once alternative C is introduced.

The inequality in (1) holds independently of any distributional assumptions the analyst makes about the utility shocks ε_i, including allowing the shocks to be correlated. Hence, it is impossible in the random utility framework to have a choice reversal as found in the fourth row of Table 1.

Moreover, the random utility framework does not allow phantom alternatives to have any effect on choice. When the only choosable options are A and B, the probability that A is chosen is always equal to $\mathbb{P}[U_A > U_B]$, independently of how many phantom alternatives are presented. Hence, the random utility framework constrains the choice probabilities that generate the first and last rows of Table 1 to be identical. The statistically significant difference in those rows provides strong evidence that the true data generating process lies beyond the random utility framework.

Confronting the data with the random utility framework often leads to the conclusion that subjects are irrational. For example, the choice behavior presented in Table 1 seems incompatible with the existence of stable, complete and

transitive preferences over the choice alternatives. Lea and Ryan (2015) interpret the choice data as arising from a reversal of the underlying preferences:

> *Female túngara frogs reversed their preferences in the presence of an irrelevant alternative in two separate experiments and thus violate a key assumption of mate choice models derived from decision theory. (Lea and Ryan, Science, Aug 2015)*

In the next section, we present an alternative to the random utility framework, and explain how context-dependent choice may arise from the maximization of stable utility values, under limited sampling.

4 Model: choice under limited sampling

The main assumption in the Bayesian probit model is that decision makers choose under *limited sampling*, i.e., they choose with imperfect information about the value of the options in any given menu (Natenzon, 2015). We assume the decision maker optimally uses the limited information available to choose the alternative with the highest expected value.

Multiple factors contribute to uncertainty and may lead to errors in mating choices among animals. Mating choices are made in complex, dynamic environments; individuals exhibit complex traits; time spent contemplating available mate options can increase the risk exposure to predators; organisms have limited cognitive resources; and so on. These factors naturally lead to limited sampling: organisms obtain imperfect evidence about the value of each alternative before making a choice.

Formally, the value of the options in the population is assumed to be identically and independently distributed according to a Gaussian distribution $\mu_i \sim \mathcal{N}(m, 1/s)$. Female frogs see every choice problem, a priori, as a collection of independent draws from a population of male frogs whose utility (or "darwinian fitness") is normally distributed with mean m and variance $1/s$.

In every choice trial, túngara frogs obtain some information about the value of available mates based on the sound of their call. We model the information

obtained about the utility of each option from the sound of their call as noisy signals with a jointly distributed Gaussian distribution $X_i = \mu_i + \varepsilon_i$, where μ_i is the true value of alternative i and ε_i is a random perception error with $\mathbb{E}[\varepsilon_i] = 0$ and $\text{Var}(\varepsilon_i) = 1/t$ for every i. We allow errors to be correlated, with $\sigma_{ij} := \text{Corr}(\varepsilon_i, \varepsilon_j) \in (-1, 1)$ for all i, j.

After obtaining a vector X of signal realizations, one for each available alternative, the decision maker chooses the alternative i that maximizes $\mathbb{E}[\mu_i|X]$. Hence, the decision maker uses all the information obtained from the noisy signals to update the prior and chooses the alternative with the highest posterior mean.

Allowing Bayesian updating introduces a fundamental departure from the random utility model. For example, the multinomial probit is a classic random utility model which shares the assumption of Gaussian distributed error terms. While in the random utility framework the decision maker chooses alternative i to maximize X_i, the Bayesian probit decision maker chooses i to maximize $\mathbb{E}[\mu_i|X]$. When signals are correlated, the alternative with the highest signal need not be the same as the alternative with the highest posterior mean.

In the random utility framework, the probability that option i is ranked above option j is given by $\mathbb{P}\{X_i > X_j\}$ in every menu of alaternatives. In contrast, in the Bayesian probit the option i is ranked above j when $\mathbb{E}\{\mu_i|X\} > \mathbb{E}\{\mu_j|X\}$. The distribution of $\mathbb{E}\{\mu_i|X\}$ depends on the available signals X and will change from one menu to another. In particular

$$\mathbb{E}\{\mu_A|X_A, X_B\} \neq \mathbb{E}\{\mu_A|X_A, X_B, X_C\}$$

whenever X_C is correlated with X_A. Hence, the presence of a signal about C can influence the probability of choosing A over B.

The parameters of the prior are never identified from choice data and can without loss of generality be normalized to $\mu_i \sim \mathcal{N}(0, 1)$:

Lemma 1. *The Bayesian probit with prior $\mathcal{N}(\hat{m}, 1/\hat{s})$, utility parameters $\hat{\mu}_i \in \mathbb{R}$, precision $\hat{t} > 0$ and correlation parameters $-1 < \hat{\sigma}_{ij} < 1$ is observationally*

equivalent to the Bayesian probit with prior $\mathcal{N}(0,1)$*, utility* $\mu_i = \sqrt{\hat{s}}(\hat{\mu}_i - \hat{m})$*, precision* $t = \hat{t}/\hat{s}$ *and correlation* $\sigma_{ij} = \hat{\sigma}_{ij}$*.*

Lemma 1 has two implications for the interpretation of parameter estimates. With the normalized prior $\mathcal{N}(0,1)$, the utility parameter μ is measured in standard deviations from the population mean. For example, $\mu_i = -0.5$ means that the darwinian fitness of frog i is half a standard deviation below the population mean. Likewise, the precision parameter $t = \hat{t}/\hat{s}$ measures the precision of the information obtained by the decision maker in units of precision in the population distribution. This is another departure from the random utility framework, where utility parameters have only ordinal (but not cardinal) meaning.

The choice frequencies generated in each menu of alternatives by the Bayesian probit are a random choice rule P parameterized by a utility vector $\mu = (\mu_A, \mu_B, \mu_C)$ and a covariance matrix $(1/p)\Sigma$ where

$$\Sigma = \begin{bmatrix} 1 & \sigma_{AB} & \sigma_{AC} \\ \sigma_{AB} & 1 & \sigma_{BC} \\ \sigma_{AC} & \sigma_{BC} & 1 \end{bmatrix}$$

We let $P(A, \{A, B, C\})$ denote the probability that option A is chosen when A, B, C are presented, etc.

5 Revealed Preference and Revealed Similarity

The binary comparison choice probabilities in the Bayesian probit with parameters μ, p, Σ are given by:

$$P(i, \{i, j\}) = \Phi\left(\frac{\sqrt{p}}{\sqrt{2}} \times (\mu_i - \mu_j) \times \frac{1}{\sqrt{1 - \sigma_{ij}}}\right) \tag{2}$$

where Φ is the standard Gaussian cumulative distribution function. The binary choice formula (2) shows that the ability of the decision maker to compare and correctly discriminate among the pair of options i, j increases with the overall

precision of the signals p, the difference in value $|\mu_i - \mu_j|$ and the signal correlation σ_{ij}. We refer to σ_{ij} as the *similarity* parameter following the literature in psychology and psychophysics (see Appendix A).

The binary choice probabilities in 2 are equivalent to the choice probabilities generated by classic binary multinomial probit under with the same parameters, under the assumption of equal variances (Natenzon, 2016).

We say that an alternative i is *revealed preferred* to j, and we write $i \succ j$ whenever $P(i, \{i, j\}) > 1/2$. The binary choice frequencies on the first three rows of Table 1 reveal the ranking $B \succ A \succ C$, where all the binary comparisons are easily seen to be statistically significant. From equation (2) it is clear that the Bayesian probit and the classic multinomial probit generate the revealed preference ranking obtained from the data if and only if $\mu_B > \mu_A > \mu_C$.

We say that a pair of alternatives $\{i, j\}$ is *easier to discriminate* than a pair $\{k, \ell\}$ if

$$\left| P(i, \{i, j\}) - \frac{1}{2} \right| > \left| P(k, \{k, \ell\}) - \frac{1}{2} \right|.$$

Hence, a pair is easier to discriminate whenever the choice frequencies among the options in the pair are more extreme (i.e., closer to zero and one). The first three rows of Table 1 reveal that the pair $\{A, C\}$ is easier to discriminate than $\{B, C\}$, which, in turn, is easier to discriminate than $\{A, B\}$.

Recall $\succ$ denotes the revealed preference relation. Let $i \sim j$ whenever P(i,{i,j})=1/2 and write $P(i, \{i, i\}) = 1/2$ so that $i \sim i$ for all i. Finally, let $i \succsim j$ whenever $i \succ j$ or $i \sim j$ hold. We say that a pair $\{i, j\}$ is *revealed more similar* than the pair k, ℓ whenever (i) $k \succsim i \succ j \succsim \ell$; and (ii) $\{i, j\}$ is easier to discriminate than $\{k, \ell\}$. Whenever $k = i$ above that i is revealed more similar to j than to ℓ; and whenever $j = \ell$ above we say that j is revealed more similar to i than to k.

According to the first three rows of Table 1, the pair $\{A, C\}$ is revealed more similar than the pair $\{B, C\}$. In other words, C is revealed more similar to A than to B. To see this note that (i) $B \succ A \succ C$ and (ii) the pair $\{A, C\}$ is easier to discriminate than any other pair. It follows from equation (2) that the Bayesian probit accomodates the revealed similarity relation only if $\sigma_{AC} > \sigma_{BC}$.

Note that since $\{B, C\}$ is easier to discriminate than $\{A, B\}$, no other pairs are revealed more similar according to the definition above. In other words, we cannot infer just based on the binary choice formula (2) how σ_{AB} relates to the other two correlation parameters. It is possible that the pair $\{A, B\}$ is the hardest to discriminate because A and B are very close in value (i.e., $|\mu_A - \mu_B|$ is small), or because σ_{AB} is low, or both.

We made the revealed preference and revealed similarity statements above based only on the choice frequencies of Table 1 and the binary choice formula (2). In particular, we did not refer to any measurable attributes of the alternatives or the decision makers. In principle, our model can shed light on individual choice data independently of the number of measurable attributes that are available to the analyst. Likewise, we keep the estimation exercise "attribute free" in the next section to avoid unnecessary distractions. But it should be noted that, just as in the traditional probit model, the analyst can incorporate a vector of observable attributes into the analysis by formulating and testing additional assumptions specifying the dependence of each parameter μ, p, Σ on the observed characteristics.

6 Structural estimation of preference

The choice frequencies presented in Table 1 have six degrees of freedom. The Bayesian probit model has seven parameters: μ_A, μ_B, μ_C are the utilities of each choice option, $\sigma_{AB}, \sigma_{AC}, \sigma_{BC}$ reflect the comparability of each pair of options, and p is the precision of the signals. We fit the model by maximum likelihood, imposing two parameter restrictions:

$$\sigma_{BC} = 0 \text{ and } \mu_B = 1.96 \tag{3}$$

The restricted version of the model has five degrees of freedom. Given Lemma 1, the restriction $\mu_B = 1.96$ fixes the utility of frog B to the utility of a frog in the 97th percentile of the population distribution. Alternative restrictions to (3) change the pointwise estimates but do not change any of the qualitative results. We estimate the model under different sets of restrictions in the Appendix.

Parameter	Estimate	(St. Dev.)
μ_A	0.4540	(.)
μ_B	1.9600	n/a
μ_C	-0.5661	(.)
σ_{AB}	0.1547	(.)
σ_{AC}	0.9516	(.)
σ_{BC}	0.0000	n/a
p	0.0749	(.)
Log Like	-265.8687	

Table 2: Point estimates for utility μ_i, similarity σ_{ij} and information precision p in the Bayesian probit model imposing the additional restrictions in equation (3).

The parameter estimates of Table 2 tell a rational story of utility maximization under limited sampling. The three alternatives can be ranked in a complete and transitive manner according to value. Alternative B is the best alternative, with $\mu_B = 1.96$, which means B is almost two standard deviations above the population mean in terms of 'darwinian fitness'. Hence, frog B is better than 97% of the population. Frog A comes in second place, approximately half a standard deviation above the population mean (better than 67% of the population). Finally, frog C is the worst, being more than half a standard deviation below the population mean (better than 28% of the population).

Given this rational benchmark, we interpret every instance in which an inferior option was chosen in the dataset as a *mistake* resulting from limited sampling. The precision parameter $p = 0.07$ gives a measure of the limitation. The estimated correlation parameters give a measure of the pairwise similarity or comparability of the options. Alternatives A and C are the most similar, with correlation $\sigma_{AC} = 0.95$. Alternatives B and C are the least similar, with estimated correlation σ_{BC} equal zero. These estimates of similarity/comparability are in line with the experimental design, where alternative C is placed closer to alternative A than to B across every simulated characteristic. The next table shows these estimates provide a good fit to the experimental data.

6.1 Comparison to random utility

Table 3 compares the choice frequencies in the data to the estimated choice probabilities for the Bayesian probit (BP) and for the random utility model (RUM). Estimates were obtained by maximum likelihood. We estimated a five-parameter model of the entire RUM family (this is possible due to the simplicity of the data and to Lemma 2 below). Hence the fit of the RUM model in Table 3 is, by definition, superior to the fit of any of the special cases of RUM such as logit, probit, nested logit, mixed logit, and so on.

	data			BP			RUM		
Menu	A	B	C	A	B	C	A	B	C
$\{A, B\}$	.37	.63	–	.37	.63	–	.48	.52	–
$\{B, C\}$	–	.69	.31	–	.69	.31	–	.69	.31
$\{A, C\}$	.84	–	.16	.84	–	.16	.83	–	.17
$\{A, B, C\}$	.55	.28	.17	.57	.26	.17	.48	.35	.17
$\{A, B, \not{C}\}$	.61	.39	–	.60	.40	–	.48	.52	–

Table 3: Comparison of relative frequencies of choice in the original data (left), estimated choice probabilities for the Bayesian probit (middle) and estimated choice probabilities for the random utility model (right).

Table 3 illustrates the difficulty of RUM models to capture the choice frequency reversals in the data. First, the RUM model restricts the probability of choosing A from $\{A, B\}$ to be the same independently of the presence of the phantom alternative C. Thus, the RUM model restricts the first row and the last row of Table 3 to be identical. Moreover, in the RUM model the probability of choosing A from $\{A, B\}$ can only increase when the decoy alternative C is added to the menu. Thus, the RUM model restricts the probability of choosing A in the first row of Table 3 to be at least as large as the probability of choosing A in the fourth row of Table 3.

In the random utility model (RUM), the utility of each choice alternative is

a random variable

$$\begin{aligned} U_A &= u_A + \varepsilon_A \\ U_B &= u_B + \varepsilon_B \\ U_C &= u_C + \varepsilon_C \end{aligned} \tag{4}$$

where u_i is the deterministic component of the utility of alternative i and ε_i is a stochastic taste shock for $i = A, B, C$. Under the additional assumption that the stochastic taste shocks are joint normally distributed, we obtain the multinomial probit model; if the shocks are iid Gumbel distributed, we obtain the logit model; assuming shocks have a joint Generalized Extreme Value distribution, we obtain generalizations of logit such as nested logit, cross nested logit, and so on (see any standard textbook, e.g. Train (2009)).

Instead of separately testing the fit of a random utility model under different assumptions about the distribution of taste shocks, we will use the following lemma to find a single best fit among the entire set of random utility models:

Lemma 2 (Block and Marschak, 1960)**.** *Every RUM is equivalent to a probability measure over the $n!$ strict rankings over the n choice alternatives.*

Proof. See Block and Marschak (1960), Theorem 3.1. □

Let $p_{ABC} \in [0, 1]$ denote the probability of the strict ranking $A \succ B \succ C$ and analogously denote the probability of every other strict ranking over the three alternatives. Since there is a total of $3! = 6$ such rankings, every RUM can be described by the five parameters $p_{ABC}, p_{ACB}, p_{BAC}, p_{BCA}, p_{CAB} \in [0, 1]$ with $p_{ABC} + p_{ACB} + p_{BAC} + p_{BCA} + p_{CAB} \leq 1$.

Under the assumption that choice trials are independent, the likelihood of

obtaining the sample data in Table 1 is maximized at

p_{ABC}	0.343298
p_{ACB}	0.140951
p_{BAC}	0.345591
p_{BCA}	0.000000
p_{CAB}	0.000000
p_{CBA}	0.170160
Log-likelihood	−271.044906

6.2 Model selection

The Akaike information criterion (AIC) offers an estimate of the information loss when a given model is used to represent the data Akaike (1974). The AIC is equal to $2k - 2\ln(L)$ where k is the number of free parameters and L is the maximum value of the likelihood function for the model. It rewards goodness of fit, measured by the likelihood function, and includes a penalty that increases in the number of parameters to discourage overfitting. The table below shows that the Bayesian Probit has the lowest value of AIC. The AIC for the Bayesian probit is calculated with a penalty for $k = 7$ parameters, which means it has an even lower AIC with the restriction imposed by equation (3).

Model	k	loglike	AIC
BProbit	7	−265.86	543.74
RUM	5	−271.04	552.09
Logit	2	−275.75	555.49
Probit	4	−274.32	556.65

7 Out-of-sample prediction

We present two out-of-sample prediction exercises using the published data of Lea and Ryan (2015). First, we estimate the model based on the odd numbered cases in the dataset, and assess how well the fitted model predicts the choices for

the even numbered cases. Second, we estimate the model excluding the choice trials for a particular menu, and compare the predicted choice probabilities to the sampled probabilities in that menu. We do the second exercise separately for every menu of alternatives. In both prediction exercises our model performs significantly better out-of-sample than any random utility model.

7.1 Half-sample estimation and prediction

	data (even)			Bayesian P			RUM			Logit			Probit		
Menu	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
$\{A,B\}$	.39	.61	–	.36	.64	–	.45	.55	–	.47	.53	–	.45	.55	–
$\{A,C\}$	.87	–	.13	.80	–	.20	.82	–	.18	.79	–	.21	.82	–	.18
$\{B,C\}$	–	.56	.44	–	.79	.21	–	.82	.18	–	.81	.19	–	.82	.18
$\{A,B,C\}$	.65	.20	.15	.47	.40	.13	.45	.37	.18	.41	.48	.11	.45	.37	.18
$\{A,B,\not C\}$	.64	.36	–	.57	.43	–	.45	.55	–	.47	.53	–	.45	.55	–
RSS	–			96.3			100.1			101.2			100.1		

Table 4: Out-of-sample prediction with each model estimated using odd-numbered choice trials. The table compares data from even-numbered choice trials to the pseudo out-of-sample predictions from the Bayesian Probit, the RUM family as a whole, and two particular cases of RUM, Logit and Probit. We restrict the Bayesian probit according to equation (3). Estimation details are presented in the Appendix.

7.2 Out-of-sample choice menus

Table 7.2 presents the prediction of a pseudo out-of-sample exercise in which all the choice trials for a single menu of alternatives is treated as out-of-sample. For example, the first column of Table 7.2 compares the actual data to the prediction results when the every model is estimated using the data for every menu except $\{A, B\}$. An asterisk (*) in the table means the model does not have enough empirical bite to predict a single choice probability. The Bayesian probit performs better than the other models in every instance, except for the menu $\{A, B\}$, where every model does poorly, and for $\{A, B, C\}$, where it doesn't offer

	data (odd)			Bayesian P			RUM			Logit			Probit		
Menu	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
$\{A,B\}$	.36	.64	–	.41	.59	–	.52	.48	–	.58	.42	–	.54	.46	–
$\{A,C\}$	.82	–	.18	.83	–	.17	.85	–	.15	.75	–	.25	.86	–	.14
$\{B,C\}$	–	.82	.18	–	.59	.41	–	.56	.44	–	.68	.32	–	.56	.44
$\{A,B,C\}$	.45	.35	.20	.64	.26	.10	.52	.32	.15	.48	.35	.16	.45	.44	.10
$\{A,B,\not{C}\}$	.58	.43	–	.68	.32	–	.52	.48	–	.58	.42	–	.54	.46	–
SSR	–			92.8			95.7			93.8			96.7		

Table 5: Out-of-sample prediction with each model estimated using even-numbered choice trials. The table compares data from odd-numbered choice trials to the pseudo out-of-sample predictions from the Bayesian Probit, the RUM family as a whole, and two particular cases of RUM, Logit and Probit. We restrict the Bayesian probit according to equation (3). Estimation details are presented in the Appendix.

a single prediction. The only model that has enough empirical bite in every menu at this level of generality is the Logit. In applications model acquire the necessary level of empirical bite because the analyst analyzes models choices conditional on a vector of observables.

Menu	A	B	A	C	B	C	A	B	C	A	B	$\not{C}$
Data	.37	.63	.84	.16	.69	.31	.55	.28	.17	.61	.39	–
BP	.68	.32	.73	.27	.68	.32	$*$	$*$	$*$	.61	.39	–
Logit	.64	.36	.69	.31	.81	.19	.44	.43	.13	.49	.51	–
Probit	.64	.36	.57	.43	.83	.17	$*$	$*$	$*$	.46	.54	–
RUM	.61	.39	$*$	$*$	$*$	$*$	$*$	$*$	$*$	.42	.58	–

Table 6: Out-of-sample prediction with each model estimated excluding all choice trials for a single menu of alternatives. We compare observed data of each menu to the pseudo out-of-sample predictions from the Bayesian Probit, the RUM family as a whole, and two particular cases of RUM, Logit and Probit. We restrict the Bayesian probit to have utility parameters lying in the 95% confidence interval according to the decision maker's prior. Estimation details are presented in the Appendix.

8 Rationality, Monty Hall and evolution

Are the observed choices rational? We draw two main lessons from the analysis of the data using the Bayesian probit model.

First, the data is compatible with the maximization of the complete and transitive ranking $B \succ A \succ C$, under limited sampling. Hence, the choice data is compatible with the narrow definition of 'rational' used in microeconomic decision theory: there exists a stable, complete and transitive preference underlying the observed choices.

Second, we can interpret the choice reversals observed in the lab as a direct consequence of the experimental design. The observed choice behavior is compatible with a decision procedure that maximizes the expected value of the chosen alternative when options are drawn independently from the same population. In contrast, the experimental design did not draw options A, B and C randomly and independently from the population. Instead, these options were carefully *designed* to emulate the types of choice reversals observed in the marketing literature. This is in line with a classic interpretation of many results in behavioral economics: decision makers utilize heuristics or rules of thumb that perform well on average outside of the lab (Tversky and Kahneman, 1974). By carefully manipulating the choice environment in the lab, experiments are able to tease out the biases that result from the employment of these heuristics.

Why can't frogs figure out the lab may be different from a typical choice situation? Amphibians have evolved for millions of years. Any decision procedure that mimics Bayesian updating has an evolutionary advantage by getting closer to maximizing the expected value (or "darwinian fitness") of their mating partners. The behavior of frogs can be seen as procedurally rational, in the sense of being optimal for random encounters with potential mates in nature, while at the same time completely failing to respond to the special conditions of the lab environment.

Our explanation for the optimality of the observed choice behavior is analogous to the solution to the classic Monty Hall problem (Selvin, 1975). In the

Monty Hall problem a prize is equally likely to be hidden in one three boxes A, B, C. The contestant initially points to a box. Monty, the game host, has to open one of the two remaining boxes. Since Monty knows where the prize is, Monty always opens an empty box. Suppose for the sake of the example that the contestant initially pointed to box B, and that Monty opened the empty box C. The problem asks if the contestant should switch from the initial pick B to the unopened box A if given the chance. The answer is yes: the strategy of never switching boxes gives $1/3$ probability of winning the prize, while the strategy of always switching boxes results in $2/3$ probability of winning.

The optimality of switching boxes in the Monty Hall problem is shown by Bayesian updating. Monty's action reveals no information about the initial pick B, but makes it very easy to compare the two remaining boxes A and C. When all options are identical a priori, a Bayesian decision maker optimally chooses among options that are easier to compare.

Many people have trouble understanding the solution to the Monty Hall problem. Why would frogs be able to solve it? To implement the optimal strategy, frogs don't need the ability to explicitly calculate conditional probabilities. Any heuristic that favors options that are easier to compare leads to better mating choices and better darwinian fitness, just like the pure heuristic of always switching doors wins more often in the Monty Hall problem than the pure heuristic of never switching doors. Evolutionary pressure may therefore lead to a hard-wired bias towards choosing among more options that are easier to compare. This mechanism may at least in part help explain the prevalence of the attraction and compromise effects in experiments involving choices by slime molds, bees, birds, frogs, monkeys, and humans.

9 Related Literature

There is a vast literature documenting systematic frequency choice reversals among human subjects.

The specific choice reversals commonly referred to as the attraction and com-

promise effect have also been found among Rhesus macaques (Parrish et al., 2015), honeybees, grayjays (Shafir et al., 2002), hummingbirds (Bateson et al., 2002), and even a unicellular slime mold (Latty and Beekman, 2011). The slime mold is noteworthy because it has no brain; it doesn't have a single neuron.

Our model incorporates several insights from behavioral economics. The idea that rational subjects make optimal inferences from the menu as an explanation for choice paradoxes is proposed in Kamenica (2008).

10 Concluding Remarks

We conclude that the Bayesian probit may be a useful tool for discrete choice estimation for datasets on human subjects in which choices are context-dependent. In contrast to the random utility framework currently used in most discrete choice estimation applications, the Bayesian probit allows the analyst to identify complete and transitive preferences that underlie the data, which in turn makes the data amenable to standard welfare analysis.

By suggesting a particular mechanism for how choice reversals may arise, the Bayesian probit can also guide the design of future experiments. In some experiments in the literature —see, for example, the choice over lotteries experiment of Soltani et al. (2012)— the order in which a subject is presented with choice problems, and the position of the choice options on the computer screen for each choice problem are randomly generated. In these cases, the assumption of a symmetric prior may be a good approximation to the subject's ex-ante information about the value of the choice options in any given choice trial. Future experimental work could attempt to manipulate the prior of the subjects and draw inferences between the level of symmetry in the prior and the magnitude of context-dependence in the subject's choices.

A Similarity and Correlation

Cognitive tasks in experimental psychology and psychophysics ask decision makers to choose the largest geometrical figure, the heaviest object, the loudest sound, the darkest shade of gray, and so on. A common finding in these experiments is that decision makers do a better job discriminating among a pair of options i, j when the difference in value is greater. Another common finding is that, keeping the value of the alternatives i, j constant, the ability to discriminate the options improves when the alternatives are more similar. This regularity is known in the psychology literature at least since the experimental work of Tversky and Russo (1969):

> *The similarity between stimuli has long been considered a determinant of the degree of comparability between them. In fact, it has been hypothesized that for a fixed difference between the psychological scale values, the more similar the stimuli, the easier the comparison or the discrimination between them.*

For example, consider the visual task of choosing the triangle with the largest area in Figures 1 and 2. Decision makers typically find it easier to choose, and make less mistakes, in Figure 2. Triangle j is identical in Figures 1 and 2. Triangle i on the left in Figure 1 is very different from triangle i' on the left in Figure 2, but i and i' have exactly the same area. Hence, the difference in area between i and j in Figure 1 is equal to the difference between i' and j in Figure 2. The triangles in Figure 2 are easier to compare because they are more similar.

For a second visual example, suppose we ask subjects to choose which star has more points in Figures 3 and 4 (I am grateful to David K. Levine for suggesting this example.) The star on the left is the same in both Figures. The start on the right has the same number of points in both Figures. Again, subjects usually find it easier to choose and make less mistakes in Figure 4 where the pair of stars is more similar.

In both examples, our model captures the similarity of the alternatives with the correlation parameter. The more similar pair has a higher value of correlation.

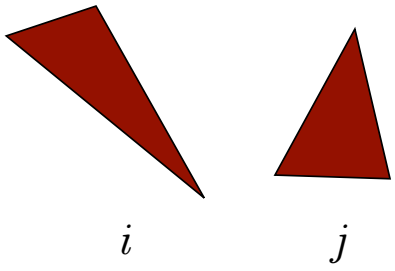

Figure 1: Which triangle has the largest area?

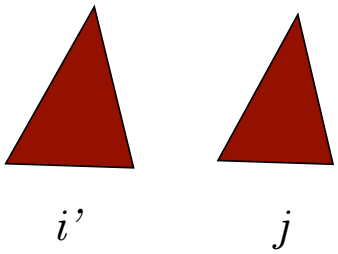

Figure 2: Which triangle has the largest area?

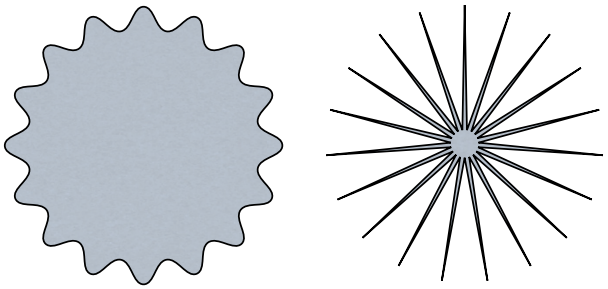

Figure 3: Which star has more points?

Finally, consider an example from a familiar setting of choice problems in economics. Suppose each choice object is a simple lottery (p, m) described by a probability $0 < p < 1$ of winning and a monetary prize $m \geq 0$. Such lotteries are commonly offered in experimental work, e.g. Soltani, De Martino and Camerer

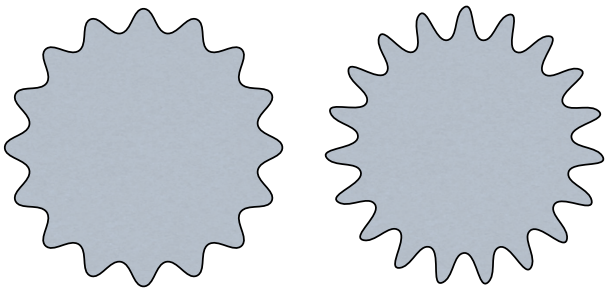

Figure 4: Which star has more points?

(2012). Suppose revealed preference analysis determines for an economic agent that lotteries B, C, D, E are on the same indifference curve while lottery A is superior, as depicted in Figure 5. Tversky and Russo's idea applied to this setting implies the intuitive ranking

$$\rho(A, E) > \rho(A, D) > \rho(A, C) > \rho(A, B) > 1/2.$$

The difference in utility is the same in every pairwise choice, but mistakes are more likely when the options are less similar. Here, similarity means distance according to some metric on an Euclidean space of observed attributes. Fixing the indifference curves, options are harder to compare when they are more distant in the attribute space. In Figure 5, options A and B are the most difficult to compare, while options A and E are the easiest. Note that option A strictly dominates option E, offering a larger prize and a larger probability of winning. The same difference in utility becomes more transparent when options are more similar. Strict dominance can be interpreted as a particularly extreme form of similarity.

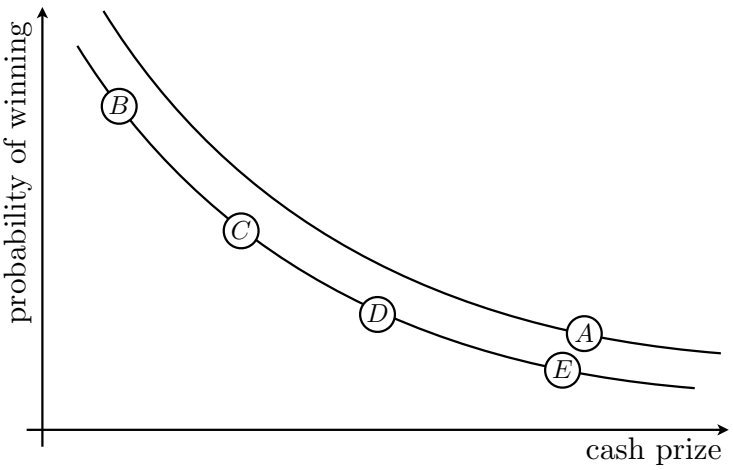

Figure 5: The two indifference curves represent the decision maker's true preferences of over simple money lotteries. Lotteries B, C, D, E lie on the same indifference curve, while lottery A is superior. In pairwise choice tasks, mistakes are more likely when the options are less similar. The comparison of A versus B is the hardest, while the comparison of A and E is the easiest.

B RUM estimates

C Logit estimates

D Probit estimates

E Bayesian probit estimates

References

Akaike, Hirotugu, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, 1974, *19* (6), 716–723.

Bateson, Melissa, Susan D Healy, and T Andrew Hurly, "Irrational choices in hummingbird foraging behaviour," *Animal Behaviour*, 2002, *63* (3), 587–596.

Block, H. D. and Jacob Marschak, *Random Orderings and Stochastic Theories of Responses* number 66, Stanford, CA: Stanford University Press,

Huber, Joel and Christopher Puto, "Market Boundaries and Product Choice: Illustrating Attraction and Substitution Effects," *The Journal of Consumer Research*, 1983, *10* (1), pp. 31–44.

_ , J. W. Payne, and C. Puto, "Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis," *Journal of Consumer Research*, 1982, *9* (1), 90–98.

Kamenica, Emir, "Contextual inference in markets: On the informational content of product lines," *The American Economic Review*, 2008, *98* (5), 2127–2149.

Latty, Tanya and Madeleine Beekman, "Irrational decision-making in an amoeboid organism: transitivity and context-dependent preferences," *Proceedings of the Royal Society of London B: Biological Sciences*, 2011, *278* (1703), 307–312.

Lea, Amanda M and Michael J Ryan, "Irrationality in mate choice revealed by túngara frogs," *Science*, 2015, *349* (6251), 964–966.

Natenzon, Paulo, "Random Choice and Learning," Theoretical Economics, revise and resubmit June 2016.

Parrish, Audrey E, Theodore A Evans, and Michael J Beran, "Rhesus macaques (Macaca mulatta) exhibit the decoy effect in a perceptual discrimination task," *Attention, Perception, & Psychophysics*, 2015, *77* (5), 1715–1725.

Selvin, Steve, "A Problem in Probability," *The American Statistician*, 1975, *29* (1), 67–71.

Shafir, Sharoni, Tom A Waite, and Brian H Smith, "Context-dependent violations of rational choice in honeybees (Apis mellifera) and gray jays

(Perisoreus canadensis)," *Behavioral Ecology and Sociobiology*, 2002, *51* (2), 180–187.

Simonson, Itamar, "Choice Based on Reasons: The Case of Attraction and Compromise Effects," *The Journal of Consumer Research*, 1989, *16* (2), pp. 158–174.

Soltani, Alireza, Benedetto De Martino, and Colin Camerer, "A Range-Normalization Model of Context-Dependent Choice: A New Model and Evidence," *PLoS computational biology*, 2012, *8* (7), e1002607.

Train, Kenneth, *Discrete Choice Methods with Simulation*, 2nd ed., Cambridge University Press, 2009.

Tversky, Amos and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science*, 1974, *185* (4157), 1124–1131.

_ **and J. Edward Russo**, "Substitutability and similarity in binary choices," *Journal of Mathematical Psychology*, 1969, *6* (1), 1–12.