What Can Be Learned from Behavior? Predictive Ability in Discrete Choice Environments*

María José Boccardi

NYU Abu Dhabi †

November 2, 2017

Abstract

Revealed preference restrictions provide testable implications for many theories of consumption behavior. Often, empirical evidence finds violations to these conditions which raises the question of how severe these are. The severity of the violations does not only depend on the extent of the observed deviations, but also on the sensitivity of the test to detect them. This paper provides a joint treatment for the severity of the violations and the sensitivity of the test in discrete choice environments by assessing the amount of information about preferences that can be inferred from data while allowing for errors. The proposed approach allows to compare across (limited) data sets and different models of behavior, addressing the concern raised by De Clippel and Rozen (2014) about extensibility of bounded rationality models.

1 Introduction

Axiomatizations provide testable implications for many theories of consumption, but often empirical evidence finds violations to these conditions leading to the rejection of the model. This has been extensively studied with respect to the utility maximization model by either providing measures of the extent of the observed violations (if any), or proposing alternative models of behavior. This paper extends the predictive approach proposed by Boccardi (2017) to discrete choice environments where menus need not to have a budget structure. This extension allows the researcher to asses the performance across behavioral models and within models for different data sets by providing a measure that accounts for: (i) the severity of the violations (if any); (ii) the

*Click here for an updated version

†e-mail: *Maria.Jose.Boccardi@nyu.edu*

amount of preference information that can be inferred from the data and therefore the sensitivity of the test; and (iii) the extensibility of the predictions for unobserved environments.

A series of goodness of fit measures have been proposed in the literature to assess the severity of the violations, if any, for the utility maximization model. Some of these measures rely on the number of violations (Famulari (1995), Houtman and Maks (1985)) but these do not provide a sense of the significance of such deviations. Cost based measures size the significance of the departures from rationality relying on the budget structure of the menus (Afriat (1967), Varian (1990)) but these are extremely sensitive to one 'bad' violation and cannot be extended to other menu settings. 'Hybrid' measures reflect both, the extent and number of violations (Echenique *et al.* (2011), Dean and Martin (2016)) relying on the budget structure of the menus. Apesteguia and Ballester (2015) proposes a measure of the severity of the violations in terms of the welfare loss, recovering the preference relation that minimizes that metric. This approach does not rely on the budget set structure of the menus. I show that the procedure proposed by Apesteguia and Ballester (2015) recovers the true underlying preferences when choices are observed from a sufficiently informative subset of all possible menus. Moreover, it can be employed to recover the fundamentals for other theories of consumption.

For the utility maximization model, revealed preference theory provides an axiomatization that allows to test for the consistency of the model even in limited data sets. However, the axiomatization for many bounded rationality models assume complete datasets. When dealing with empirical datasets it is not possible to control how exhaustive these are, and for experiments the data requirements may be overwhelming. Dealing with incomplete datasets implies that the information recovered from choices may not be uniquely pinned down resulting in (potential) prediction uncertainty. I refer to this source uncertainty as 'model uncertainty'. This multiplicity results in low power when testing the theory, since the relative size of consistent behavior may not be restrictive enough. Boccardi (2017) shows that model uncertainty reflects the likelihood to detect violations to the model if behavior is not generated by it, conditional on observed behavior. This is important because standard measures of power fail to account for the actual pattern of choices observed, masking significant differences on the amount of information that can be learned from data. Intuitively, consider a decision maker choosing from $\{a, b, c\}$ and $\{a, c\}$ and assume utility maximization model. This design is ex-ante powerful since the probability of detecting a violation if choices were to be generated uniformly at random is $\frac{1}{3}$. However, if b is chosen from $\{a, b, c\}$ then any choice from $\{a, c\}$ would be consistent. An ex-ante assessment of power is incapable of capturing this difference, while a conditional power assessment is meant to capture it.

To measure the predictive performance of the model, I first recover the revealed preference information from data by minimizing an Apesteguia and Ballester (2015)-like measure of incon-

sistency. Then, I predict behavior in each menu for each possible choice rule that is consistency with the recovered revealed preference information. The (potential) multiplicity of choice rules is incorporated in predictions by assuming that each such rule is equally likely to be the true underlying process that generated behavior, reflecting model uncertainty. The difference between predicted and observed choices provides an estimate for the error process that rationalizes such choice rule, reflecting a second source of uncertainty due to possible deviations from the considered theory. Therefore, this construction, reflects two sources of uncertainty when predicting behavior: (i) 'model uncertainty' and (ii) 'error uncertainty'. I measure the performance of the model given observed behavior as a negative function of the uncertainty to predict behavior in each possible menu given observed choices.

I show the performance of the proposed measure in simulation exercises. The predictive precision measure reflects both the severity of the observed deviations (as a reflection of the assumed error process) and the amount of information that can be inferred from data. Additionally, this measure has power to differentiate rational behavior with errors from random behavior as long as there is enough information to identify underlying preferences (more than 25% of possible menus are observed). Finally, I show that the recovered errors are significantly different from a random process for the observed menus.

For the data on choices under risk from Harless and Camerer (1994) I show how the proposed approach can be used for model selection. The uncertainty measure relates to the approach adopted by the authors by favoring more 'parsimonious' models, i.e. models that restrict the pattern of choices that they allow for.

By allowing for multiple choices from the same menu to be observed with different frequencies and accounting for these when inferring information from data, the proposed approach also accommodates for theories that rely on a stochastic choice structure of the data. Finally, I discuss how the proposed approach can be used for bounded rationality models and for the comparison between these and the utility maximization model; addressing some of the concerns raised by De Clippel and Rozen (2014) for bounded rationality models by studying their empirical performance.

Outline Next section introduces the predictive ability approach for the context of discrete choice environments by constructing the predictive distribution. Section 2 also proposed a measure of the predictive ability of the model based on the information content of predictive distribution. Section 4 shows the application of the proposed approach to the data gathered in Harless and Camerer (1994) for models of generalized expected utility. Section 5 shows the power of the proposed measures in a simulation exercise for different error processes and probability of observing choices from any given menu. Section 3 shows the extension of the predictive approach to bounded rationality models, showing that, the proposed measures can aid on the selection of the theory to

model observed behavior. Section 7 offers a brief summary of the relevant literature.

2 Predictive Ability

A decision maker (DM) faces a series of decision problems in the form of a menu of alternatives and she must choose one alternative from the menu. She is assumed to be an utility maximizer: presented with a menu, she chooses the maximal alternative given her own preference order. As expected, preferences are not observed but choices are. Adopting a revealed preference approach, the model imposes acyclicality of the preference information revealed from data. In this section I develop the predictive ability methodology for the utility maximization model; but this approach extends to other behavioral models.

2.1 Set-up

Let $\mathbf{X}$ be the universe of alternatives with $|\mathbf{X}| = k < \infty$. The DM faces a sequence of decision problems where she must choose an alternative from a set $A \subseteq \mathbf{X}$ with $|A| \geq 2$ and let $2^{\mathbf{X}*} \equiv \left\{A \in 2^{\mathbf{X}} : |A| \geq 2\right\}$ be the set of all such possible menus.[1] The behavior of a DM is summarized by a collection of observations f; where $f(a, A)$ denotes de frequency with which the DM faced menu A and chose a. I define $\mathcal{O}$ as the collection of all possible observations, i.e. $\mathcal{O} \equiv \left\{(a, A) \text{ such that } a \in A \text{ and } A \subseteq 2^{\mathbf{X}*}\right\}$. This set-up explicitly allows for incomplete data sets.

The utility maximization model establishes that there exists a preference relation P that is a linear order on $\mathbf{X}$, that is a asymmetric, transitive and complete binary relation. The empirical implications of the model then can be summarized on the set $\mathcal{P}$ of all possible linear orders on $\mathbf{X}$. A collection of observations f is said to be *rationalizable* given $\mathcal{P}$ if all observations in the data can be generated by the maximization of the same preference relation, that is, if there exists a $P \in \mathcal{P}$ such that $f(a, A) > 0$ implies that aPx for all $x \in A$. Formally,

Definition 1 (Rationalizable) *Let $\mathbf{X}$ be the universe of alternatives with $|\mathbf{X}| = k < \infty$ and let $\mathcal{P}$ be the set of all linear orders on $\mathbf{X}$. Let f be the observed collection of observations. f is said to be rationalized by a preference relation $P \in \mathcal{P}$ if $f(a, A) > 0$ implies that aPx for all $x \in A$ and all $A \in \mathcal{A}$.*

For incomplete data sets, there may be more than one preference relation that rationalizes the data. The degree of identification of preferences can be characterized in terms of the cardinality of $\mathcal{P}(f)$. Formally,

[1] Note that excluding the empty set and all menus of one element can be done without loss of generality for any complete theory, and it is done since no information can be inferred from such observations. Note that the presented approach is can be equivalently employed for a complete data set $\bar{f}$ where $\bar{f}(\emptyset, \emptyset) = 0$ and $f(a, A) = \frac{\bar{f}(a,A)}{\sum_{a \in \mathbf{X}} \bar{f}(a,a)}$.

Definition 2 (Identification of Preferences) *Let $\mathcal{P}(f)$ be the preference information inferred from data,*

$$\mathcal{P}(f) \equiv \{P \in \mathcal{P} : f(a, A) > 0 \Rightarrow aPx \text{ for all } x \in A, \text{ for all } A \in \mathcal{A}\}$$

then

1. *If $|\mathcal{P}(f)| = 1$, preferences are **'just identified'** or **'uniquely identified'**;*
2. *If $|\mathcal{P}(f)| > 1$, then preferences are **'not uniquely identified'** or there is a **'multiplicity'** of binary relations that rationalize the data; and*
3. *If $|\mathcal{P}(f)| = 0$, observed data is **'not rationalizable'** or **'inconsistent'**.*

Remark 1 *If $\mathcal{A} = 2^{\mathbf{X}*}$, then $|\mathcal{P}(f)| \leq 1$.[2] However, if $\mathcal{A} \subset 2^{\mathbf{X}*}$ then $|\mathcal{P}(f)| \in \{0, 1, 2, \dots\}$.*

If preferences are just identified, predictions are unique for any menu. If $|\mathcal{P}(f)| > 1$ there is uncertainty when predicting behavior for some unobserved menus for which different linear orders in $\mathcal{P}(f)$ select different alternatives as maximals. I refer to this uncertainty as 'model uncertainty' and it is studied in Section 2.2. The case when data is not rationalizable, $|\mathcal{P}(f)| = 0$, is addressed in Section 2.3 extending the approach to allow for errors.

2.2 Model Uncertainty

Let the collection of observations be rationalizable ($\mathcal{P}(f) \neq \emptyset$). Then, there is a (potentially incomplete) binary relation that can directly be inferred from data f. Formally,

Definition 3 (Binary relation inferred from f) *The binary relation inferred from f, $R(f)$, is given by*

$$R(f) \in \mathbf{X} \times \mathbf{X} : (x, y) \in R(f) \Leftrightarrow f(x, A) > 0 \text{ for } x, y \in A \text{ and some } A \in \mathcal{A}$$

Note that $R(f)$ needs not to be an element of $\mathcal{P}$ since it (i) may not be complete; and/or (ii) may not be consistent with the model. Then, a collection of observations f is rationalizable in $\mathcal{P}$ if there exists an extension[3], $R^{ext}(f)$, such that $R^{ext}(f) \in \mathcal{P}$. Let $\mathcal{R}^{ext}(f)$ be the set of all such extensions. Each order in $\mathcal{R}^{ext}(f)$, predicts that the chosen alternative is maximal according to that order, that is

$$\hat{f}(a, A) = \hat{f}_{2^{X*}}(A) \times \mathbb{1}\left[a \; R^{ext}(f) \; b \;\; \forall b \in A\right] \tag{1}$$

[2] Observing all subsets of X is a strong condition, note milder conditions may suffice depending on the theory under condition. For example, for utility maximization this condition can be swap to a condition imposing that all menus of two and three elements are observed.

[3] R^{ext} is an *'extension'* of R if $R \subseteq R^{ext}$.

where $\hat{f}_{2^{X*}}(A)$ is the expected frequency of menu A in $\mathcal{O}$. Unless further information is provided, I assume that each menu in 2^{X*} has, when constructing predictions, ex-ante equal likelihood. Formally,

Assumption 1 (Uniform distribution across menus) *All menus $A \in 2^{\mathbf{X}*}$ are 'equally likely' for the sake of predictions, i.e. $\hat{f}_{2^{X*}}(A) = \frac{1}{2^{|\mathbf{X}|}-|\mathbf{X}|-1} = \gamma_{\mathbf{X}}$.*

Definitions are presented for the general case unless otherwise stated. If there exists a unique extension to the preference information inferred from data that rationalizes behavior, $|\mathcal{R}^{ext}(f)| = 1$, predictions are uniquely defined by equation (1) for all menus. Otherwise predictions are not unique for some menu A. Consider the following example.

Example 1 (Predictions with 'multiplicity' of preferences that rationalize f) *Let $\mathbf{X} \equiv \{a, b, c\}$ and $f(a, \{a, b\}) = \frac{1}{2}$, $f(c, \{b, c\}) = \frac{1}{2}$. Given f, $|\mathcal{R}^{ext}(f)| = 2$ since $R(f) \equiv \{(a, b); (c, b)\}$; and there exists two possible extensions $R_1^{ext}(f) = \{(a, b); (c, b); (a, c)\}$ and $R_2^{ext}(f) = \{(a, b); (c, b); (c, a)\}$. Consider the predictions for $A = \{a, c\}$, $\hat{f}^{R_1^{ext}(f)}(a, A) = \hat{f}_{2^{X}*}(A) > 0$; while $\hat{f}^{R_2^{ext}(f)}(a, A) = 0$.*

Without additional information, I assume that all possible extensions of the revealed preference information are equally likely. Formally,

Definition 4 (Predictions for $2^{\mathbf{X}*}$) *Given a collection of observations f, predictions are given by*

$$\hat{f}^{R(f)}(a, A) = \frac{\gamma_f(a, A)}{r} \times \hat{f}_{2^{\mathbf{X}*}}(A) \tag{2}$$

for all $a \in A$ with $A \in 2^{\mathbf{X}*}$ and where $\gamma_f(a, A) \equiv \sum_{R_i \in \mathcal{R}^{ext}(f)} \mathbb{1}[a \ R_i \ b \ \forall \ b \in A \setminus \{a\}]$, with $\sum_{A \in 2^{X*}} \hat{f}_{2^{\mathbf{X}*}}(A) = 1$ and $r = |\mathcal{R}^{ext}(f)|$.

Note that $\frac{\gamma_f(a,A)}{r}$ is the relative likelihood of a being maximal according to some well behaved preference relation in A that extends the information inferred from the observed collection f. If no order in $\mathcal{R}^{ext}(f)$ sets $a \in A$ as the maximal element in A, then $\gamma_f(a, A) = 0$. On the other end, if all extensions set a as the maximal element in A then $\gamma_f(a, A) = r$. The following example illustrates this construction.

Example 2 (Predictions with $R^T(f)$ not complete) *Let $\mathbf{X} = \{a, b, c, d, e\}$ and let $f(a, \{a, b, c\}) = \alpha$ and $f(d, \{a, c, d, e\}) = 1 - \alpha$ with $\alpha > 0$. The inferred preference relation from data is given by $R(f) = \{(a, b); (a, c); (d, a); (d, c); (d, e)\}$ and its transitive extension is given by $R^T(f) = \{(a, b); (a, c); (d, a); (d, c); (d, e); (\mathbf{d}, \mathbf{b})\}$, which is not complete. Then*

- $\hat{f}(d, A_d) = \hat{f}_{2^{\mathbf{X}*}}(A_d)$ *with* $A_d = \{A \in 2^{\mathbf{X}*} : d \in A\}$

- $\hat{f}(a, A_{a*}) = \hat{f}_{2^{\mathbf{X}*}}(A_{a*})$ *with* $A_{a*} = \{A \in 2^{\mathbf{X}*} : a \in A$ *and* $\{d, e\} \cap A = \emptyset\}$

However predictions for all other subsets are not unique. Out of the 120 possible linear orders on $\mathbf{X}$ *only 8 extend* $R^T(f)$*; i.e.* $|\mathcal{R}^{ext}(f)| = 8$.[4] *Consider for example* $A = \{b, c, e\}$*, predictions are given by,*

$$\hat{f}^{R(f)}(b, A) = \frac{1}{4}\hat{f}_{2^{\mathbf{X}*}}(A); \quad \hat{f}^{R(f)}(c, A) = \frac{1}{4}\hat{f}_{2^{\mathbf{X}*}}(A) \quad \textit{and} \quad \hat{f}^{R(f)}(e, A) = \frac{1}{2}\hat{f}_{2^{\mathbf{X}*}}(A).$$

2.2.1 'Model Uncertainty'

Definition 4 provides the machinery to construct predictions over all possible menus. If preferences are just identified, then there is no model uncertainty. On the other extreme, there is complete uncertainty or no information to predict behavior when all alternatives in each menu are predicted to be seen with the same probability; that is $\hat{f}(x, A) = \frac{1}{|A|}\gamma_{\mathbf{X}}$ for all $x \in A$ and all $A \in 2^{\mathbf{X}*}$. This can only occurs if $f = \emptyset$. Any observation provides information to narrow down predictions. The following example illustrates how adding more observations reduces predictive uncertainty.

Example 3 (From no information to zero model uncertainty) *Let* $\mathbf{X} \equiv \{a, b, c\}$*. Under assumption 1, if* $f = \emptyset$ *then*

$$\hat{f}(a, A) = \frac{1}{|A|}\gamma_{\mathbf{X}} = \frac{1}{|A|}\frac{1}{4} \quad \forall a \in A \textit{ and all } A \in 2^{\mathbf{X}*}$$

Any observation improves predictions by reducing $|R^{ext}(f)|$*. For example, consider* $f(a, \{a, b\}) = 1$*, and* $f(x, A) = 0$ *for all* $x \in A$ *and all* $A \in 2^{\mathbf{X}*}$*. Under assumption 1 predictions update to*

- $\hat{f}(a, \{a, b\}) = \frac{1}{4}$ *and* $\hat{f}(b, \{a, b\}) = 0$*;*
- $\hat{f}(x, A) = 0$ *for* $x \in A$ *and* $A \in \{\{a, c\}; \{b, c\}\}$ *and*
- $\hat{f}(a, \{a, b, c\}) = \frac{2}{3}\frac{1}{4} = \frac{1}{6}$*,* $\hat{f}(c, \{a, b, c\}) = \frac{1}{3}\frac{1}{4} = \frac{1}{12}$ *and* $\hat{f}(b, \{a, b, c\}) = 0$

On the other extreme consider

$$f' \equiv \left\{ f(a, \{a, b\}) = \frac{1}{3}; f(a, \{a, c\}) = \frac{1}{3}; f(b, \{b, c\}) = \frac{1}{3} \right\}$$

The only complete and transitive preference relation that rationalizes these observations is given by $R^{ext}_f = (a, b, c)$*; and predictions are given by*

[4] Namely these linear orders are given by $R^{ext}_1(f) = (d, e, a, b, c)$, $R^{ext}_2(f) = (d, e, a, c, b)$, $R^{ext}_3(f) = (d, a, e, b, c)$, $R^{ext}_4(f) = (d, a, e, c, b)$, $R^{ext}_5(f) = (d, a, b, c, e)$, $R^{ext}_6(f) = (d, a, b, e, c)$, $R^{ext}_7(f) = (d, a, c, b, e)$ and $R^{ext}_8(f) = (d, a, c, e, b)$.

- $\hat{f}(a,A)=\frac{1}{4}$ *and* $\hat{f}(x,A)=0$ *for all* $A\in 2^{\mathbf{X}*}$ *such that* $a\in A$ *and* $x\in A\setminus\{a\}$ *and*
- $\hat{f}(b,\{b,c\})=\frac{1}{4}$ *and* $\hat{f}(c,\{b,c\})=0$

The example illustrates how as the number of observations increases – measured by $\frac{|\left\{A\in 2^{\mathbf{X}*}:\sum_{a\in A}f(a,A)>0\right\}|}{2^{\mathbf{X}}-|\mathbf{X}|-1}$–, the uncertainty to predict behavior within menus (weakly) diminishes. The uncertainty for predicting behavior for each given menu can be size by the entropy of the predictive distribution, that is,

$$\begin{aligned} Entropy(\hat{f}^{R(f)}|A) &= -\frac{1}{\hat{f}_{2^{\mathbf{X}*}}(A)}\left[\sum_{a\in A}\hat{f}^{R(f)}(a,A)ln\left(\hat{f}^{R(f)}(a,A)\right)\right]+ln\left(\hat{f}_{2^{\mathbf{X}*}}(A)\right)\\ &= -\sum_{a\in A}\frac{\gamma_f(a,A)}{r}\ln\left(\frac{\gamma_f(a,A)}{r}\right)\end{aligned}$$

Example (Example 3 continued) *For Example 3 the case of 'no information' corresponds to the case of 'maximal entropy', i.e. if* $\hat{f}^{R(f)}(a,A)=\frac{1}{|A|}\gamma_{\mathbf{X}}$ *for all* $a\in A$*, then* $\hat{f}^{R(f)}(a,A|A)=\frac{\hat{f}(a,A)}{\hat{f}_{2^{\mathbf{X}*}}(A)}=\frac{1}{|A|}$ *is a uniform distribution over the alternatives in menu* A*, and its entropy is maximal, since* $Entropy(\hat{f}^{R(f)}|A)=H(\hat{f}^{R(f)}|A)=\sum_{a\in A}\frac{1}{|A|}ln(|A|)=ln(|A|)$.

When preferences are just identified, then $\hat{f}^{R(f)}(a,A)=\hat{f}_{2^{\mathbf{X}*}}(A)$ *for* $a\in A$ *maximal according to inferred preferences, and* $\hat{f}(x,A)=0$ *for* $x\in A\setminus\{a\}$*. In this case the entropy is zero, since* $Entropy(\hat{f}^{R(f)}|A)=H(\hat{f}^{R(f)}|A)=\sum_{x\in A\setminus\{a\}}0\times ln(0)+1\times ln(1)=0$.

An intermediate case is, for example, when $f(a,\{a,b\})=1$*. For this case* $H(\hat{f}^{R(f)}|\{a,b\})=0$*,* $H(\hat{f}^{R(f)}|\{a,c\})=H(\hat{f}^{R(f)}|\{b,c\})=ln(|A|)=ln(2)$ *while* $H(\hat{f}^{R(f)}|\{a,b,c\})=-\frac{2}{3}ln\left(\frac{2}{3}\right)-\frac{1}{3}ln\left(\frac{1}{3}\right)=ln(3)-\frac{2}{3}ln(2)<ln(3)$.

Overall uncertainty is defined as the entropy of the predictive distribution over $\mathcal{O}$.

Definition 5 ('Overall Entropy') *Let* f *be a collection of observations and let* $\hat{f}^{R(f)}$ *be the predictive distribution as given in Definition 4 for any menu* $A\in 2^{\mathbf{X}*}$*. Then the 'Overall Entropy' is defined as*

$$H(\hat{f}^{R(f)})\equiv Entropy(\hat{f}^{R(f)})=-\sum_{A\in 2^{\mathbf{X}*}}\sum_{a\in A}\hat{f}^{R(f)}(a,A)ln\left(\hat{f}^{R(f)}(a,A)\right).$$

Proposition 1 shows the relation between the entropy of the predictive distribution for each menu A, and the overall entropy of the predictive distribution.

Proposition 1 ('Menu Entropy' and 'Overall Entropy') *Let $\hat{f}^{R(f)}$ be the predicted distribution of choices constructed as in Definition 4 Then,*

$$\begin{aligned} H(\hat{f}^{R(f)}) &= E_{\hat{f}_{2^{\mathbf{X}_*}}}\left[H(\hat{f}^{R(f)}|A)\right] - E_{\hat{f}_{2^{\mathbf{X}_*}}}\left[\ln(\hat{f}_{2^{\mathbf{X}_*}})\right] \\ &= E_{\hat{f}_{2^{\mathbf{X}_*}}}\left[H(\hat{f}^{R(f)}|A)\right] + H(\hat{f}_{2^{\mathbf{X}_*}}) \end{aligned}$$

where $E_{\hat{f}_{2^{\mathbf{X}_*}}}[\cdot]$ it is the expectation with respect to the probability distribution over menus for the sake of prediction, i.e. $E_{\hat{f}_{2^{\mathbf{X}_*}(A)}}[X(A)] = \sum_{A\in 2^{\mathbf{X}_*}} \hat{f}_{2^{\mathbf{X}_*}} X(A)$.

Otherwise stated all proofs are relegated to Appendix A.

Under Assumption 1, Equation (3) collapses to

$$H(\hat{f}^{R(f)}) = \gamma_{\mathbf{X}}\left[\sum_{A\in 2^{\mathbf{X}_*}}\sum_{a\in A}\frac{\gamma_f(a,A)}{r}\ln\left(\frac{\gamma_f(a,A)}{r}\right)\right] + \ln\gamma_{\mathbf{X}}$$

Note that, as defined above, the Overall Entropy reflects $H(\hat{f}_{2^{\mathbf{X}_*}}(A))$ which is not in itself a source of uncertainty. This is due to the structure of the construction of the predictive distribution to reflect the uncertainty to predict behavior for all menus. Moreover, $H(\hat{f}_{2^{\mathbf{X}_*}}(A))$ changes with the number of alternatives in $\mathbf{X}$ irrespective of the information contained in the collection of observations f. I normalize the entropy to only reflect changes in $E_{\hat{f}_{2^{\mathbf{X}_*}}}\left[H(\hat{f}^{R(f)}|A)\right]$ and lie in the interval $[0,1]$. Formally,

Definition 6 ('Normalized Entropy') *Let the conditions be given as in Definition 5. The 'Normalized Entropy' is given by*

$$NE_{\mathbf{X}}(\hat{f}^{R(f)}) = \frac{H(\hat{f}^{R(f)}) - H(\hat{f}_{2^{\mathbf{X}_*}})}{ME_{\mathbf{X}} - H(\hat{f}_{f=\emptyset}(A))} = \frac{E_{\hat{f}_{2^{\mathbf{X}_*}(A)}}\left[H(\hat{f}^{R(f)}|A)\right]}{E_{\hat{f}_{f=\emptyset}}(A)\left[H(\hat{f}_{f=\emptyset}|A)\right]}$$

where $ME_{\mathbf{X}} = H(\hat{f}_{f=\emptyset})$ and

$$\hat{f}_{f=\emptyset}(a,A) = \frac{1}{|A|}\gamma_{\mathbf{X}}$$

for all $(a,A)\in\mathcal{O}$.

Remark 2 *Under Assumption 1*

$$NE_{\mathbf{X}}(\hat{f}^{R(f)}) = \frac{\sum_{A\in 2^{\mathbf{X}_*}} H(\hat{f}^{R(f)}|A)}{\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} ln\,(n)}$$

with $C_n^{|\mathbf{X}|} = \frac{|\mathbf{X}|!}{n!(|\mathbf{X}|-n)!}$.

Normalized entropy reflects changes in the per-menu-entropy, and their impact depends on the size of the grand set $\mathbf{X}$; since

$$\frac{\partial NE_{\mathbf{X}}(\hat{f}^{R(f)})}{\partial H(\hat{f}^{R(f)}|A)} = \frac{1}{E_{\hat{f}_{f=\emptyset}}(A)\left[H(\hat{f}_{f=\emptyset}|A)\right]^2} > 0$$

for all $A \in 2^{\mathbf{X}*}$.

Additional information that reduces the entropy of one or more menus – in the sense that reduces the cardinality of the set of extensions– reduces the overall entropy. On one extreme, if there is no 'model uncertainty' then $NE_{\mathbf{X}}(\hat{f}^{R(f)})|_{\text{'no uncertainty'}} = 0$; on the other extreme if there is 'no information' then $NE_{\mathbf{X}}(\hat{f}^{R(f)}) = 1$. More information either keeps $\mathcal{R}^{ext}(f)$ unaltered or reduces it cardinality by discarding some preferences relations that are not longer consistent with the observed collection f, therefore reducing uncertainty.

Proposition 2 (Normalized Entropy) *Let f be a collection of observations, and $\hat{f}^{R(f)}$ be the predictive distribution as in Definition 4 and let $NE_{\mathbf{X}}(\hat{f}^{R(f)})$ be the normalized predictive uncertainty as in Definition 6. Then,*

$$NE_{\mathbf{X}}\left(\hat{f}^{R(f)}\right) \in [0,1]$$

2.3 Error Uncertainty

Empirical evidence finds violations supporting the presence of errors/mistakes in consumers' decisions.[5] Errors may occur because of the lack of attention, misunderstanding of the choice problem, mistakes when selecting the desire alternative, inability to implement the optimal choice among others. In this paper, I abstract from the source of the error but assume that they may occur with certain probability. The question remains with respect to how significant and systematic the observed deviations are and whether they support or not rationality as the model of behavior. In this section, I extend the approach to allow for error, under the assumption that errors are such that they assign decreasing probability to alternatives worse ranked by the underlying preference relation. Formally,

Assumption 2 (Error structure) *Let $\mathbf{R} \subseteq \mathbf{X} \times \mathbf{X}$ be a complete and transitive preference relation that generates behavior. For each $A \in 2^{\mathbf{X}*}$, let $g^{\mathbf{R}}(\cdot, A) : A \to \{1, \ldots, |A|\}$ be the (injective) rank function, i.e. $g^{\mathbf{R}}(a, A) = 1 \Leftrightarrow a = m(\mathbf{R}, A)$, $g^{\mathbf{R}}(a, A) = 2 \Leftrightarrow a = m\left(\mathbf{R}, A \setminus m(\mathbf{R}, A)\right)$,*

[5] From the empirical point of view, several studies have shown the existence of GARP violations in different settings, cross section data (Blundell *et al.* (2003), Beatty and Crawford (2011) and Echenique *et al.* (2011), among others); laboratory experiments (Andreoni and Miller (2002), Sippel (1997), Choi *et al.* (2007), among others); as well as in field experiments (for example in Harbaugh *et al.* (2001)).

etc; where $x = m(R, X) \Leftrightarrow xRa$ *for all* $a \in X \setminus \{x\}$. *Then, for all* $A \in 2^{\mathbf{X}*}$ ***'error process'*** *is assumed to be such that* $g^{\mathbf{R}}(a, A) < g^{\mathbf{R}}(b, A)$ *implies* $\eta^{\mathbf{R}}_{(a,A)} \geq \eta^{\mathbf{R}}_{(b,A)}$; *where* $\eta^{\mathbf{R}}_{(a,A)}$ *is the probability of* a *being chosen from set* A. *Then, the expected frequency distribution over menu* A *given preference* $\mathbf{R}$ *is given by*

$$E(f(a, A)^{\mathbf{R}|model}|A) = \eta^{\mathbf{R}}_{(a,A)}$$

with $\sum_{a \in A} \eta^{\mathbf{R}}_{(a,A)} = 1$ *for all* $A \in 2^{\mathbf{X}*}$.

Assumption 2 implicitly defines the (discrete) distribution of the error process. Note that, with probability $\eta^{\mathbf{R}}_{m(\mathbf{R},A),A}$ the error is equal to zero, that is, the decision maker chooses the alternative that maximizes her underlying utility function. Errors are defined as the magnitude of the rank deviations from optimal choice. Therefore, its maximum value for menu A is $|A| - 1$ and this is observed with the lowest probability among all possible values for the error process. In particular, the implied distribution for the error process by Assumption 2 is given by,

$$\begin{aligned}
\varepsilon &= 0 \text{ with probability } \eta^{\mathbf{R}}_{(m(\mathbf{R},A),A)} \\
&\vdots \\
&= g^{\mathbf{R}}(a, A) - 1 \quad \text{with probability } \eta^{\mathbf{R}}_{(a,A)} \\
&\vdots \\
&= |A| - 1 \text{ with probability } \eta^{\mathbf{R}}_{(b,A)} \text{ where } a \ \mathbf{R} \ b \text{ for all } a \in A \setminus \{b\}.
\end{aligned}$$

Assumption 2 relies on menus being countable and does not require a budget structure. Under this assumption, the expectation of the errors within menu weighted by the probability used for prediction sake is given by,

$$\begin{aligned}
\mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}}\left[\varepsilon_{(a,A)}\right] &= \sum_{A \in 2^{\mathbf{X}*}} \hat{f}_{2^{\mathbf{X}*}}(A) \sum_{a \in A} \left[g^{\mathbf{R}}(a, A) - 1\right] \eta^{\mathbf{R}}_{(a,A)} \\
&= \sum_{(a,A) \in \mathcal{O}} |\{x \in A : x\mathbf{R}a\}| \left(\hat{f}_{2^{\mathbf{X}*}}(A) \times \eta^{\mathbf{R}}_{(a,A)}\right)
\end{aligned}$$

Example 4 *Let* $\mathbf{X} = \{a, b, c\}$ *and let* $\mathbf{R}$ *be the underlying complete and preference relation such that* $a\mathbf{R}b\mathbf{R}c$. *Assume that* $\hat{f}_{2^{\mathbf{X}*}}(A) = \frac{1}{4}$ *for all* $A \in 2^{\mathbf{X}*}$, *and let the error process be such that,*

$$\begin{aligned}
\eta^{\mathbf{R}}_{(a,\{a,b,c\})} &= \frac{5}{8}; \quad \eta^{\mathbf{R}}_{(b,\{a,b,c\})} = \frac{1}{4}; \quad \eta^{\mathbf{R}}_{(c,\{a,b,c\})} = \frac{1}{8}; \\
\eta^{\mathbf{R}}_{(a,\{a,b\})} = \eta^{\mathbf{R}}_{(b,\{b,c\})} = \eta^{\mathbf{R}}_{(a,\{a,c\})} &= \frac{5}{7}; \quad \eta^{\mathbf{R}}_{(b,\{a,b\})} = \eta^{\mathbf{R}}_{(c,\{b,c\})} = \eta^{\mathbf{R}}_{(a,\{a,c\})} = \frac{2}{7}
\end{aligned}$$

Then

$$\begin{aligned}\mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}(A)}\left[\varepsilon_{(a,A)}\right] &= \hat{f}_{2^{\mathbf{X}*}}(\{a,b,c\})\left[0\times\frac{5}{8}+1\times\frac{1}{4}+2\times\frac{1}{8}\right]+\hat{f}_{2^{\mathbf{X}*}}(\{a,b\})\left[0\times\frac{5}{7}+1\times\frac{2}{7}\right]\\ &+ \hat{f}_{2^{\mathbf{X}*}}(\{b,c\})\left[0\times\frac{5}{7}+1\times\frac{2}{7}\right]+\hat{f}_{2^{\mathbf{X}*}}(\{a,c\})\left[0\times\frac{5}{7}+1\times\frac{2}{7}\right]=\frac{19}{46}\end{aligned}$$

The definition for the error process in Assumption 2 coincides with the 'tremble model' discussed in Apesteguia and Ballester (2015) by f_{TM-pro} allowing for menu specific trembles, under the assumption that the maximal alternative in the menu is chosen with probability greater than $\frac{1}{2}$. The alternative specifications discussed in Apesteguia and Ballester (2015) can be alternatively assumed and the identification results and construction proposed in Section 2.4 can be extended to contemplate these alternative error structures.

2.3.1 Identification of Underlying Preferences

Underlying preferences can be recovered as the 'Swaps Preference' defined by Apesteguia and Ballester (2015). Consider the expression for the expectation of the error process now as a function of a preference relation $\mathbf{R}$.

$$\mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}}\left[\varepsilon^{\mathbf{R}}_{(a,A)}\right]=\sum_{A\in 2^{\mathbf{X}*}}\sum_{a\in A}|\{x\in A: x\mathbf{R}a\}|\left(\hat{f}_{2^{\mathbf{X}*}}(A)\times\eta^{\mathbf{R}}_{(a,A)}\right) \tag{3}$$

Under the assumption that all menus are observed with equal probability, i.e. $\hat{f}_{2^{\mathbf{X}*}}(A)=\gamma_{\mathbf{X}}=\frac{1}{2^{|\mathbf{X}|}-|\mathbf{X}|-1}$ for all menus A,

$$\begin{aligned}\mathbb{E}_{f_{2^{\mathbf{X}*}}}\left[\varepsilon^{\mathbf{R}}_{(a,A)}\right] &= \gamma_{\mathbf{X}}\sum_{A\in 2^{\mathbf{X}*}}\sum_{a\in A}\left(g^{\mathbf{R}}(a,A)-1\right)\eta^{\mathbf{R}}_{(a,A)}\\ &= \gamma_{\mathbf{X}}\left(\sum_{A\in 2^{\mathbf{X}*}}\sum_{a\in A}g^{\mathbf{R}}(a,A)\times\eta^{\mathbf{R}}_{(a,A)}\right)-1\end{aligned}$$

Assume now that a researcher has observed the collection f generated by $\mathbf{R}$ plus an error process that satisfies Assumption 2; but mistakenly infers that the underlying process generating the data is $\mathbf{R}'$ with $\mathbf{R}'\neq\mathbf{R}$. Given assumption 2, the researcher is assigning higher probability to bigger errors. Fixing the observed collection of observations f, I show that any change to the estimated preference order changes the estimated error, affecting the term $\sum_{a\in A}g^{\mathbf{R}}(a,A)\times\eta^{\mathbf{R}}_{a,A}$ in the above expression through $g^{\mathbf{R}}(a,A)$. Consider the following example.

Example (Example 4 continued) *Assume the conditions of Example 4 and let* $\hat{f}_{2^{\mathbf{X}*}}(A) = \frac{1}{4}$ *for all* $A \in 2^{\mathbf{X}*}$*. In this case the true underlying preferences* $\mathbf{R}$ *is such that* $a\mathbf{R}b\mathbf{R}c$*, and therefore the expected sum of errors is given by, then*

$$\begin{aligned}\mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}}\left[\varepsilon_{(a,A)}\right] &= \sum_{(a,A)\in\mathcal{O}} |\{x \in A : x\mathbf{R}a\}| \left(\hat{f}_{2^{\mathbf{X}*}(A)} \times \eta^{\mathbf{R}}_{(a,A)}\right) \\ &= \frac{1}{4}\left[\eta^{\mathbf{R}}_{(b,\{a,b,c\})} + 2\eta^{\mathbf{R}}_{(c,\{a,b,c\})} + \eta^{\mathbf{R}}_{(b,\{a,b\})} + \eta^{\mathbf{R}}_{(c,\{b,c\})} + \eta^{\mathbf{R}}_{(c,\{a,c\})}\right]\end{aligned}$$

Assume that the researcher mistakenly assumes that the underlying order is given by $\mathbf{R}'$ *such that* $a\mathbf{R}'c\mathbf{R}'b$*, then given the same data generating process, the sum of expected errors is given by*

$$\begin{aligned}\mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}}\left[\varepsilon_{(a,A)'}\right] &= \sum_{(a,A)\in\mathcal{O}} |\{x \in A : x\mathbf{R}a\}| \left(\hat{f}_{2^{\mathbf{X}*}}(A) \times \eta^{\mathbf{R}}_{(a,A)}\right) \\ &= \frac{1}{4}\left[2\eta^{\mathbf{R}}_{(b,\{a,b,c\})} + \eta^{\mathbf{R}}_{(c,\{a,b,c\})} + \eta^{\mathbf{R}}_{(b,\{a,b\})} + \eta^{\mathbf{R}}_{(b,\{b,c\})} + \eta^{\mathbf{R}}_{(c,\{a,c\})}\right]\end{aligned}$$

where the difference is given by

$$\mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}(A)}\left[\varepsilon_{(a,A)'} - \varepsilon_{(a,A)}\right] = \frac{1}{4}\left[\eta^{\mathbf{R}}_{(b,\{a,b,c\})} - \eta^{\mathbf{R}}_{(c,\{a,b,c\})} + \eta^{\mathbf{R}}_{(b,\{b,c\})} - \eta^{\mathbf{R}}_{(c,\{b,c\})}\right]$$

Since the underlying generating process is given by $\mathbf{R}$ *and the error process is consistent with Assumption 2,* $\eta^{\mathbf{R}}_{(b,\{a,b,c\})} \geq \eta^{\mathbf{R}}_{(c,\{a,b,c\})}$ *and* $\eta^{\mathbf{R}}_{(b,\{b,c\})} \geq \eta^{\mathbf{R}}_{(c,\{b,c\})}$ *and thus* $\mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}(A)}\left[\varepsilon_{(a,A)'}\right] \geq \mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}(A)}\left[\varepsilon_{(a,A)}\right]$*.*

Claim 1 (Identification of Preferences) *Let the observed collection* f *being generated from preferences* $\mathbf{R}$ *with an error process consistent with Assumption 2, but the researcher mistakenly assumes that preferences are given by* $\mathbf{R}'$*, then*

$$\mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}}\left[\varepsilon^{\mathbf{R}'}_{(a,A)}\right] - \mathbb{E}_{\hat{f}_{2^{\mathbf{X}*}}}\left[\varepsilon^{\mathbf{R}}_{(a,A)}\right] \geq 0$$

and the inequality is strict if the process $\eta^{\mathbf{X}}$ *is assumed to be strictly decreasing.*

Underlying preferences are therefore identified, and can be estimated from data by minimizing the empirical counterpart of expression (3). Since $\hat{f}_{2^{\mathbf{X}*}}(A) \times \eta^{\mathbf{R}}_{(a,A)}$ can be understood as the expected frequency of a being chosen from A; its empirical counterpart is just $f(a, A)$. Now the empirical counterpart of equation 3 is given by

$$\sum_{(a,A)\in\mathcal{O}} |\{x \in A : x\mathbf{R}a\}| f(a, A) \tag{4}$$

and $\mathbf{R}$ can be estimated as the minimization of this expression. Therefore, Assumption 2 provides the rationale for AB 'swaps index' and 'swaps preference relation', namely

$$I(f) = \min_{\mathbf{R}} \sum_{(a,A)\in\mathcal{O}} f(a,A) \left|\{x \in A : x\mathbf{R}a\}\right| \tag{5}$$

$$\hat{\mathbf{R}}_{\varepsilon}(f) \in arg\min_{\mathbf{R}} \sum_{(a,A)\in\mathcal{O}} f(a,A) \left|\{x \in A : x\mathbf{R}a\}\right| \tag{6}$$

The connection between these measures and others proposed in the literature has been studied in Apesteguia and Ballester (2015).

Example (Example 4 continued) *Assume the conditions of Example 4 and let* $\hat{f}_{2^{\mathbf{X}*}}(A) = \frac{1}{4}$ *for all* $A \in 2^{\mathbf{X}*}$ *and the observed collection* f *as specified above. There are 6 possible linear orders that could rationalize the data; let* R_{ijk} *be the linear order such that* $iRjRk$ *with* $i,j,k \in \mathbf{X}$*. Then the swaps index for each of these candidate preference orders are given by,*

$$\begin{aligned}
I(f)|R_{abc} &= \frac{19}{46}; & I(f)|R_{acb} &= \frac{107}{124} > \frac{19}{46};\\
I(f)|R_{bac} &= \frac{121}{124} > \frac{19}{46}; & I(f)|R_{bca} &= \frac{173}{124} > \frac{19}{46};\\
I(f)|R_{cab} &= \frac{159}{124} > \frac{19}{46}; & I(f)|R_{cba} &= \frac{204}{124} > \frac{19}{46};
\end{aligned}$$

therefore $\hat{\mathbf{R}}_{\varepsilon}(f) = R_{abc}$.

The swaps index defined in equation 5 provides a measure of the inconsistency of the error process, by reflecting the expectation of the error process.[6] Moreover, it can be interpreted as a measure, in sample, of the certainty to predict behavior based on $\hat{\mathbf{R}}_{\varepsilon}(f)$ given that there are some 'mistakes' that have been made. For the just identified case, the swaps theory deliver a unique prediction for each of the observed budget sets. However, observed data may not coincide with the predictions of the data, if $I(f) \neq 0$. By summarizing the level of inconsistency of observed data relying solely on the revealed preference information inferred from data, this index provides a measure of the certainty to predict behavior based on the recovered theory.

The stochastic satisfies $P-$monotonicity as defined in AB if $\eta_{(m(R,A),A)} > \frac{1}{2}$ for all menus; therefore the inferred preference relation is a swaps preference and, if all menus are observed with positive probability, then it is the unique swaps preference of f, as follows from Theorem 1 in Apesteguia and Ballester (2015); and their axiomatization carries through.

[6]Note that since by construction the error process is assumed to be always positive, the mean (expectation) of the error process corresponds to the mean absolute deviation.

2.4 Predictions based on the stochastic extension

The stochastic extension presented above induces a new source of uncertainty which I refer as 'Error Uncertainty'. For each complete and transitive extension of recovered preferences, expected behavior is a probability distribution on the set of alternatives available in the menu. Based on this stochastic extension the predictive distribution should reflect both, the uncertainty due to (potential) partial identification of the underlying preferences –'model uncertainty'–, and the uncertainty due to possible mistakes –'error uncertainty'–.

In case of incomplete data sets we can only recover $\eta_{(a,A)}$ for those menus that are observed. In order to extend the prediction analysis to menus that have not yet been observed, I propose further structure on the error process assuming that the relative probability of observing different magnitudes for the error process only depends on the relative ranking of the alternatives and the size of the menu. Formally,

Assumption 3 (Common relative likelihood of errors (CRLE)) *Let* $\eta^{\mathbf{X}} : \{0,1,\ldots,|\mathbf{X}|-1\} \to [0,1]$ *be an injective and (strictly) decreasing function, i.e.* $\eta^{\mathbf{X}}(n)(>) \geq \eta^{\mathbf{X}}(n+1)$.[7] *Then,*

$$\eta^{\mathbf{R}}_{(a,A)} = \frac{\eta^{\mathbf{X}}(g^{\mathbf{R}}(a,A)-1)}{\sum_{b\in A}\eta^{\mathbf{X}}(g^{\mathbf{R}}(b,A)-1)}$$

where $g^{\mathbf{R}}(a,A)$ *is the ranking function as in Assumption 2. Without loss of generality further assume that* $\sum_{i=0}^{|X|-1}\eta^{\mathbf{X}}(i)=1$.

The estimation can be performed by employing maximum likelihood methods either dependent on the underlying preference relation, or it can be assumed that there is a unique function distribution of the error process irrespective of the preference order considered, case in which $\widehat{\eta}^{\mathbf{X}}_{R,f} = \widehat{\eta}^{\mathbf{X}}_{f}$ for all $R \in \mathcal{R}_{\varepsilon}(f)$, and then pool all the observations. This is particularly recommended when the number of observations is small with respect to the number of parameters to be estimated ($|\mathbf{X}|-1$); and further assumptions may be required.[8]

There is a connection between the Inconsistency index for any recovered preference order and the variance of the process implied by the structure of $\eta^{\mathbf{X}}$; under assumptions 2 and 3,

$$E(I(f)) = \sum_{n=2}^{|\mathbf{X}|}\frac{\beta_n}{\eta^{\mathbf{X}}_n}\sum_{i=1}^{n}(i-1)\eta^{\mathbf{X}}(i-1) \tag{7}$$

[7] Note that $\eta^{\mathbf{X}}(n)$ can be interpreted as the relative likelihood of making errors of size n where the grand set is observed. For this interpretation to follow though you need to assume that $\sum_{i=0}^{|\mathbf{X}|-1}\eta^{\mathbf{X}}(i)=1$; normalization that can be done without loss of generality.

[8] For example, if η_k is not identified and $\sum_{A\in\mathcal{A}}\sum_{i=k+1}^{|\mathbf{X}|} f(x : g^{\mathbf{R}}(i,A),A)=0$, then one can impose $\eta_k,\eta_{k+1},\ldots,\eta_{|X|}=0$. Moreover, before normalization, if η_k is not identified but $\hat{\eta}_{k-1}=\alpha$ and $\hat{\eta}_{k+1}=\beta$ with $\alpha>\beta$, then $\hat{\eta}_k=\frac{\alpha+\beta}{2}$.

where $\beta_n = \sum_{A \in 2^{\mathbf{X}*} | f(A) > 0} \mathbb{1}[|A| = n]$ and $\eta_n^{\mathbf{X}} = \sum_{i=1}^{n} \eta^{\mathbf{X}}(i-1)$. Equation 7 shows that if $\eta^X(\cdot)$ is such that the error process assigns higher probability to worse ranked alternatives (i.e. $\eta^X \to U_X$) then $E(I(f))$ would also be higher.

The following definition combines the estimation of the error process with Definition 4 to construct the predictive distribution based on the stochastic extensions.

Definition 7 (Predictive distribution based on 'Stochastic Extension') *Under Assumption 2, let $\widehat{\eta}^{\mathbf{X}}_{R,f}$ be the estimation of the error process. Let f be the collection of observations, with $f \subseteq \mathcal{O}$; then the predictive distribution over $\mathcal{O}$ is given by,*

$$\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a,A) \;=\; \hat{f}_{2^{\mathbf{X}*}}(A) \left[\sum_{R \in \mathcal{R}^{ext}_\varepsilon(f)} \frac{1}{|\mathcal{R}^{ext}_\varepsilon(f)|} \frac{\widehat{\eta}^{\mathbf{X}}_{R_i,f}(g^{R_i}(a,A)-1)}{\sum_{j=1}^{|A|} \widehat{\eta}^{\mathbf{X}}_{R_i,f}(j-1)} \right]$$

for all $(a,A) \in \mathcal{O}$ with

$$\mathcal{R}^{ext}_\varepsilon(f) = \{R \in \mathcal{P} : \exists R_\varepsilon(f) \subseteq R \text{ with } R_\varepsilon(f) \in \mathcal{R}_\varepsilon(f)\}$$

where

$$\mathcal{R}_\varepsilon(f) = \{\mathbf{R} \subset \mathbf{X} \times \mathbf{X} : \mathbf{R} \in arg\min_{\mathbf{R}} \sum_{(a,A) \in \mathcal{O}} f(a,A) \, |\{x \in A : x\mathbf{R}a\}|\}$$

Example 5 ('Model uncertainty') *Assume the conditions of Example 4 and let $f_{2^{\mathbf{X}*}}(A) = \frac{1}{4}$ for all $A \in 2^{\mathbf{X}*}$. Then, $\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a,A) = f$ with R such that $aRbRc$ and*

$$\widehat{\eta}^{\mathbf{X}}_{R,f}(i) = \begin{cases} \frac{5}{8} & \text{if } i = 0 \\ \frac{1}{4} & \text{if } i = 1 \\ \frac{1}{8} & \text{if } i = 2 \end{cases}$$

and $\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a,A) = f$.

If the observed collection is given by $f'(a,\{a,b\}) = \frac{5}{7}$, $f'(b,\{a,b\}) = \frac{2}{7}$, $f'(a,\{a,c\}) = \frac{5}{7}$ and $f'(c,\{a,c\}) = \frac{2}{7}$; then further assumptions are necessary to estimate $\widehat{\eta}^{\mathbf{X}}_{R,f}$. The preference order recovered is $R_{f'}$ such that $aR_{f'}b$ and $aR_{f'}c$. There are two possible complete and transitive extensions $R^{ext1}_{f'}$ and $R^{ext2}_{f'}$ such that $aR^{ext1}_{f'}bR^{ext1}_{f'}c$ and $aR^{ext2}_{f'}cR^{ext2}_{f'}b$.

Further structure (I) *One possible assumption is that* $\widehat{\eta}^{\mathbf{X}}_{R,f}(2) = 0$ *which results in*

$$\widehat{\eta}^{\mathbf{X}}_{R,f}(i) = \begin{cases} \frac{5}{7} & \text{if } i = 0 \\ \frac{2}{7} & \text{if } i = 1 \\ 0 & \text{if } i = 2. \end{cases}$$

Under this assumption, the predictions for the unseen menus are given by

$$\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(i, \{a,b,c\}) = \begin{cases} \frac{5}{7} f(\{a,b,c\}) & \text{if } i = a \\ \frac{1}{7} f(\{a,b,c\}) & \text{if } i = b \\ \frac{1}{7} f(\{a,b,c\}) & \text{if } i = c \end{cases}$$

and

$$\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(i, \{b,c\}) = \begin{cases} \frac{1}{2} f(\{b,c\}) & \text{if } i = b \\ \frac{1}{2} f(\{b,c\}) & \text{if } i = c \end{cases}$$

Further structure (II) *An alternative assumption is that* $\widehat{\eta}^{\mathbf{X}}_{R,f}(2) = \alpha \widehat{\eta}^{\mathbf{X}}_{R,f}(1)$ *with* $\alpha \in (0,1)$. *In this example, if* $\alpha = .5$ *we recover the exact distribution when data is complete. Then*

$$\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(i, \{a,b,c\}) = \begin{cases} \frac{5}{8} f(\{a,b,c\}) & \text{if } i = a \\ \frac{3}{16} f(\{a,b,c\}) & \text{if } i = b \\ \frac{3}{16} f(\{a,b,c\}) & \text{if } i = c \end{cases}$$

and

$$\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(i, \{b,c\}) = \begin{cases} \frac{1}{2} f(\{b,c\}) & \text{if } i = b \\ \frac{1}{2} f(\{b,c\}) & \text{if } i = c \end{cases}$$

if $\widehat{\eta}^{\mathbf{X}}_{R,f}(2) = \alpha \widehat{\eta}^{\mathbf{X}}_{R,f}(1)$ *and* $\alpha = .5$.

If the collection of observations f is rationalizable without error, then Definition 7 coincides with Definition 4, since $|\mathcal{R}_\varepsilon(f)| = 1$, extending the prediction for rationalizable data to allow for mistakes; reflecting not only Model Uncertainty but also Error Uncertainty. The above definition highlights the two sources of uncertainty when predicting behavior: 'Model Uncertainty' due to many possible partial orders that rationalize the data while allowing for error (as long as the data set is incomplete) and to many possible complete linear orders that extend each partial order recovered from data; and 'Error Uncertainty' due to the chance of observing 'mistakes'. In

particular,

$$\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a,A) = \hat{f}_{2^{\mathbf{X}}*}(A)\left[\sum_{R\in\mathcal{R}^{ext}_\varepsilon(f)} \underbrace{\frac{1}{|\mathcal{R}^{ext}_\varepsilon(f)|}}_{\text{'Model Uncertainty'}} \underbrace{\frac{\widehat{\eta}^{\mathbf{X}}_{R_i,f}(g^{R_i}(a,A)-1)}{\sum_{j=1}^{|A|}\widehat{\eta}^{\mathbf{X}}_{R_i,f}(j-1)}}_{\text{'Error Uncertainty'}}\right] \tag{8}$$

2.5 Measuring Predictive Ability

Previous section extends the predictive distribution to reflect the uncertainty due to observed mistakes. Here I extend the proposed measure to reflect the 'Error uncertainty', and then I propose a (positive) measure of 'Predictive Ability'.

Definition 8 (Uncertainty to Predict Behavior: Model and Error Uncertainty) *Let $\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a,A)$ be the predictive distribution as in Definition 7. The 'Prediction Uncertainty' (PU) that reflects both model and error uncertainty is given by*

$$PU_f \equiv H\left(\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}\right)$$

As in Proposition 1,

$$PU_f = E_{\hat{f}_{2^{\mathbf{X}}*}}\left[H\left(\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}|A\right)\right] + H(\hat{f}_{2^{\mathbf{X}}*})$$

I normalize this measure to remove the dependency with respect to the probability distribution across menus, formally,

Definition 9 (Normalized PU: Model and Error Uncertainty) *Let $\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a,A)$ be the predictive distribution as in Definition 7. The 'Normalized Prediction Uncertainty' (NPU) that reflects both model and error uncertainty is given by*

$$NPU_f \equiv \frac{PU_f - H(\hat{f}_{2^{\mathbf{X}}*})}{H(\hat{f}_{f=\emptyset}) - H(\hat{f}_{2^{\mathbf{X}}*})} = \frac{E_{\hat{f}_{2^{\mathbf{X}}*}}\left[H\left(\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}|A\right)\right]}{\gamma_{\mathbf{X}}\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} ln\,(n)}$$

Even when extending the predictive distribution to allow for error, the maximum entropy corresponds to the case of no information, since it induces a uniform distribution on the menu leading to the maximum level of uncertainty when predicting behavior. On the other hand, the minimum entropy is attained when there is no 'Error Uncertainty' and no 'Model Uncertainty', i.e. $\eta^{\mathbf{X}}(0)=1$, $\eta^{\mathbf{X}}(i)=0$ for all $i>0$ and $|\mathcal{R}^{ext}_\varepsilon(f)|=1$.

Proposition 3 (Properties of PU_f) *Let $\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a,A)$ be the predictive distribution as in Definition 7; and let PU_f the predictive uncertainty measure as in Definition 8, then*

- $\min_{f:\mathcal{O}\to[0,1]} H\left(\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}\right) = H(\hat{f}_{2^{\mathbf{X}*}}) = -ln\left(\gamma_{\mathbf{X}}\right)$ *with* $\gamma_{\mathbf{X}} = \frac{1}{2^{|\mathbf{X}|}-|\mathbf{X}|-1}$. *The minimum is attained when there is no 'Model' and no 'Error Uncertainty', i.e.* $\hat{f}(a,A) = \hat{f}(A)\mathbb{1}\left[a \in A : a\mathbf{R}b\ \forall b \in A \setminus \{a\}\right]$ *for all* $(a,A) \in \mathcal{O}$ *where* $\mathbf{R}$ *is a complete linear order in* $\mathbf{X}\times\mathbf{X}$*; therefore* $\min NPU_f = 0$
- $\max_{f:\mathcal{O}\to[0,1]} H\left(\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}\right) = E_{\hat{f}_{2^{\mathbf{X}*}}}\left[\max H\left(\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}|A\right)\right] + H(\hat{f}_{2^{\mathbf{X}*}}) = \gamma_{\mathbf{X}}\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} ln\left(n\right) - ln\left(\gamma_{\mathbf{X}}\right)$ *with* $\gamma_{\mathbf{X}} = \frac{1}{2^{|\mathbf{X}|}-|\mathbf{X}|-1}$ *and* $C_n^{|\mathbf{X}|} = \frac{|\mathbf{X}|!}{n!(|\mathbf{X}|-n)!}$. *The maximum entropy is attained when there is no information, i.e.* $\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}|A = U_A$*; therefore* $maxNPU_f = 1$.

The measures of predictive uncertainty presented above reflect the amount of information (or lack of) when predicting behavior for all possible economic scenarios. Therefore, the predictive ability of the model given the observed collection f can be measured by a negative function of the predictive uncertainty, that is, the lower is the uncertainty to predict behavior the better the ability of the model when predicting behavior given observed choices.

Definition 10 (Measure of Predictive Ability) *Let NPU_f be the normalized predictive uncertainty measure given in Definition 9. Then, the 'Measure of Predictive Ability' is defined as the complement of the NPU measure, that is*

$$PAM_f = 1 - NPU_f$$

PAM_f reflects both Model and Error uncertainty. One may expect that as 'steepness' of the error process decreases, i.e. $\eta^X(\cdot) \to U_{|\mathbf{X}|}$ the overall uncertainty increases. Given the 'error process', an increase in $|\mathcal{R}^{ext}_\varepsilon(f)|$ is expected to cause an increase in the uncertainty to predict behavior. This intuitively hints to measures of these two sources of uncertainty to estimate their impact on overall uncertainty.

2.5.1 Error Uncertainty

Conditional on the potential multiplicity of complete preferences that rationalize choices, uncertainty to predict behavior is driven by the dispersion of the error process. Intuitively, the 'steeper' the estimated process $\hat{\eta}$ is, the lower the uncertainty when predicting behavior. For example, consider two alternative processes $\eta(i) \in [0,1]$ and $\theta(i) \in [0,1]$ for all $i \in \{0,1,\ldots,n\}$ with $\sum_{i=0}^n \eta(i) = \sum_{i=0}^n \theta(i) = 1$; $\eta(j) = \theta(j) + \epsilon$; $\eta(k) = \theta(k) - \epsilon$ for $j < k$ and $\eta(i) = \theta(i)$ for all $i \neq j,k$. Note that η is a mean preserving spread of θ, therefore, by the concavity of

the logarithmic function we have that $H(\eta) = -E_\eta(\ln(\eta)) > -E_\theta(\ln(\theta)) = H(\theta)$; that is, the 'steeper' the function the lower the entropy.

Definition 11 ('Error Uncertainty Measure') *Let $\hat{\eta}^{\mathbf{X}}(\cdot)$ be the estimated error process that satisfies Assumption 3. For any given menu $A \in 2^{\mathbf{X}*}$, I define the 'Error Uncertainty Measure' as the entropy of the distribution implied by $\eta^{\mathbf{X}}$ on $\{0, \ldots, |A|-1\}$, i.e.*

$$\text{'Error Uncertainty Measure'}_A = -\sum_{i=1}^{|A|} \frac{\eta^{\mathbf{X}}(i-1)}{\sum_{j=1}^{|A|}\eta^{\mathbf{X}}(j-1)} \ln\left[\frac{\eta^{\mathbf{X}}(i-1)}{\sum_{j=1}^{|A|}\eta^{\mathbf{X}}(j-1)}\right]$$

If there is no 'Model Uncertainty', i.e. $|R_\varepsilon^{ext}(f)| = 1$ then the predictive distribution is given by,

$$\begin{aligned}
\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a,A) &= \hat{f}_{2^{\mathbf{X}*}}(A)\left[\sum_{R\in\mathcal{R}_\varepsilon^{ext}(f)} \underbrace{\frac{1}{|\mathcal{R}_\varepsilon^{ext}(f)|}}_{=1} \frac{\widehat{\eta}^{\mathbf{X}}_{R_i,f}(g^{R_i}(a,A)-1)}{\sum_{j=1}^{|A|}\widehat{\eta}^{\mathbf{X}}_{R_i,f}(j-1)}\right] \\
&= \hat{f}_{2^{\mathbf{X}*}}(A)\left[\frac{\widehat{\eta}^{\mathbf{X}}_{\mathbf{R}}(g^{\mathbf{R}}(a,A)-1)}{\sum_{j=1}^{|A|}\widehat{\eta}^{\mathbf{X}}_{\mathbf{R},f}(j-1)}\right]
\end{aligned}$$

and the normalized predictive uncertainty is

$$NPU_f = \frac{E_{\hat{f}_{2^{\mathbf{X}*}}}\{\text{'Error Uncertainty Measure'}_A\}}{\gamma_{\mathbf{X}}\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} ln\,(n)}$$

2.5.2 Model Uncertainty

The other source of uncertainty is driven by the (potential) multiplicity of complete preference relation that rationalize observed data, i.e. $|\mathcal{R}_\varepsilon^{ext}(f)| > 1$.

Definition 12 ('Model Uncertainty Measure') *Let $\mathbf{X}$ be the grand set of alternatives with $|\mathbf{X}| < \infty$ and let $\mathcal{O} \equiv \{(a,A)$ such that $a \in A$ and $A \in 2^{\mathbf{X}^*}\}$. Let f be the collection of observations, with $f \subseteq \mathcal{O}$; and $\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a,A)$ be the predictive distribution as in Definition 7. For any given menu $A \in 2^{\mathbf{X}*}$, I define the 'Model Uncertainty Measure' as the as the entropy of the distribution over preference orders on A implied by the distribution over $\mathcal{R}^{ext}(f)$*

$$\text{'Model Uncertainty Measure'}_A = -\sum_{R_A\in\mathcal{R}\cap A\times A} P(R_A)\ln(R_A)$$

with $P(R_A) = \sum_{R \in \mathcal{R}^{ext}(f)} \frac{1}{|\mathcal{R}^{ext}(f)|} \mathbb{1}[R \cap A \times A \equiv R_A]$

To develop the intuition for this measure note that there are two channels for changes in information to have an effect on predictions: an intensive and an extensive margin. Fixing errors, predictions for any given menu A may change either because (i) there are fewer swaps preferences that rationalize the data with error (extensive margin); and/or (ii) at least one of the swaps preference orders contain more information in the sense that there are fewer complete linear orders that extend it (intensive margin). Taking a menu A for which either of these channels had an impact on the implied distribution over linear orders, such a change moves mass probability from linear orders that now have zero probability to all or a subset of the remaining linear orders in the support; decreasing the entropy of the distribution over linear orders in A.

In the case of no 'Error Uncertainty' that is, if $\eta^{\mathbf{X}}(0) = 1$ and $\eta^{\mathbf{X}}(i) = 0$ for all $i \in \{1, 2, \ldots, |\mathbf{X}| - 1\}$, the predictive distribution is given by,

$$\begin{aligned} \hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}}(a, A) &= \hat{f}_{2^{\mathbf{X}}*}(A) \left[\sum_{R \in \mathcal{R}^{ext}_\varepsilon(f)} \frac{1}{|\mathcal{R}^{ext}_\varepsilon(f)|} \underbrace{\frac{\widehat{\eta}^{\mathbf{X}}_{R_i,f}(g^{R_i}(a, A) - 1)}{\sum_{j=1}^{|A|} \widehat{\eta}^{\mathbf{X}}_{R_i,f}(j-1)}}_{=\mathbb{1}[aR_ib \forall b \in A \setminus \{a\}]} \right] \\ &= \hat{f}_{2^{\mathbf{X}}*}(A) \times P(R_A) \times \mathbb{1}[a \; R_A \; b \; \forall \; b \in A \setminus \{a\}] \end{aligned}$$

and the normalized predictive uncertainty is then given by,

$$\begin{aligned} NPU_f &= \frac{E_{\hat{f}_{2^{\mathbf{X}}*}} \left[H\left(\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}} | A \right) \right]}{\gamma_{\mathbf{X}} \sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} ln\,(n)} = \frac{E_{\hat{f}_{2^{\mathbf{X}}*}} \left\{ E_{\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}} | A} \left[\ln \left[\hat{f}^{R_\varepsilon(f)}_{\widehat{\eta}^{\mathbf{X}}_{R,f}} | A \right] \right] \right\}}{\gamma_{\mathbf{X}} \sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} ln\,(n)} \\ &= \frac{E_{\hat{f}_{2^{\mathbf{X}}*}} \{ \text{'Model Uncertainty Measure'}_A \}}{\gamma_{\mathbf{X}} \sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} ln\,(n)} \end{aligned}$$

3 Bounded Rationality Models

De Clippel and Rozen (2014) (DR henceforth) shows that for bounded rationality models, rationalization does not guarantee that the theory has bite when predicting behavior for unobserved menus. In this section, I revisit the examples proposed by DR and I compare the predictive ability of those models with respect to the utility maximization model that allows for errors in choices.

Bounded rationality models have risen to explain behavior that cannot be rationalized by standard models. However, when enhancing the 'fit' of the model, one may allow for 'too much',

therefore hurting the power of the model. In this section I compare the limited attention and categorization models with the performance of the utility maximization model if we were to allow for error. This application shows that, models of bounded rationality, in general, have higher data requirements to identify primitives, and therefore may lack of power to predict behavior.

3.1 Extending the Predictive Ability Approach

The approach presented above can be extended to other theories of consumption behavior. Consider a theory $\mathcal{T}$, such that each $T \in \mathcal{T}$ describes a collection of observations that is consistent with the theory $\mathcal{T}$. A collection of observations f is rationalized by some rule under theory $\mathcal{T}$, if a is prescribed to be chosen from A by theory $T \in \mathcal{T}$, whenever $f(a, A) > 0$. Let $T(f)$ be the (potentially incomplete) preference information that can be inferred from data, that is, $T(f) \subseteq T$ with $T \in \mathcal{T}$. Let $\mathcal{T}(f)$ the set of all complete extensions (i.e. predictions for all elements in $\mathcal{O}$) such that $T(f) \subseteq \mathcal{T}(f)$, that is

$$\mathcal{T}(f) = \{T \in \mathcal{T} : f(a, A) > 0 \Rightarrow a \; T \; x \; \forall x \in A \setminus \{a\} \; \forall A \in \mathcal{A}\} \tag{9}$$

For the utility maximization model, the existence of an incomplete preference relation that rationalizes the observed collection implies the non-emptiness of $\mathcal{T}(f)$ while this needs not to be true for models of bounded rationality.

Definition 13 (Predictions for $2^{\mathbf{X}*}$) *For a theory $\mathcal{T}$, let $\mathcal{T}(f)$ be the collection of (complete) decision rules that is consistent with the observed collection f under theory $\mathcal{T}$ as in equation (9). Then,*

$$\hat{f}^{\mathcal{T}(f)}(a, A) = \frac{\gamma_f^{\mathcal{T}}(a, A)}{t} \times \hat{f}_{2^{X*}}(A) \tag{10}$$

for all $a \in A$ with $A \in 2^{\mathbf{X}*}$ and where $\gamma_f^{\mathcal{T}}(a, A) \equiv \sum_{T \in \mathcal{T}(f)} \mathbb{1}[a \; T \; b \; \forall \; b \in A \setminus \{a\}]$ and $t = |\mathcal{T}(f)|$.

3.2 Limited Attention

The limited attention model assumes that the DM facing a menu A selects the best alternative given a preference order P over a consideration set $\Gamma(A) \subset A$, assuming that not considered alternatives do not change the consideration set. In particular,

$$f(a, A) > 0 \Rightarrow a = \arg\max_P \Gamma(A) \text{ and } \Gamma(A) \subset B \subset A \Rightarrow \Gamma(B) = \Gamma(A) \text{ for all } A, B \in 2^{X^*}$$

Example 6 (Example 3 in DR) *Consider the following slight modification of Example 3 in DR*

A	ae	ef	abd	ade	bde	af	be
$\{x \in A : f(x, A)\}$	e	f	d	a	b	f	b

Limited Attention *From the analysis in DR it follows that the revealed preference information from data under the limited attention model is: (i) aPd, (ii) dPa or bPe and (iii) dPb or aPe. By relying on the test on out of sample predictions proposed by DR, the only revealed preference relation that can be extended such that predictions are non-empty for all possible menus, reduces to aPd, bPe and aPe. When predicting choices for $\tilde{A} = \{b, e, f\}$ the model predicts that either b or f can be chosen.*

In this case, assuming equal ex-ante probability, the predictive ability measure of the model for the menu $\{b, e, f\}$ is given by

$$PAM_f|A = 1 - \frac{H\left(\hat{f}^{R_\varepsilon(f)}_{\hat{\eta}^{\mathbf{X}}_{R,f}}|\{b,e,f\}\right)}{H\left(\hat{f}_{f=\emptyset}|\{b,e,f\}\right)} = 1 - \frac{-\left(\frac{1}{2}\ln\frac{1}{2} + \frac{1}{2}\ln\frac{1}{2}\right)}{-\left(\frac{1}{3}\ln\frac{1}{3} + \frac{1}{3}\ln\frac{1}{3} + \frac{1}{3}\ln\frac{1}{3}\right)} = 1 - \frac{0.6931}{1.0986} = 0.3691$$

Utility Maximization *Note that the above data cannot be rationalized by a well-behaved preference relation, since, for example $f(e, \{a, e\}) > 0$ and $f(a, \{a, d, e\}) > 0$ generate a violation to acyclicality. Assuming an error process that satisfies assumption 2, and under the assumption that the DM faced each menu only once, or that the probability of observing each menu is $\frac{1}{7}$, $|\mathcal{R}_\varepsilon(f)| = 9$. Since all $R \in \mathcal{R}_\varepsilon(f)$ can be recovered by swapping three times one alternative for a dominated one, I am pooling the observation across all menus and all preference orders. In particular, we have that $\hat{\eta}^{\mathbf{X}}_1 = \frac{4}{7}$ and $\hat{\eta}^{\mathbf{X}}_2 = \frac{3}{7}$. When constructing the predictive distribution for $\{b, e, f\}$ it is important to notice that 3 out of the 9 orders in $\mathcal{R}_\varepsilon(f)$ prescribe $bRfRe$ while the other 6 prescribe $fRbRe$. Then, the predictive distribution for $\{b, e, f\}$ is given by*

$$\hat{f}^{R_\varepsilon(f)}_{\hat{\eta}^{\mathbf{X}}_{R,f}}(x, \{b,e,f\}) = \begin{cases} \hat{f}_{2^{\mathbf{X}_*}}(\{b,e,f\})\left[\frac{1}{3}\hat{\eta}^{\mathbf{X}}_1 + \frac{2}{3}\hat{\eta}^{\mathbf{X}}_2\right] = \frac{10}{21}\hat{f}_{2^{\mathbf{X}_*}}(\{b,e,f\}) & \text{if } x = b \\ \hat{f}_{2^{\mathbf{X}_*}}(\{b,e,f\})\left[\frac{1}{3}\hat{\eta}^{\mathbf{X}}_2 + \frac{2}{3}\hat{\eta}^{\mathbf{X}}_1\right] = \frac{11}{21}\hat{f}_{2^{\mathbf{X}_*}}(\{b,e,f\}) & \text{if } x = f \\ 0 & \text{if } x = e \end{cases}$$

and the predictive ability measure conditional for this set is given by

$$PAM_f|A = 1 - \frac{H\left(\hat{f}^{\mathcal{R}_\varepsilon(f)}_{\hat{\eta}^{\mathbf{X}}_{\mathcal{R},f}}|\{b,e,f\}\right)}{H\left(\hat{f}_{f=\emptyset}|\{b,e,f\}\right)} = 1 - \frac{-\left(\frac{10}{21}\ln\frac{10}{21} + \frac{11}{21}\ln\frac{11}{21}\right)}{-\left(\frac{1}{3}\ln\frac{1}{3} + \frac{1}{3}\ln\frac{1}{3} + \frac{1}{3}\ln\frac{1}{3}\right)} = 1 - \frac{0.6920}{1.0986} = 0.3701$$

Therefore, the predictive ability for the utility maximization model is slightly higher.

3.3 Categorization and Rationalization

Rationalization and Categorization are observationally equivalent to the maximization of a complete, asymmetric relation P over a psychological filter Ψ. The filter is such that satisfies that the DM pays attention to it in any subset in which it is contained. The filter probability implies that the DM certainly pays attention to alternatives in

$$\Psi(A) \equiv \{b \in B : f(b, B) > 0 | A \subset B, B \in \mathcal{A}, b \in A\}$$

However, in limited data sets ψ may not qualify as a filter since it may be empty-valued for some out of sample menus.

Example 7 (Example 4 in DR) *Consider the following observed menus and choices*

A	ab	ad	ae	bd	be	de	abd	abe	bde
$\{x \in A : f(x, A)\}$	b	d	a	d	e	e	b	a	d

Rationalization/Categorization *Observed choices yield dP^*b, bP^*a, aP^*e, eP^*d, eP^*b and dP^*a which is a complete order. Moreover it imposes the following restrictions (i) dPa or dPb from $R = \{a, b, d\}$ and dP^*b, (ii) bPa or bPe, from $R = \{a, b, e\}$ and bP^*a; and (iii) ePb or ePd from $R = \{b, d, e\}$ and eP^*d. The out of sample prediction for $\{a, d, e\}$ is that either a or d can be picked, but e would not be consistent.*

In this case, assuming equal ex-ante probability, the predictive ability measure of the model for the menu $\{a, d, e\}$ is given by

$$PAM_f|A = 1 - \frac{H\left(\hat{f}^{R_\varepsilon(f)}_{\hat{\eta}^{\mathbf{X}}_{R,f}}|\{b, e, f\}\right)}{H\left(\hat{f}_{f=\emptyset}|\{b, e, f\}\right)} = 1 - \frac{-\left(\frac{1}{2}\ln\frac{1}{2} + \frac{1}{2}\ln\frac{1}{2}\right)}{-\left(\frac{1}{3}\ln\frac{1}{3} + \frac{1}{3}\ln\frac{1}{3} + \frac{1}{3}\ln\frac{1}{3}\right)} = 1 - \frac{0.6931}{1.0986} = 0.3691$$

Utility Maximization *Note that the above data cannot be rationalized by a well-behaved preference relation, since, for example $f(d, \{b, d\}) > 0$ and $f(b, \{a, b, d\}) > 0$ generate a violation to acyclicality. However, it is easy to check that, under the assumption that all menus are observed with equal probability $\frac{1}{9}$, the only well-behaved preference order that minimizes the swap index, and it is given by $R_\varepsilon(f)$ such that $dR_\varepsilon(f)bR_\varepsilon(f)aR_\varepsilon(f)e$; and the estimation of the error process delivers $\hat{\eta}^{\mathbf{X}}_1 = \frac{5}{9}$ and $\hat{\eta}^{\mathbf{X}}_2 = \frac{4}{9}$. Then, the predictive distribution for $\{a, d, e\}$ is given by*

$$\hat{f}^{R_\varepsilon(f)}_{\hat{\eta}^{\mathbf{X}}_{R,f}}(x, \{a, d, e\}) = \begin{cases} \hat{f}_{2^{\mathbf{X}_*}}(\{a, d, e\})\left[\eta^{\mathbf{X}}_2\right] = \frac{4}{9}\hat{f}_{2^{\mathbf{X}_*}}(\{a, d, e\}) & \text{if } x = a \\ \hat{f}_{2^{\mathbf{X}_*}}(\{a, d, e\})\left[\hat{\eta}^{\mathbf{X}}_1\right] = \frac{5}{9}\hat{f}_{2^{\mathbf{X}_*}}(\{a, d, e\}) & \text{if } x = d \\ 0 & \text{if } x = e \end{cases}$$

and the predictive ability measure conditional for this set is given by

$$PAM_f|A = 1 - \frac{H\left(\hat{f}^{\mathcal{R}_\varepsilon(f)}_{\hat{\eta}^{\mathbf{x}}_{\mathcal{R},f}}|\{a,d,e\}\right)}{H\left(\hat{f}_{f=\emptyset}|\{a,d,e\}\right)} = 1 - \frac{-\left(\frac{4}{9}\ln\frac{4}{9} + \frac{5}{9}\ln\frac{5}{9}\right)}{-\left(\frac{1}{3}\ln\frac{1}{3} + \frac{1}{3}\ln\frac{1}{3} + \frac{1}{3}\ln\frac{1}{3}\right)} = 1 - \frac{0.6870}{1.0986} = 0.3747$$

3.4 Model Comparison

Examples 6 and 7 show that even when theories of bounded rationality can perfectly explain observed behavior, the much simpler utility maximization model provides more informative predictions if we allow for error. These examples illustrate that, when dealing with limited datasets is important to understand the empirical content of considered models beyond their fit. More complex models of behavior may help understanding deviations but, the additional complexity may induce additional uncertainty when predicting behavior for unobserved economic environments.

4 Empirical Application: GEU Models

The proposed predictive approach can easily be extended to other models of consumption behavior and therefore used for model comparison. I consider the data discussed in Harless and Camerer (1994), but I restrict the analysis to the following theories: Expected Utility, Fan-out, Fan-in, Mixed Fanning, Reference Dependence with concave indifference curves; Reference Dependence with convex indifference curves; and Prospect theory. The authors discuss the importance of not just focusing on fit but also on parsimonia, understood as the number of feasible pattern allowed by each theory. Harless and Camerer (1994) explores the predictive ability of a series of choice models, which by allowing for errors in choices, explores all the information contained in observed behavior; penalizing for both: (i) systematic variation in unpredicted patterns; and (ii) the number of patterns that a given theory allows for.

Tables 2-6 in the Appendix summarize the results for the data analyzed in Harless and Camerer (1994) reporting the main criteria proposed by the authors and the measures proposed in this paper. The tables also report the ranking prescribed by these criteria in the horse race among the generalized expected utility models. The results show that the predictive measure trades off parsimonia, measured as the number of patterns allowed by each theory, with the error rate that is required by each of them to rationalized behavior. Therefore, the predictive measures follow the same intuition as the predictive odds criteria proposed by Harless and Camerer (1994). The main difference between predictive odds and predictive uncertainty is that the measure proposed in this paper penalize for the number of patterns allowed by each theory even when these are observed with zero frequency in sample.

5 Power vs. Uniformly Random Behavior: Simulations

In this section I present the results of simulation exercise where $\mathbf{X} = \{a, b, c, d, e\}$ where choices are generated as the result of the maximization of the preference order $a\ R\ b\ R\ c\ R\ d\ R\ e$ with different error process and a probability of menus to be observed that ranged from 0.1 to 1. The simulations were run for fully rational behavior, random generated behavior, and five alternative error processes as described in Table 1. For each of these treatments, 6 different assumptions on the observability of the menus are considered, $p \equiv \{0.10, 0.25, 0.50, 0.75, 0.90, 1\}$ where p_i is the probability of observing menu A for all menus in $2^{\mathbf{X}*}$ under treatment i.

Figure 1 shows the distributions of predictive uncertainty for different levels of observability of menus for all assumptions on the data generating process. Table 1 reports the p-values for the test of difference in means with respect to randomly generated behavior for the same probability of observing menus. difference in means and p-values in [] are reported; while Table 7 in Appendix B, shows the means and standard deviation for the normalized measure of predictive precision. The test results for the differences in means confirm that, as long as the dataset is sufficiently informative –in the sense that $P(f(A) > 0) > \frac{1}{4}$, the predictive ability measure can distinguish rational behavior with error with respect to random uniform behavior for processes generated by an error process significantly different to random errors for sets with more than three alternatives.

Table 8 displays the mean and standard deviations per η treatment for the error uncertainty measure, where the distribution of η is estimated across all swaps preferences. Table 9 displays the same statistics for model uncertainty measure for all treatments, for the same estimated distribution of the eta.

When regressing the normalized measure of predictive uncertainty (or predictive ability) with respect to the two sources of uncertainty the estimates of the correlation are, as expected, fairly stable and significant for all cases but when the probability of observing menus is 0.1 and the assumption on the error process is either random or disperse. The results for these estimates are shown in Figure 2. The results are displayed per assumption on the probability of observing menus. The expected pattern to be observed is increasing coefficients within subfigure for error uncertainty while decreasing across figures for model uncertainty. Figure 3 displays the same coefficient across assumptions for the error process where the expected pattern is the opposite.

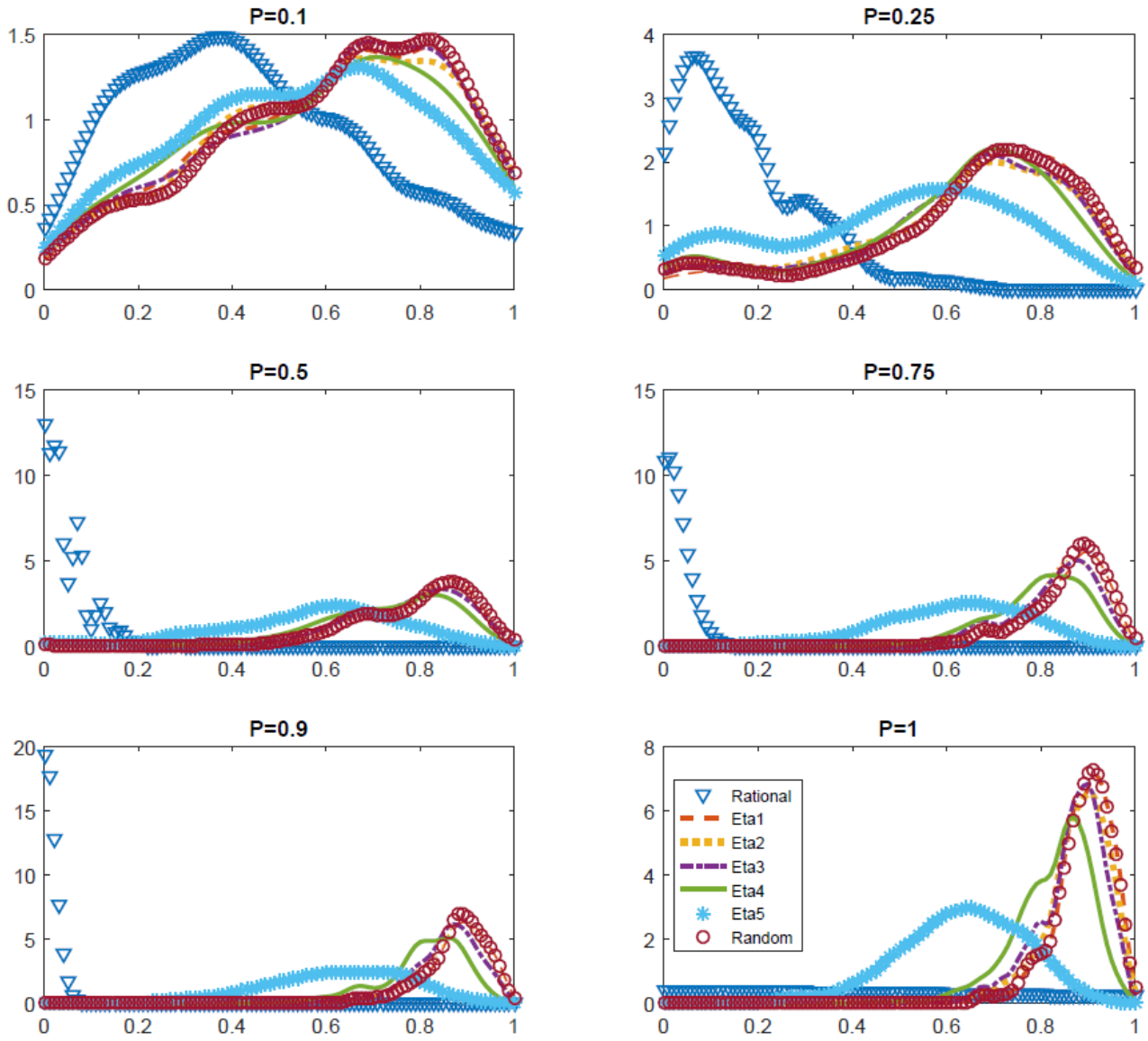

Figure 1: Distribution of Predictive Uncertainty per level of observability of menus and assumption on the error process.

Table 1: Test for differences in means for Predictive Ability Measure with respect to random behavior.

Assumption $\eta^{\mathbf{X}}$	Probability of observing choices from $A \in 2^{X*}$					
	0.10	**0.25**	**0.50**	**0.75**	**0.90**	**1.00**
η_1^X	0.0053	-0.0109	-0.0074	-0.0012	0.0031	0.0000
[.25;.20;.20;.20;.15]	[0.6481]	[0.2866]	[0.2418]	[0.7757]	[0.3238]	[0.9866]
η_2^X	0.0086	0.0064	0.0091	0.0165	0.0122	0.0148
[.35;.25;.20;.15;.05]	[0.4585]	[0.0136]	[0.1692]	[0.0002]	[0.0001]	[0.0000]
η_3^X	0.0078	0.0201	0.0165	0.0280	0.0255	0.0250
[.40;.30;.20;.10;0]	[0.5043]	[0.0545]	[0.0123]	[0.0000]	[0.0000]	[0.0000]
η_4^X	0.0307	0.0435	0.0514	0.0594	0.0671	0.0653
[.50;.30;.20;0;0]	[0.0089]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
η_5^X	0.0531	0.1616	0.2157	0.2407	0.2534	0.2594
[.70;.20;.075;.025,0]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
η_6^X	0.1644	0.4794	0.7338	0.8398	0.8726	0.8901
(Rational)	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]

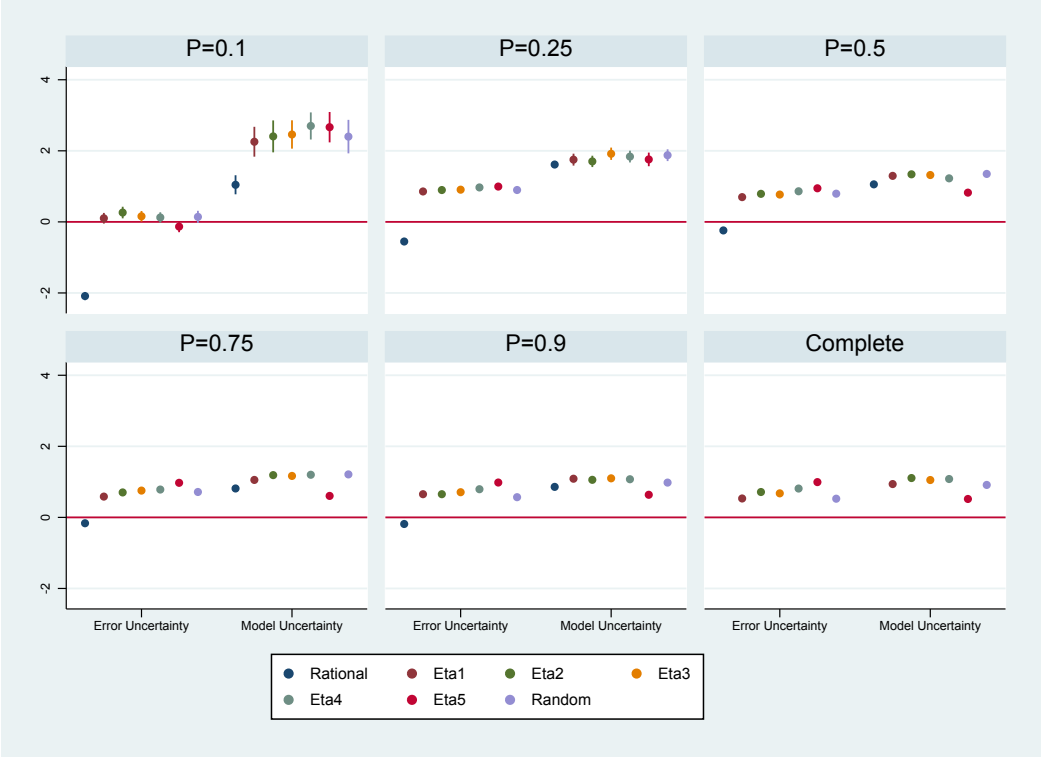

Figure 2: Correlations Predictive Uncertainty, Model Uncertainty and Error Uncertainty. Confidence intervals at 95% for all treatments.

6 Discussion and Extensions

6.1 Experimental Design

The predictive approach can be used to design experiments to better identify underlying behavior. Given an incomplete collection of observations, additional questions can be efficiently selected to maximize the predictive ability of the extended data for the considered model. Assuming that new observations are drawn from the same data generating process as f, one may select additional menus to observed choices from by maximizing the expected impact on model uncertainty. This procedure follows the one proposed by Boccardi (2017) and the intuition of the optimal placement index of Andreoni et. al (2013) for data from budget sets.

6.2 Random Utility Models

Model Uncertainty resembles Random Utility models where all 'utility models' in $\mathcal{R}^{ext}(f)$ have equal relative probably. However, the nature of this uncertainty is fundamentally different from classical random utility models. The assumption insofar is that there exist a unique preference relation that generates observed behavior with potential errors); but this may only be partially recovered from data. Partial identification of the underlying complete linear order induces uncertainty with respect to the unique preference relation that generates data; where all processes that are consistent with the collection of observations are deemed to be equally likely.

In the limit, preferences are fully identified, i.e. if $\mathcal{A} = 2^{\mathbf{X}*}$, $|\mathcal{R}^{ext}(f)| = 1$. On the contrary, under RUM many alternative utility orders may coexist, and observed choices can be the result of any of these preference orders, with a specified likelihood.

Correctly identifying the process generating the data is not only important in order to improve predictions but also to fully understand the welfare implications of the violations. Recovering the revealed preference information by minimizing the inconsistency index would confound the existence of multiple preference orders that jointly rationalized the data with errors if the utility maximization model is assumed, inducing misspecification errors.

Liang (2016) proposes a statistical regularization approach to select the number of distinct orderings that rationalize the data by establishing a data driven trade-off that seeks for parsimonious models (penalizing for the 'complexity' of the model) and fit (penalizing for the unexplained portion of the data by the considered orderings). This approach proposes a methodology to select the number of alternative by simultaneously minimizing the number of orderings and the number of implied choice errors for those orderings by establishing a linear trade off between these two.

Allowing for a multiplicity of preferences implies an additional level for 'Model Uncertainty' that, in principle, does not vanish with size of the sample. Given a series of candidate models,

Bayesian Model Averaging provides a framework to compute an 'average' predictive distribution by constructing a weighted average of the predictive distributions delivered by each of the candidate models, where the weights are constructed to reflect the empirical evidence of the relative likelihood of each of the considered models.

Definition 14 (Average Predictive Distribution) *Consider a set of K of candidate orderings $\mathcal{R}^K \equiv \{R_1, \ldots, R_K\}$ and let $w^k(f)$ be the weight for model k that reflects the relative likelihood of ordering R_k and such that $w^k(f) \in [0,1]$ for all $k \in \{1, \ldots, K\}$ and $\sum_{j=1}^{K} w^j(f) = 1$. Let $\hat{f}_{\chi}^{R^k(f)}(a, A)$ be the predictive distribution constructed for ordering R_k as in Definition 7 where $\chi(\cdot)$ the estimated error process and $k \in \{1, \ldots, K\}$. Then the 'average predictive distribution' is defined as*

$$\hat{f}_{APD(\mathcal{R}^K)}(a, A) = \sum_{j=1}^{K} w^k(f) \hat{f}_{\chi}^{R^k(f)}(a, A)$$

Definition 15 (Weights to reflect multiplicity) *Consider a set of K of candidate orderings $\mathcal{R}^K \equiv \{R_1, \ldots, R_K\}$ with weights to reflect multiplicity that can be constructed as*

$$w^k_{\text{'multiplicity'}}(f) = \frac{\frac{1}{I(f)|R_k}}{\sum_{j=1}^{J} \frac{1}{I(f)|R_j}}$$

However, it is important to highlight that in order to extend the construction to accommodate for multiple preference orders, restrictive assumptions should be consider on the error process and the nature of this multiplicity to allow for separately identify the different components of this extension.

6.3 'DGP Uncertainty'

An alternative approach to soften the unique ordering assumption has a similar flavor though a slightly different interpretation. Instead of assuming that there are multiple preferences that can be context-dependent or accommodate for pooling cross-sectional data, the researcher may simply consider that the swaps preference maybe the most likely underlying ordering under assumptions, but it is not the only likely outcome. Consider again example 4. The true underlying order $a\mathbf{R}b\mathbf{R}c$ which induces the lowest inconsistency index (~ 0.413); however the ordering $a\mathbf{R}'c\mathbf{R}'b$ induces the second lowest inconsistency index ~ 0.863. Therefore the researcher knows that, under maintain assumptions on the structure of the error process, $\mathbf{R}$ is the most likely order; but may not want to discard the information contained but alternative orderings that may have generated behavior.

The construction as in Definition 14 can be employed to reflect 'DGP uncertainty' where weights are given by the relative likelihood of the considered process to be the true underlying

process. The example above hints that the inconsistency measure provides a (negative) metric, in sample, of the likelihood of a preference relation to have generated the data, the bigger the errors implied by an ordering, the less likely the ordering is the one generating the data.

7 Relation to Existing Literature

This paper relates to the literature on revealed preference testing by extending previous approaches to: (i) account for fit and power; (ii) be applicable in a discrete choice environment where menus do not respond to budget structures; (iii) be used for other models of behavior to allow for model comparison and (iv) design experiments to maximize the certainty when predicting behavior.

The present approach relates to the welfare measure proposed by Apesteguia and Ballester (2015) by employing their machinery to recover the revealed preference information from data. Extending the model to allow for error I provide a theoretical justification to the swap preference to construct the predictive distribution as in Boccardi (2017) reflecting uncertainty due to the error process and the (potential) multiplicity of preferences that rationalize observed (projected) behavior.

Recovered errors measure the uncertainty related to the literature on goodness of fit. That is, errors size the extent of the adjustments to observed data in order for it to be rationalized by a well-behaved preference relation. There is a vast literature on goodness of fit but it mostly relies on the budget structure of the problem, as in Afriat (1967), Varian (1990), Echenique *et al.* (2011) and Dean and Martin (2016). This paper does not assume such a menu structure and therefore is closer in spirit to counting measures as in Houtman and Maks (1985) and Famulari (1995). Apesteguia and Ballester (2015) accounts for the number of deviations penalizing for the welfare implications of the number of swaps that are required to rationalized observed behavior.

A problem that plagues the literature on goodness of fit is that results can be hardly interpreted when dealing with limited datasets, due to the potential lack of power. In the discrete choice environment, there is a direct relation between incomplete data sets and the power of the test to detect violations. The most prominent approach in the literature is to compare the results to those that would have been obtained if data were to be generated uniformly at random, Bronars (1987). This paper follows the approach proposed by Boccardi (2017) by exploiting conditional power instead. The predictive ability measure reflects conditional power in terms of the potential multiplicity of preferences that rationalize behavior induced by incomplete datasets. Additionally, I show that when the assume error process is 'steep enough' and the incompleteness of the data is limited, the proposed measure has power to differentiate rational behavior from uniformly generated one.

Finally, this paper relates to the observation that, for model of bounded rationality, the empirical accuracy of the model for limited data sets does not imply extensibility of the model, De Clippel and Rozen (2014). The proposed approach accounts for this possibility and allows for the comparison across different models with different levels of complexity.

8 Conclusion

This paper extends the predictive approach proposed by Boccardi (2017) to models of consumption choices from finite menus of alternatives. Predictions reflect the uncertainty about preferences due to partial recoverability from data and potential errors that need to be allowed for to rationalize data. Therefore, measuring the uncertainty when predicting behavior over all menus provides a measure of both: (i) fit, through the dispersion of the error process; and (ii) power, as a negative function of the degree of partial observability. I show the performance of the proposed measures in controlled simulation exercises and its application to generalized expected utility models for the data from Harless and Camerer (1994). By relying on the predictive performance of the model, the proposed approach addresses the concerns raised by De Clippel and Rozen (2014) with respect to the extensibility of model of bounded rationality for out of sample menus. This is done by focusing on those underlying functionals that are extensible to all menus from the grand set of alternatives.

References

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International Economic Review*, **8** (1), pp. 67–77.

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

Apesteguia, J. and Ballester, M. A. (2015). A measure of rationality and welfare. *Journal of Political Economy*, **123** (6), 1278–1310.

Beatty, T. K. and Crawford, I. A. (2011). How demanding is the revealed preference approach to demand? *The American Economic Review*, **101** (6), 2782–2795.

Blundell, R. W., Browning, M. and Crawford, I. A. (2003). Nonparametric engel curves and revealed preference. *Econometrica*, **71** (1), pp. 205–240.

Boccardi, M. J. (2017). Predictive ability and the fit-power trade-off in theories of consumer behavior.

Bronars, S. G. (1987). The power of nonparametric tests of preference maximization. *Econometrica: Journal of the Econometric Society*, pp. 693–698.

Cherepanov, V., Feddersen, T. and Sandroni, A. (2013). Rationalization. *Theoretical Economics*, **8** (3), 775–800.

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

De Clippel, G. and Rozen, K. (2014). Bounded rationality and limited datasets.

Dean, M. and Martin, D. (2016). Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics*, **98** (3), 524–534.

Echenique, F., Lee, S. and Shum, M. (2011). The money pump as a measure of revealed preference violations. *Journal of Political Economy*, **119** (6), 1201–1223.

Famulari, M. (1995). A household-based, nonparametric test of demand theory. *The Review of Economics and Statistics*, pp. 372–382.

Harbaugh, W. T., Krause, K. and Berry, T. R. (2001). Garp for kids: On the development of rational choice behavior. *American Economic Review*, pp. 1539–1545.

Harless, D. W. and Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica: Journal of the Econometric Society*, pp. 1251–1289.

Houtman, M. and Maks, J. (1985). Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve methoden*, **19**, 89–104.

Kalai, G., Rubinstein, A. and Spiegler, R. (2002). Rationalizing choice functions by multiple rationales. *Econometrica*, **70** (6), 2481–2488.

Liang, A. (2016). How many choice contexts are there?

Manzini, P. and Mariotti, M. (2012). Categorize then choose: Boundedly rational choice and welfare. *Journal of the European Economic Association*, **10** (5), 1141–1165.

Masatlioglu, Y., Nakajima, D., Ozbay, E. Y. *et al.* (2012). Revealed attention. *American Economic Review*, **102** (5), 2183–2205.

Sippel, R. (1997). An experiment on the pure theory of consumer's behaviour*. *The Economic Journal*, **107** (444), 1431–1444.

VARIAN, H. R. (1990). Goodness-of-fit in optimizing models. *Journal of Econometrics*, **46** (1), 125–140.

A Proofs

A.1 Proof of Proposition 1

Proof. The entropy of the predictive distribution for any particular menu $A \in 2^{\mathbf{X}*}$ can be written as

$$\begin{aligned}
Entropy(\hat{f}|A) &= -\sum_{a\in A} \frac{\hat{f}(a,A)}{\hat{f}(A)} ln\left(\frac{\hat{f}(a,A)}{\hat{f}(A)}\right) \\
&= -\frac{1}{\hat{f}(A)} \sum_{a\in A} \hat{f}(a,A)\left[ln\left(\hat{f}(a,A)\right) - ln\left(\hat{f}(A)\right)\right] \\
&= -\frac{1}{\hat{f}(A)} \left[\sum_{a\in A} \hat{f}(a,A) ln\left(\hat{f}(a,A)\right) - \sum_{a\in A} \hat{f}(a,A) ln\left(\hat{f}(A)\right)\right] \\
&= -\frac{1}{\hat{f}(A)} \left[\sum_{a\in A} \hat{f}(a,A) ln\left(\hat{f}(a,A)\right)\right] + ln\left(\hat{f}(A)\right) \frac{1}{\hat{f}(A)} \sum_{a\in A} \hat{f}(a,A) \\
&= -\frac{1}{\hat{f}(A)} \left[\sum_{a\in A} \hat{f}(a,A) ln\left(\hat{f}(a,A)\right)\right] + ln\left(\hat{f}(A)\right)
\end{aligned}$$

while the overall entropy is given by

$$\begin{aligned}
Entropy(\hat{f}) &= -\sum_{a\in A, A\in 2^{\mathbf{X}*}} \hat{f}(a,A) ln\left(\hat{f}(a,A)\right) \\
&= -\sum_{A\in 2^{\mathbf{X}*}} \sum_{a\in A} \hat{f}(a,A) ln\left(\hat{f}(a,A)\right)
\end{aligned}$$

Then,

$$
\begin{aligned}
Entropy(\hat{f}) &= -\sum_{A\in 2^{\mathbf{X}*}} \left[-\hat{f}(A)\left(Entropy(\hat{f}|A) - ln(\hat{f}(A))\right)\right] \\
&= \sum_{A\in 2^{\mathbf{X}*}} \hat{f}(A)\left[\left(Entropy(\hat{f}|A) - ln(\hat{f}(A))\right)\right] \\
&= E_{\hat{f}(A)}\left[Entropy(\hat{f}|A)\right] - E_{\hat{f}(A)}\left[ln(\hat{f}(A))\right] \\
&= E_{\hat{f}(A)}\left[Entropy(\hat{f}|A)\right] + Entropy(\hat{f}(A))
\end{aligned}
$$

where $E_{\hat{f}(A)}[\cdot]$ it is the expectation with respect to the probability distribution over menus for the sake of prediction, i.e. $E_{\hat{f}(A)}[X(A)] = \sum_{A\in 2^{\mathbf{X}*}} \hat{f}(A)X(A)$. ■

A.2 Proof of Proposition 2

Proof. To prove that $NE(\hat{f}) \in [0,1]$ first note that

$$
NE(\hat{f}) = \frac{\sum_{A\in 2^{\mathbf{X}*}} H(\hat{f}|A)}{\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} \ln(n)}
$$

Since $\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} \ln(n)$ does not depend on observed information but on $\mathbf{X}$ the lower and upper bounds for $NE(\hat{f})$ depend uniquely on $\sum_{A\in 2^{\mathbf{X}*}} H(\hat{f}|A)$.

It is straightforward to see then that the lower bound of $NE(\hat{f})$ is given when there is 'no uncertainty', i.e. predictions are certain for all $A \in 2^{\mathbf{X}*}$. If that is the case, then $H(\hat{f}|A) = 0$ for all menus $A \in 2^{\mathbf{X}*}$; while for any other predictive distribution $H(\hat{f}|A) \geq 0$ for all A with $H(\hat{f}|A) > 0$ for some $A \in 2^{\mathbf{X}*}$; and $\gamma_{\mathbf{X}} = \frac{1}{2^{|\mathbf{X}|}-|\mathbf{X}|-1} > 0$ for all $\mathbf{X}$ such that $|\mathbf{X}| > 2$. Then, $NE(\hat{f}) \geq 0$ and, $NE(\hat{f}) = 0 \Leftrightarrow H(\hat{f}|A) = 0$ for all $A \in 2^{X*}$ which only occurs when $|\mathcal{R}^{ext}(f)| = 1$.

On the other hand, the upper bound of $NE(\hat{f})$ is given when there is total uncertainty or 'no information' for all menus $A \in 2^{\mathbf{X}*}$, since the 'Menu Entropy' is maximal for each menu $A \in 2^{\mathbf{X}*}$.

That is, for any given menu $H(\hat{f}|A) \leq \ln(|A|)$, therefore, under Assumption 1

$$\begin{aligned} NE(\hat{f}) &= \frac{\sum_{A\in 2^{\mathbf{X}_*}} H(\hat{f}|A)}{\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} \ln(n)} \\ &\leq \frac{\sum_{A\in 2^{\mathbf{X}_*}} ln(|A|)}{\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} \ln(n)} \\ &= \frac{\sum_{n=2}^{|\mathbf{X}|} \left[\sum_{A\in 2^{\mathbf{X}_*}} \mathbb{1}\left[|A|=n\right]\right] ln(n)}{\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} \ln(n)} \\ &= \frac{\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} \ln(n)}{\sum_{n=2}^{|\mathbf{X}|} C_n^{|\mathbf{X}|} \ln(n)} = 1 \end{aligned}$$

Moreover $H(\hat{f}|A) = \ln(|A|) \Leftrightarrow \hat{f}(a,A) = \frac{1}{|A|}$ for all $a \in A$ and $A \in 2^{X*}$. ■

Proof of Claim 1.

$$\begin{aligned} \mathbb{E}_{A\in\mathcal{D}}\left[\varepsilon^{\mathbf{R}'}_{(a,A)}\right] - \mathbb{E}_{A\in\mathcal{D}}\left[\varepsilon^{\mathbf{R}}_{(a,A)}\right] &= \gamma_{\mathbf{X}} \left(\sum_{A\in 2^{\mathbf{X}_*}} \sum_{a\in A} \left(g^{\mathbf{R}'}(a,A) - g^{\mathbf{R}}(a,A) \right) \times \eta^{\mathbf{R}}_{(a,A)} \right) \\ &= \gamma_{\mathbf{X}} \left(\sum_{A\in 2^{\mathbf{X}_*}} \mathbb{1}\left[\{a,b\}\subseteq A\right] \sum_{c\in\{a,b\}} \left(g^{\mathbf{R}'}(c,A) - g^{\mathbf{R}}(c,A) \right) \times \eta^{\mathbf{R}}_{(c,A)} \right) \end{aligned}$$

Since only the ranking among a, b is changed, $\Delta g(a,A) = -\Delta g(b,A)$ where $\Delta g(c,A) \equiv g^{\mathbf{R}'}(c,A) - g^{\mathbf{R}}(c,A) = 1$, therefore

$$\begin{aligned} \mathbb{E}_{A\in\mathcal{D}}\left[\varepsilon^{\mathbf{R}'}_{(a,A)}\right] - \mathbb{E}_{A\in\mathcal{D}}\left[\varepsilon^{\mathbf{R}}_{(a,A)}\right] &= \gamma_{\mathbf{X}} \left(\sum_{A\in 2^{\mathbf{X}_*}} \mathbb{1}\left[\{a,b\}\subseteq A\right] \left(\Delta g(a,A) \times \eta^{\mathbf{R}}_{(a,A)} + \Delta g(b,A) \times \eta^{\mathbf{R}}_{(b,A)} \right) \right) \\ &= \gamma_{\mathbf{X}} \left(\sum_{A\in 2^{\mathbf{X}_*}} \mathbb{1}\left[\{a,b\}\subseteq A\right] \Delta g(a,A) \left(\eta^{\mathbf{R}}_{(a,A)} - \eta^{\mathbf{R}}_{(b,A)} \right) \right) > 0 \end{aligned}$$

where the last inequality follows from $\Delta g(a,A) \equiv g^{\mathbf{R}'}(a,A) - g^{\mathbf{R}}(a,A) > 0$ and $\left(\eta^{\mathbf{R}}_{(a,A)} - \eta^{\mathbf{R}}_{(b,A)}\right) > 0$ by assumption 2. ■

B Tables and Figures

B.1 Harless and Camerer (1994) data

The following tables replicate the results presented in Harless and Camerer (1994) together with the computation of the average leave one out computation for the predictive uncertainty and the normalized predictive uncertainty; as well as the ranking prescribed for the selected criteria for

the considered theories. The criteria considered are: (i) the percentage of explained choices by the patterns allowed by each theory; (ii) the proportion of explained choices with respect to the percentage of allowed patterns; (iii) the ratio between the percentage of explained choices and the maximum percentage of choices that could have been explained given the number of allowed patterns; (iv) the z-statistic for the likelihood of predicted patterns versus randomly generated behavior; (v) the Selten score for the difference between explained choices by the theory and what would have been explained if choices were to be generated uniformly at random; (vi) the error rate that should be allowed in order to rationalize observed behavior by each of these theories; (vii) the posterior odds for the model with respect to expected utility as computed by Harless and Camerer (1994); (viii) the distance between the distribution implied by the model and the observed distribution of data (a negative function of the performance of the model); and (ix) the distance between the distribution of choices implied by each theory and random behavior (a positive function of the power of the model).

Table 2: Table V.- Harless real gains from unit triangle interior. Outcomes: \$0,\$3,\$6. Probabilities: $S_1(.84, .14, .02)$; $R_1(.89, .01, .10)$; $S_3(.04, .94, .02)$; $R_3(.09, .81, .10)$; $S_5(.44, .14, .42)$; $R_5(.49, .01, .50)$; $S_7(.04, .14, .82)$; $R_7(.09, .01, .90)$ - Summary of the results

Criteria	EU	Fan-out	Fan-in	MF	RDcave	RDvex	PT
	Results						
% Explained	.310	.429	.631	.583	.905	.571	.857
% Allowed	.063	.125	.375	.375	.813	.563	.563
# patterns	2	6	6	13	9	9	5
% Max	.310	.429	.774	.774	.964	.893	.893
Z-stat	5.245	6.051	5.376	4.702	2.922	.870	5.924
Selten	.247	.304	.256	.208	.092	.009	.295
Error	.366	.219	.166	.216	.092	.216	.108
χ^2-stat	59.92	29.22	15.37	29.18	6.88	29.18	8.55
Avg. LOO	.656	1.095	1.411	1.392	2.056	1.567	1.845
NEU	.886	.855	.885	.837	.885	.839	.885
	Ranking						
% Explained	7	3	4	1	5	2	6
Expl/Allowed	1	2	4	6	7	5	3
Expl/Max	1	4	5	3	7	2	6
Z-stat	1	3	4	6	7	2	5
Selten	1	3	4	6	7	2	5
Error	7	3	4	1	4	2	4
PO vs EU - HC	1	2	5	6	7	3	4
Avg. LOO	1	2	4	3	7	5	6
Dist. true	4	3	4	1	4	2	4
Dist. rnd	7	3	4	1	4	2	4

Table 3: Table VI Harless real losses from unit triangle interior. Outcomes: -$4,-$2,$0. Probabilities: $S_1(.8, .18, .02)$; $R_1(.88, .02, .10)$; $S_3(.02, .96, .02)$; $R_3(.1, .8, .1)$; $S_5(.41, .18, .41)$; $R_5(.49, .02, .49)$; $S_7(.02, .18, .80)$; $R_7(.10, .02, .88)$

Criteria	EU	Fan-out	Fan-in	MF	RDcave	RDvex	PT
	Results						
% Explained	.279	.380	.595	.886	.595	.684	.317
% Allowed	.125	.375	.375	.813	.563	.563	.188
# patterns	2	6	6	13	9	9	3
% Max	.279	.646	.646	.949	.823	.823	.392
Z-stat	3.785	.614	4.720	2.505	1.627	3.298	3.420
Selten	.154	.005	.220	.074	.032	.121	.129
Error	.281	.281	.222	.141	.248	.234	.362
χ^2-stat	18.72	18.72	7.46	3.08	11.31	10.19	22.79
Avg. LOO	1.266	1.456	1.866	2.369	1.772	1.962	1.256
NEU	.965	.965	.942	.930	.950	.948	.969
	Ranking						
% Explained	7	5	3	1	3	2	5
Expl/Allowed	1	7	3	5	6	4	2
Expl/Max	1	7	3	2	6	4	5
Z-stat	2	7	1	5	6	4	3
Selten	2	7	1	5	6	4	3
Error	5	5	2	1	4	3	7
PO vs EU - HC	1	4	2	7	6	5	3
Avg. LOO	2	3	5	7	4	6	1
Dist. true	6	6	2	1	4	3	4
Dist. rnd	6	6	2	1	4	3	5

Table 4: Table VII Chew and Waller: \$0,-\$40,\$100. Probabilities: $S_o(0,1,0)$; $R_o(.5,0,.5)$; $S_i(0,1,0)$; $R_i(.05,.9,.05)$; $S_l(.9,.1,0)$; $R_l(.95,0,.05)$; $S_h(0,.1,.9)$; $R_h(.05,0,.95)$

Criteria	EU	Fan-out	Fan-in	MF	RDcave	RDvex	PT
	Results						
% Explained	.212	.697	.323	.879	.374	.838	.798
% Allowed	.125	.375	.375	.813	.563	.563	.625
# patterns	2	6	6	13	9	9	10
% Max	.374	.758	.758	.990	.899	.899	.929
Z-stat	3.07	6.86	-2.33	2.65	-4.80	6.14	4.55
Selten	.087	.322	-.052	.066	-.189	.276	.173
Error	.339	.178	.339	.121	.458	.143	.158
χ^2-stat	81.99	15.16	89.89	10.35	90.19	11.15	14.17
Avg. LOO	1.204	1.824	1.376	2.143	1.763	2.056	1.959
NEU	.985	.866	.985	.858	1.000	.859	.863
	Ranking						
% Explained	7	4	6	1	5	2	3
Expl/Allowed	2	1	6	5	7	3	4
Expl/Max	5	2	6	3	7	1	4
Z-stat	4	1	6	5	7	2	3
Selten	4	1	6	5	7	2	3
Error	5	4	5	1	7	2	3
PO vs EU - HC	5	1	6	4	7	2	3
Avg. LOO	1	4	2	7	3	6	5
Dist. true	5	4	7	1	6	2	3
Dist. rnd	5	3	6	2	7	1	4

Table 5: Table VIII - Sopher and Gigliotti - Common consequence Hypothetical large gains on unit triangle boundary: \$0,\$1M,\$5M. Probabilities: $S_1(0,1,0)$; $R_1(.01,.89,.1)$; $S_2(.89,.11,0)$; $R_2(.9,0,.1)$; $S_3(0,.11,.89)$; $R_3(.01,0,.99)$; $S_4(.79,.11,.1)$; $R_4(.8,0,.2)$; $S_5(.01,.89,.1)$; $R_5(.02,.78,.2)$

Criteria	EU	Fan-out	Fan-in	MF	RDcave	RDvex	PT
	Results						
% Explained	.220	.538	.258	.936	.667	.376	.796
% Allowed	.063	.250	.250	.656	.406	.406	.563
# patterns	2	8	8	21	13	13	18
% Max	.333	.747	.747	.968	.887	.887	.952
Z-stat	5.906	8.666	1.230	8.489	8.093	-2.395	7.521
Selten	.158	.288	.008	.279	.260	-.030	.233
Error	.299	.198	.299	.061	.147	.280	.126
χ^2-stat	189.90	97.40	189.90	15.36	53.91	186.11	48.93
Avg. LOO	.949	1.593	1.074	2.535	1.951	1.364	2.174
NEU	.921	.865	.921	.788	.826	.916	.820
	Ranking						
% Explained	7	4	6	1	3	5	2
Expl/Allowed	1	2	6	4	3	7	5
Expl/Max	5	4	7	1	3	6	2
Z-stat	5	1	6	2	3	7	4
Selten	5	1	6	2	3	7	4
Error	7	5	7	2	4	6	3
PO vs EU - HC	5	3	6	2	1	7	4
Avg. LOO	1	4	2	7	5	3	6
Dist. true	6	4	6	1	3	5	2
Dist. rnd	6	4	6	1	3	5	2

Table 6: Table IX. Sopher and Gigliotti - Common consequence Hypothetical large gains on unit triangle interior: \$0,\$1M,\$5M. Probabilities: $S_1(.01, .98, .01)$; $R_1(.02, .87, .11)$; $S_2(.80, .19, .01)$; $R_2(.81, .08, .11)$; $S_3(.01, .19, .80)$; $R_3(.02, .08, .90)$; $S_4(.70, .19, .11)$; $R_4(.71, .08, .21)$; $S_5(.02, .87, .11)$; $R_5(.03, .76, .21)$

Criteria	EU	Fan-out	Fan-in	MF	RDcave	RDvex	PT
	Results						
% Explained	.424	.484	.538	.853	.734	.630	.669
% Allowed	.063	.250	.250	.656	.406	.406	.219
# patterns	2	8	8	21	13	13	7
% Max	.446	.745	.745	.973	.870	.870	.712
Z-stat	10.205	7.491	8.628	6.719	9.465	7.292	12.263
Selten	.361	.234	.288	.197	.327	.224	.450
Error	.186	.184	.150	.104	.111	.173	.119
χ^2-stat	102.60	102.34	81.46	43.79	49.53	95.89	52.56
Avg. LOO	1.041	1.159	1.460	2.129	1.787	1.351	1.683
NEU	.818	.818	.801	.769	.766	.810	.771
	Ranking						
% Explained	7	6	5	1	2	4	3
Expl/Allowed	1	4	3	7	5	6	2
Expl/Max	1	7	6	3	4	5	2
Z-stat	2	5	4	7	3	6	1
Selten	2	5	4	7	3	6	1
Error	7	6	4	1	2	5	3
PO vs EU - HC	2	5	4	6	3	7	1
Avg. LOO	1	2	4	7	6	3	5
Dist. true	7	6	4	1	2	5	3
Dist. rnd	6	7	4	2	1	5	3

B.2 Simulation Results

Table 7: Means for measure of normalized predictive uncertainty.

Assumption η^X	**Probability of observing choices from $A \in 2^{X*}$**					
	0.10	**0.25**	**0.50**	**0.75**	**0.90**	**1.00**
Random	0.6120	0.6524	0.7843	0.8541	0.8785	0.8925
	(0.2600)	(0.2428)	(0.1513)	(0.1017)	(0.0698)	(0.0613)
η_1^X	0.6057	0.6616	0.7910	0.8555	0.8750	0.8923
	(0.2617)	(0.2238)	(0.1356)	(0.0869)	(0.0751)	(0.0623)
η_2^X	0.6031	0.6241	0.7743	0.8373	0.8654	0.8771
	(0.2622)	(0.2456)	(0.1503)	(0.0971)	(0.0721)	(0.0721)
η_3^X	0.6036	0.6295	0.7670	0.8259	0.8520	0.8669
	(0.2638)	(0.2358)	(0.1484)	(0.1002)	(0.0817)	(0.0721)
η_4^X	0.5803	0.6056	0.7298	0.7936	0.8101	0.8262
	(0.2692)	(0.2412)	(0.1530)	(0.1065)	(0.0890)	(0.0867)
η_5^X	0.5574	0.4852	0.5638	0.6114	0.6232	0.6314
	(0.2672)	(0.2525)	(0.1937)	(0.1589)	(0.1489)	(0.1341)
η_6^X	0.4449	0.1652	0.0434	0.0109	0.0031	0.0000
(Rational)	(0.2566)	(0.1346)	(0.0506)	(0.0201)	(0.0090)	(0.0000)

Table 8: Mean and standard deviations for the error uncertainty measure.

Assumption η^X	**Probability of observing choices from $A \in 2^{X*}$**					
	0.10	**0.25**	**0.50**	**0.75**	**0.90**	**1.00**
Random	0.1976	0.5644	0.7723	0.8573	0.8861	0.9009
	(0.3048)	(0.2810)	(0.1579)	(0.1027)	(0.0694)	(0.0604)
η_1^X	0.2027	0.5714	0.7799	0.8585	0.8818	0.9009
	(0.3030)	(0.2712)	(0.1393)	(0.0866)	(0.0751)	(0.0619)
η_2^X	0.2126	0.5288	0.7624	0.8401	0.8723	0.8859
	(0.3126)	(0.2883)	(0.1543)	(0.0976)	(0.0729)	(0.0723)
η_3^X	0.2159	0.5372	0.7545	0.8292	0.8588	0.8745
	(0.3077)	(0.2755)	(0.1516)	(0.1002)	(0.0816)	(0.0720)
η_4^X	0.1815	0.5143	0.7166	0.7950	0.8149	0.8335
	(0.2955)	(0.2776)	(0.1591)	(0.1064)	(0.0886)	(0.0870)
η_5^X	0.1335	0.3699	0.5417	0.6080	0.6251	0.6349
	(0.2601)	(0.2876)	(0.1972)	(0.1596)	(0.1484)	(0.1350)
η_6^X	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
(Rational)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

Table 9: Means for measure of model uncertainty

Assumption $\eta^{\mathbf{X}}$	**Probability of observing choices from $A \in 2^{X*}$**					
	0.10	**0.25**	**0.50**	**0.75**	**0.90**	**1.00**
Random	0.2479	0.1701	0.1036	0.0772	0.0657	0.0661
	(0.0515)	(0.0898)	(0.0918)	(0.0880)	(0.0834)	(0.0843)
η_1^X	0.2450	0.1694	0.1012	0.0798	0.0697	0.0647
	(0.0535)	(0.0907)	(0.0911)	(0.0901)	(0.0845)	(0.0823)
η_2^X	0.2436	0.1668	0.1029	0.0743	0.0627	0.0594
	(0.0546)	(0.0917)	(0.0928)	(0.0864)	(0.0801)	(0.0796)
η_3^X	0.2451	0.1688	0.1056	0.0687	0.0627	0.0634
	(0.0540)	(0.0872)	(0.0902)	(0.0836)	(0.0813)	(0.0789)
η_4^X	0.2460	0.1668	0.0944	0.0681	0.0574	0.0506
	(0.0521)	(0.0877)	(0.0872)	(0.0827)	(0.0748)	(0.0712)
η_5^X	0.2484	0.1741	0.0899	0.0535	0.0372	0.0311
	(0.2601)	(0.2876)	(0.1972)	(0.1596)	(0.1484)	(0.1350)
η_6^X	0.2485	0.1823	0.0936	0.0342	0.0118	0.0000
(Rational)	(0.0476)	(0.0775)	(0.0820)	(0.0682)	(0.0616)	(0.0000)

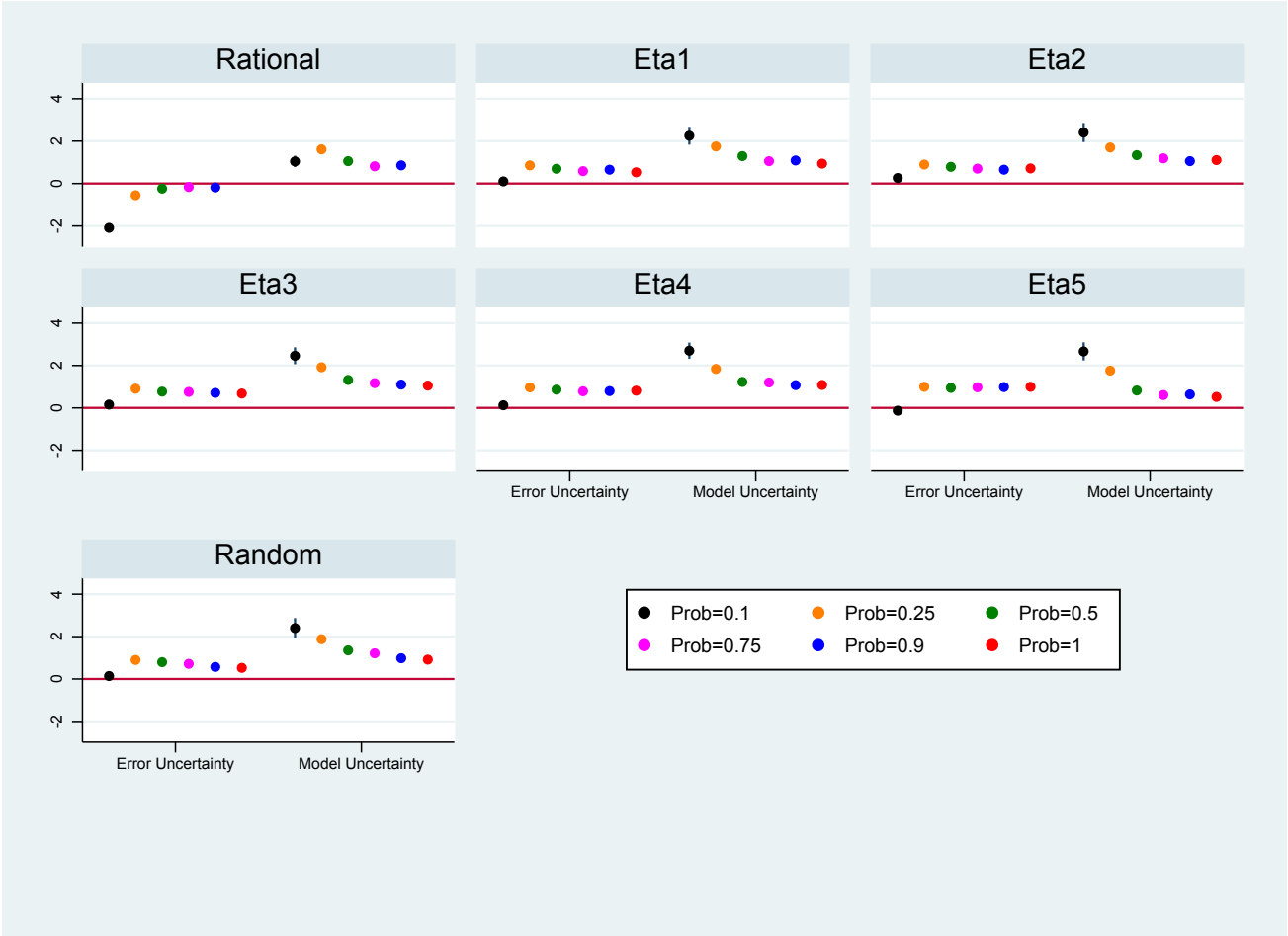

Figure 3: Correlations Predictive Uncertainty, Model Uncertainty and Error Uncertainty. Confidence intervals at 95% for all treatments.