SEPARATING PREDICTED RANDOMNESS FROM RESIDUAL BEHAVIOR*

JOSE APESTEGUIA† AND MIGUEL A. BALLESTER‡

Abstract. We propose a novel measure of goodness of fit for stochastic choice models: the maximal fraction of data that can be reconciled with the model. We do so by separating the data into two parts: one that is generated by the best specification of the model and another which represents residual behavior. We claim that the three elements involved in a separation are instrumental in understanding the data. We show how to apply our approach to any model of stochastic choice. We then study the case of four well-known models, each capturing a different notion of randomness. We illustrate our results with an experimental dataset.

Keywords: Goodness of fit; Stochastic Choice; Residual Behavior.
JEL classification numbers: D00.

1. Introduction

Choice data arising from either individual or population behavior often have a probabilistic nature. Today, there is renewed interest in obtaining a better modeling of stochastic behavior, and the literature offers a battery of models incorporating randomness in various ways.[1] In this paper we discuss a novel goodness of fit measure for stochastic choice models. The measure corresponds to the (tight) upper bound in the fraction of data that can be reconciled with the model. This approach requires to

Date: September, 2019.

*We thank Jay Lu, Paola Manzini, Marco Mariotti, Joel Sobel and Ryan Webb for very helpful comments, and Angelo Gutierrez for outstanding research assistance. Financial support by the Spanish Ministry of Science (PGC2018-098949-B-I00) and Balliol College is gratefully acknowledged.

†ICREA, Universitat Pompeu Fabra and Barcelona GSE. E-mail: jose.apesteguia@upf.edu.

‡University of Oxford. E-mail: miguel.ballester@economics.ox.ac.uk.

[1]Recently published papers include Gul and Pesendorfer (2006), Dickhaut, Rustichini and Smith (2009), Caplin, Dean and Martin (2011), Ahn and Sarver (2013), Gul, Natenzon and Pesendorfer (2014), Manzini and Mariotti (2014), Fudenberg and Strzalecki (2015), Fudenberg, Iijima and Strzalecki (2015), Matejka and McKay (2015), Barseghyan, Molinari and O'Donoghue (2016), Brady and Rehbeck (2016), Caplin and Dean (2016), Agranov and Ortoleva (2017), Apesteguia, Ballester and Lu (2017), Apesteguia and Ballester (2018), Webb (2019), and Natenzon (forthcoming).

separate the data into two parts: one which is generated by a particular specification of the model and is to be maximized, and another containing the remainder, which consists of unstructured behavior and is to be minimized. We refer to the first part as the predicted randomness given by the model and to the second as residual behavior.

The separation exercise highlights three key elements. First, the maximal fraction of data explained by the model, which represents our goodness of fit measure. This measures the ability of the model to explain actual behavior. Second, an optimal specification of the model. Clearly, when associated with large fractions of data explained, the identified specification of the model becomes a potentially useful tool in counterfactual scenarios, such as those associated with prediction problems. Third, a description of residual behavior. This facilitates a better understanding of the relationship between actual behavior and the choice model, since it endogenously enables the identification of the menus and choices for which the model fails most dramatically. This information may be relevant at the time of revising a model.

More formally, given the grand set of alternatives X, $\mathtt{SCF}$ denotes the set of all stochastic choice functions, i.e., all the possible descriptions of the probabilities of choice of each alternative in each menu in the domain. The aim is to explain data ρ, that is taken to be a stochastic choice function, in the light of model Δ, that is defined as a collection of stochastic choice functions. The model Δ describes all the possible predictions the analyst considers relevant. For example, it may encompass all the parametric specifications of the favorite choice model of the analyst, including those arising from the consideration of measurement error or unobserved heterogeneity. A triple $\langle \lambda, \delta, \epsilon \rangle$, where $\lambda \in [0, 1]$, $\delta \in \Delta$ and $\epsilon \in \mathtt{SCF}$, such that $\rho = \lambda\delta + (1 - \lambda)\epsilon$ describes a possible separation of data ρ into a fraction λ explained by the instance of the model δ and a fraction $1 - \lambda$ that corresponds to unstructured residual behavior ϵ. A separation is maximal if it provides the maximal value of λ. Proposition 1 in section 3 shows that, for any closed model Δ, maximal separations always exist, and characterizes their structure. The result shows that maximal separations are identified throughout a maxmin operation. First, for every instance of the model compute the minimum ratio, across all the observations, of data to prediction. Then, the solution is given by the instance of the model that maximizes such ratio. This is a simple method, applicable to any model, and potentially instrumental in the analysis of particular models, as it is shown later in the paper.

In section 4 we analyze four well-known stochastic choice models that predict randomness in very different ways. In all four cases, we build on Proposition 1 in order to provide tailored results describing the structure of the maximal separations of the different models. This exercise facilitates the practical implementation of maximal separations, and, by elaborating on the structural properties of the respective stochastic choice model, complements the conceptual understanding of the maximal separation approach. We start with the paradigmatic model of decision-making in economics: the deterministic choice model. In this model, the individual always selects the alternative that maximizes a preference relation, and hence there is no predicted randomness whatsoever. Thus, when a stochastic choice function is analyzed from the perspective of the deterministic model, any stochasticity in the data must be regarded as residual behavior. Given the overwhelming use of this model, it seems advisable to make it the first in our analysis of particular cases. Proposition 2 provides a simple recursive method over the sizes of the menus used to compute the maximal separations of the deterministic model.

We then turn to the study of three stochastic choice models. We start with the tremble model, where randomness represents the possibility of making mistakes at the time of choosing. In the tremble model, with probability $(1-\gamma)$ the decision-maker maximizes a preference relation, and with probability γ randomizes over all the available alternatives. Proposition 3 describes how to extend the technique developed for the deterministic model to this case. We then analyze the model of Luce (1959), which is also known as the logistic model. The Luce model incorporates randomness in the utility evaluation of the alternatives. Proposition 4 gives simplicity to the analysis of the Luce model by showing that the observations giving the minimum ratio of data to predictions in a maximal separation obey a particular structure. Finally, we study a class of random utility models incorporating randomness in the determination of the ordinal preference that governs choice. In particular, we study the class of single-crossing random utility models (Apesteguia, Ballester and Lu, 2017), that has the advantage of providing tractability, while also being applicable to a variety of economic settings. Proposition 5 gives the corresponding maximal separations, following a recursive argument over the collections of preferences in the support of the random utility model.

Section 5 reports on an empirical application of our approach. We use a previously-existing experimental dataset comprising 87 individuals making choices from binary comparisons of lotteries. We take the aggregate data of the entire population and illustrate the practicality of our results, obtaining the maximal separation results for all the models discussed in the paper. We first show that the maximal fraction of the data explained by the deterministic model, its goodness of fit, is 0.51, and that the preference relation identified in the maximal separation basically ranks the lotteries from least to most risky. The tremble model identifies exactly the same preference relation, together with a tremble probability of 0.51, which increases the fraction of data explained to 0.68. The Luce model also increases the fraction of data explained to 0.74, and identifies a utility function over lotteries that is ordinally close to the preference ranking of the deterministic and tremble models. Finally, we implement the single-crossing random utility model assuming the utility functions given by CRRA expected utility. We obtain that the fraction of data explained increases further to 0.78, with the largest mass being assigned to a preference exhibiting high levels of risk aversion.

Section 6 contrasts the maximal separation approach with other goodness of fit measures, like maximum likelihood and least squares. We argue that maximal separation, by focusing on the largest deviations from the data, is particularly accurate in the prediction involving low observed choice frequencies. We then use the experimental dataset to empirically illustrate this point, confirming that there are important complementarities between the maximal separation technique and the standard ones for gaining a deeper understanding of the data.

Section 7 concludes by discussing three aspects in the maximal separations approach. Firstly, we briefly analyze the model selection issue by discussing the case of an analyst wishing to compare the maximal fractions of data explained by different models. Secondly, we comment on the advantages and drawbacks of imposing further technical structure on the stochastic choice models. Finally, we consider the case when the notion of maximal separation is slightly modified by restricting the space of possible residual behaviors that can be combined with the randomness predicted by a model, and we conclude by suggesting some potentially fruitful ways of interpreting residual behavior.

2. Related literature

Rudas, Clogg and Lindsay (1994) developed a novel proposal in Statistics, presenting what is now known as the mixture index of fit for contingency tables. Given a multivariate frequency distribution, Rudas, Clogg and Lindsay (1994) suggest measuring the goodness of fit of a given model using a two-point mixture, by calculating the largest fraction of the population for which a distribution belonging to the model fits the data, while leaving the complementary fraction as an unstructured distribution. Rudas (1999) extends the mixture index to continuous probability distributions, and relates the optimal solution to minimax estimation.[2] The maximal separation technique imports the same logic for the study of stochastic choice functions. The latter differ from contingency tables in that they involve collections of interrelated probability distributions, one for each available menu of alternatives, where the interrelation is choice model dependent. Interestingly, Böckenholt (2006) claims that new methodologies are needed to understand the systematic behavioral violations of random utility models, and, without elaborating, suggests the mixture index of fit as one such potentially useful methodology. In this paper, we undertake this challenge by extending the methodology, not only to random utility models, but to every possible stochastic choice model, and then incorporate these ideas into Decision Theory and Economics.

In Economics, Afriat (1973) was the first in a long history of proposals for indices that measure the consistency of revealed preferences with the deterministic, rational model of choice. In a consumer setting, Afriat's suggestion was to compute the minimal amount of monetary adjustment required to reconcile all observed choices with the maximization of some preference; an idea later generalized by Varian (1990). Alternative suggestions by Houtman and Maks (1985), and more recently by Dean and Martin (2016), are to compute the maximal number of data points that are consistent with the maximization of some preference. Apesteguia and Ballester (2015) and Halevy, Persitz, and Zrill (2018) suggest consistency measures that compute the minimal welfare loss of inconsistent behavior with respect to some preference. Relevantly, Apesteguia and Ballester (2015) show, by means of an axiomatic approach, that all these measures have a common structure; they search for a preference that minimizes a given loss function, ultimately providing both a goodness of fit measure and the best possible description of behavior.[3] Since the maximal separation approach also provides

[2]The statistical literature offers a number of applications of these ideas, and develops algorithms for the implementation of the mixture index to contingency tables (see, e.g., Dayton, 2003; Liu and Lindsay, 2009).

[3]Other influential approaches provide only a goodness of fit measure; these include Swofford and Whitney (1987), Famulari (1995) and Echenique, Lee, and Shum (2011), whose proposal is to focus on the number of violations of a rationality axiom, e.g. WARP, contained in the data.

a goodness of fit measure and the best description of behavior when applied to the deterministic rational model, it shares the spirit of all these consistency measures. In Appendix D, we formally compare the maximal separation approach to the existing measures and show that it provides a distinctive, novel measure of rationality. Importantly, note that, while all these measures pertain to the analysis of the deterministic rational model, the maximal separation approach applies to any possible rational or non-rational, deterministic or stochastic model of choice.

Recently, Liang (2019) has explored whether the inconsistency part of a choice dataset can be judged as choice error or as the result of preference heterogeneity. More concretely, Liang adopts the flexible multiple-preference framework of Kalai, Rubinstein and Spiegler (2002), in which the individual can use different preferences in different menus.[4] Liang (2019) envisions inconsistencies as the consequence of two different mechanisms: (i) preference heterogeneity, represented by a large fraction of choices being explained by an, ideally, small set of preferences used by the individual á la Kalai, Rubinstein and Spiegler (2002), and (ii) error, represented by a small fraction of choices being captured by other preferences outside this set of preferences. We share with Liang (2019) an interest in identifying the part of the data that is due to error. Our approach differs in two ways: firstly, in that we do not adopt the multiple-preference framework but rather a methodology that applies to any stochastic choice model; and, secondly, as discussed above, in that we provide both a goodness of fit measure and the best description of behavior.

3. Maximal separations

Let X be a non-empty finite set of alternatives. Menus are non-empty subsets of alternatives and, in order to accommodate the diversity of existing settings, such as consumer-type domains or laboratory-type domains, we consider a non-empty arbitrary domain of menus $\mathcal{D}$. Pairs (a, A), with $a \in A$ and $A \in \mathcal{D}$ are called observations, and denoted by $\mathcal{O}$. A stochastic choice function is a mapping $\sigma : \mathcal{O} \to [0, 1]$ which, for every $A \in \mathcal{D}$, satisfies that $\sum_{a \in A} \sigma(a, A) = 1$. We interpret $\sigma(a, A)$ as the probability of choosing alternative a in menu A. We denote by $\mathtt{SCF}$ the space of all stochastic choice

[4]Crawford and Pendakur (2012) implement the approach of Kalai, Rubinstein and Spiegler (2002) using a set of data on milk purchases, finding that five preferences are enough to fully rationalize the data. Apesteguia and Ballester (2010) study the computational complexity of finding the minimal number of multiple-preferences that rationalize the data.

functions. The data are represented by means of a stochastic choice function, that we denote by ρ and that we assume to be in the interior of $\texttt{SCF}$.[5] Namely, $\rho(a, A) > 0$ for every $(a, A) \in \mathcal{O}$. A model is a non-empty closed subset Δ of $\texttt{SCF}$, representing all the possible stochastic choice functions consistent with the entertained theoretical model. We emphasize that other than its closure, we make no further restrictions on the considered model. Accordingly, the model Δ encompasses all the relevant randomness considered by the analyst. This may include a base theoretical model, and considerations on measurement error or unobserved heterogeneity. An instance of the model, that is, a particular stochastic choice function in the set of theoretically admissible ones, is typically denoted by $\delta \in \Delta$.

We say that $\langle\lambda, \delta, \epsilon\rangle \in [0, 1] \times \Delta \times \texttt{SCF}$ is a separation of data ρ whenever $\rho = \lambda\delta+(1-\lambda)\epsilon$. In a separation, we write ρ as a convex combination of the stochastic choice function δ, which contains randomness consistent with model Δ, and the stochastic choice function ϵ, which represents unstructured residual behavior. The fraction of data explained by the model in the separation is given by weight λ. We are particularly interested in explaining the largest possible fraction of data using model Δ. We say that a separation $\langle\lambda^*, \delta^*, \epsilon^*\rangle$ is maximal if there does not exist any other separation $\langle\lambda, \delta, \epsilon\rangle$ with $\lambda > \lambda^*$. The following proposition shows the existence of maximal separations and facilitates their computation.[6]

Proposition 1. *Maximal separations always exist and are characterized by:*

(1) $\lambda^* = \max\limits_{\delta\in\Delta} \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$,

(2) $\delta^* \in \arg\max\limits_{\delta\in\Delta} \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$, *and*

(3) $\epsilon^* = \frac{\rho-\lambda^*\delta^*}{1-\lambda^*}$.

In order to grasp the logic implicit in Proposition 1, let us consider the non-trivial case where $\rho \notin \Delta$. Consider any instance of the model $\delta \in \Delta$. Then, for $\langle\lambda, \delta, \epsilon\rangle$ to be a separation of ρ, the residual stochastic choice function ϵ must lie on the line defined by ρ and δ, with ρ in between δ and ϵ. Now, notice that we can always trivially

[5]This assumption is for expositional convenience; the case of ρ in the boundary of $\texttt{SCF}$ can be trivially dealt with.

[6]In order to avoid the discussion of indeterminacy in fractions throughout the text, we set the ratio $\frac{\rho(a,A)}{0}$ to be strictly larger than any real number. This is a harmless convention, since we could simply replace the expression $\min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$ with $\min_{(a,A)\in\mathcal{O},\delta(a,A)\neq 0} \frac{\rho(a,A)}{\delta(a,A)}$. Moreover, whenever $\lambda^* = 1$, ϵ^* is any stochastic choice function. All proofs are contained in the Appendix.

consider the separation $\langle 0, \delta, \rho \rangle$, where all data is regarded as residual behavior. To obtain larger values of λ with instance δ requires ϵ to depart from ρ in the opposite direction to that taken by δ. Ultimately, λ will be maximal when the residual behavior ϵ reaches the frontier of $\mathtt{SCF}$, i.e., when some observation has either probability zero or one. Indeed, we need only to consider the case $\epsilon(a, A) = 0$, i.e., $\rho(a, A) < \delta(a, A)$ or, equivalently, $\frac{\rho(a,A)}{\delta(a,A)} < 1$. This is because, if $\epsilon(a, A) = 1$ for some observation, we must also have that $\epsilon(b, A) = 0$ for any other alternative $b \in A \setminus \{a\}$. Trivially, $\epsilon(a, A) = 0$ is equivalent to $\lambda = \frac{\rho(a,A)}{\delta(a,A)}$ and hence, the frontier will be first reached by the observation that minimizes the ratio $\frac{\rho(a,A)}{\delta(a,A)}$. Since these observations will play a key role in our analysis, we provide a formal definition here. Given instance δ, the set of observations that minimize the ratio $\frac{\rho(a,A)}{\delta(a,A)}$ are called δ-critical observations and are denoted by $\mathcal{O}_\delta$. Obviously, the maximal fraction of data that can be explained with instance δ is $\min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$, or, equivalently, $\frac{\rho(a,A)}{\delta(a,A)}$ for any $(a, A) \in \mathcal{O}_\delta$. When considering all the possible instances of the model Δ, the result follows.

As already mentioned, Proposition 1 works for arbitrary domains of menus. One domain, which has received a great deal of attention in the stochastic choice literature, is that of binary menus. Since we will also be using this domain in our experimental application, it is worth mentioning that it is one in which Proposition 1 is particularly simple to apply. In essence, notice that any instance of a model will over-predict the probability of choice of one of the alternatives in each binary menu of the domain, while under-predicting the other. Thus, one instance of the model is able to explain a fraction of the data that can be computed by looking at the least over-predicted alternative among all pairs.

4. Particular models of choice

Section 3 characterizes maximal separations for every possible model Δ. We now work with specific choice models. In each case we use Proposition 1, together with the particular structure of the model being studied, to offer tighter results on maximal separations. The models we consider are the deterministic choice model, and three stochastic choice models incorporating different forms of randomness: the tremble model, the Luce model and the single-crossing random utility model. The three stochastic choice models have the deterministic model as a special case, but are mutually independent. Appendix B illustrates the application of each of the results developed here using a simple example involving three alternatives.

4.1. **Deterministic rationality.** The standard economic decision-making model contemplates no randomness whatsoever. Behavior is deterministic and described as the outcome of the maximization of a single preference relation. Thus, in the light of the deterministic model, all behavioral randomness must be regarded as residual behavior. Formally, denote by $\mathcal{P}$ the collection of all strict preference relations, that is, all transitive, complete and asymmetric binary relations on X. Maximization of $P \in \mathcal{P}$ generates the deterministic rational choice function δ_P, which assigns probability one to the maximal alternative in menu A according to preference P. We denote this alternative by $m_P(A)$, i.e., $m_P(A) \in A$ and $m_P(A)Py$ for every $y \in A \setminus \{m_P(A)\}$. Denote by `DET` the model composed of all the deterministic rational choice functions.

The following result shows that the maximal separation for `DET` can be easily computed using a simple recursive structure on subdomains of the data. For presenting the result, some notation will be useful. Given a subset $S \subseteq X$, denote by $\mathcal{D}|_S = \{A \in \mathcal{D} : A \subseteq S\}$ and $\mathcal{O}|_S = \{(a, A) \in \mathcal{O} : A \subseteq S\}$ the corresponding subdomains of menus and observations involving subsets of S. Then:

Proposition 2. *Let $\{\lambda_S\}_{S:\mathcal{D}|_S \neq \emptyset}$ and $P \in \mathcal{P}$ satisfy*

(1) $\lambda_S = \max_{a \in S} \min \Big\{ \{\rho(a, A)\}_{(a,A) \in \mathcal{O}|_S}, \lambda_{S \setminus \{a\}} \Big\}$,

(2) $m_P(S) \in \arg\max_{a \in S} \min \Big\{ \{\rho(a, A)\}_{(a,A) \in \mathcal{O}|_S}, \lambda_{S \setminus \{a\}} \Big\}$.[7]

Then, $\langle \lambda_X, \delta_P, \frac{\rho - \lambda_X \delta_P}{1 - \lambda_X} \rangle$ is a maximal separation for the deterministic model.

Proposition 2 enables a recursive computation of maximal separations for `DET`. More precisely, the algorithm constructs a maximal separation for each restriction of data ρ to a subdomain of menus $\mathcal{D}|_S$, starting with subdomains where $\mathcal{D}|_S = \{S\}$, i.e., menus for which there are no available data in proper subsets. In these menus, only the highest choice frequency of an alternative must be considered. The maximal separation can be constructed by considering the preference relation that places the alternative with the highest choice frequency above all other alternatives. For any other subdomain $\mathcal{D}|_S$, the algorithm must analyze the alternatives $a \in S$ one by one, again considering the consequences of placing a as the maximal alternative in S. It turns out to be the case that we just need to consider the following values: (i) the choice frequencies of a in

[7]Notice that equations (1) and (2) always compute a minimum over a non-empty collection of values. This is so because the computation only takes place when $\mathcal{D}|_S$ is non-empty and, hence, either $a \in A$ for some $A \subseteq S$, or $\mathcal{D}|_{S \setminus \{a\}} \neq \emptyset$.

subsets of S, and (ii) the maximal fractions over the subdomains where alternative a is not present.[8]

4.2. **Tremble model.** In tremble models, behavioral randomness is interpreted as a mistake at the moment of choice. In the simplest version, the individual contemplates a preference relation P. With probability $(1-\gamma) \in [0,1]$, the preference is maximized. With probability γ, the individual trembles and randomizes between all the alternatives in the menu.[9] This generates the tremble choice function $\delta_{[P,\gamma]}(a,A) = \frac{\gamma}{|A|}$ whenever $a \in A \setminus \{m_P(A)\}$ and $\delta_{[P,\gamma]}(m_P(A),A) = 1-\gamma\frac{|A|-1}{|A|}$. Denote by `Tremble` the model composed of all tremble choice functions. The next result describes the maximal fraction of data explained by `Tremble` and a maximal separation for `Tremble`.

Proposition 3. *Let $\{\lambda_S(\gamma)\}_{S:\mathcal{D}|_S\neq\emptyset}$ and $P(\gamma) \in \mathcal{P}$ satisfy, for every $\gamma \in [0,1]$:*

(1) $\lambda_S(\gamma) = \max\limits_{a\in S} \min \left\{ \{\frac{|A|\rho(a,A)}{(1-\gamma)|A|+\gamma}\}_{(a,A)\in\mathcal{O}|_S}, \{\frac{|A|\rho(b,A)}{\gamma}\}_{\substack{(b,A)\in\mathcal{O}|_S,\\ b\neq a\in A}}, \lambda_{S\setminus\{a\}}(\gamma) \right\}$,

(2) $m_{P(\gamma)}(S) \in \arg\max\limits_{a\in S} \min \left\{ \{\frac{|A|\rho(a,A)}{(1-\gamma)|A|+\gamma}\}_{(a,A)\in\mathcal{O}|_S}, \{\frac{|A|\rho(b,A)}{\gamma}\}_{\substack{(b,A)\in\mathcal{O}|_S,\\ b\neq a\in A}}, \lambda_{S\setminus\{a\}}(\gamma) \right\}$.

Let γ^* be the tremble value that maximizes $\lambda_X(\gamma)$. Then, $\langle \lambda_X(\gamma^*), \delta_{[P(\gamma^*),\gamma^*]}, \frac{\rho-\lambda_X(\gamma^*)\delta_{[P(\gamma^*),\gamma^*]}}{1-\lambda_X(\gamma^*)} \rangle$ is a maximal separation for the tremble model.

Given the immediate connection to the rational deterministic model, the intuition of the result is analogous to that in Proposition 2.[10]

4.3. **Luce model.** Denote by $\mathcal{U}$ the collection of strictly positive utility functions u such that, without loss of generality, $\sum_{x\in X} u(x) = 1$. Given $u \in \mathcal{U}$, a strictly positive Luce stochastic choice function is defined by $\delta_u(a,A) = \frac{u(a)}{\sum_{b\in A} u(b)}$ with $a \in A \in \mathcal{D}$. In order to accommodate the Luce model in our framework we consider the closure of the

[8] A particularly interesting example involves binary domains in which some stochastic transitivity property is satisfied. In this case, it is easy to see that the identified preference will be consistent with the stochastic revealed preference.

[9] See Harless and Camerer (1994) for an early treatment of the trembling-hand concept in the stochastic choice literature.

[10] As in the deterministic case with binary domains, where choice satisfies stochastic transitivity, the maximal separation for the tremble model identifies the preference relation that is consistent with the stochastic revealed preference. Hence, in this case, both the deterministic and the tremble models identify the same preference relation. Interestingly, this is exactly the case in our empirical application. However, as we show in Appendix B, the maximal separations for the deterministic and tremble models do not, in general, necessarily identify the same preference relation.

set of strictly positive Luce stochastic choice functions, which we denote by `Luce`.[11] We write δ_L to denote a generic, not necessarily strictly positive, Luce stochastic choice function. However, as shown in the proof of Proposition 4, there are always instances of the model of Luce identified in the maximal separations that are strictly positive, and hence, the former assumption is inconsequential.

We now describe the structure of maximal separations of `Luce`. From Proposition 1 we know that the study of a particular instance of model δ_L requires us to analyze its critical observations $\mathcal{O}_{\delta_L}$. It turns out to be the case that, under the Luce model, we only need to check for a simple condition on the set $\mathcal{O}_{\delta_L}$.

Proposition 4. *$\langle \min_{(a,A)} \frac{\rho(a,A)}{\delta^*_L(a,A)}, \delta^*_L, \frac{\rho-\lambda^*\delta^*_L}{1-\lambda^*}\rangle$ is a maximal separation for the Luce model if and only if $\mathcal{O}_{\delta^*_L}$ contains a sub-collection $\{(a_i,A_i)\}_{i=1}^I$ such that $\bigcup_{i=1}^I\{a_i\} = \bigcup_{i=1}^I A_i$.*

Proposition 4 provides a simple means to obtain maximal separations for the Luce model, which involves checking whether the critical observations of a Luce stochastic choice function satisfy a cyclical property. From here, the computation of λ^* and ϵ^* follows in the usual manner. To explain the intuition of the proof, consider a strictly positive instance of `Luce` given by $u \in \mathcal{U}$ and its critical observations $\mathcal{O}_{\delta_u}$. Proposition 4 states that if there is a subcollection $\{(a_i,A_i)\}_{i=1}^I \subseteq \mathcal{O}_{\delta_u}$ such that $\bigcup_{i=1}^I\{a_i\} = \bigcup_{i=1}^I A_i$, then δ_u is part of a maximal separation. To see this, suppose that there exists a separation using another Luce instance δ_v explaining a larger fraction of the data. It must be the case that δ_v assigns lower Luce probabilities to all the critical observations of δ_u. Clearly, reducing the Luce probability in (a_1,A_1) requires that one alternative in A_1, say a_2, is such that $v(a_2)/v(a_1) > u(a_2)/u(a_1)$. However, since there exists a critical observation of the form (a_2,A_2), we need to find another alternative in A_2, say a_3, with $v(a_3)/v(a_2) > u(a_3)/u(a_2)$. Given that $\bigcup_{i=1}^I\{a_i\} = \bigcup_{i=1}^I A_i$, this line of reasoning leads to a cycle, and consequently, we cannot improve the ρ/δ ratio of all the critical observations of δ_u, which shows its maximality. The situation is entirely different when there is $x \in \bigcup_{i=1}^I A_i \setminus \bigcup_{i=1}^I\{a_i\}$. In this case, we can find an improvement by moving the Luce values in the direction of alternative x, that is, by increasing the Luce utility of x and reducing all the rest by the same proportion. It can be shown that this logic ultimately makes all ρ/δ ratios of the critical observations of δ_u to increase.

[11] Effectively, the added stochastic choice functions have zero choice probabilities in some observations, and Luce-type behavior otherwise. See Echenique and Saito (2018) and Horan (2019) for studies of the treatment of zero choice probabilities in models à la Luce.

Notice that the proof of Proposition 4 lays down a simple algorithm with which to identify maximal separations for the Luce model. One can start with any vector of weights $u \in \mathcal{U}$ and check the stated property. If the property is satisfied, $\delta^* = \delta_u$, which leads to the computation of λ^* and ϵ^*. Otherwise, $\bigcup_{i=1}^{I} A_i \setminus \bigcup_{i=1}^{I}\{a_i\}$ is non-empty, allowing one of the alternatives in this set to be selected and the utilities to be moved along the segment $\alpha\mathbf{1}_x + (1-\alpha)u$, where $\mathbf{1}_x$ is a function assigning a value 1 to x and a value 0 to any other alternative. Eventually, this leads to a new Luce vector which explains a strictly larger fraction of the data, and the characterizing property can be re-verified. This ascending algorithm yields the maximal separation.

4.4. **Single-crossing random utility model.** In random utility models (RUMs), there exists a probability distribution μ over the set of all possible preferences $\mathcal{P}$. At the choice stage, a preference is realized according to μ, and maximized, thereby determining the choice probabilities $\delta_\mu(a, A) = \sum_{P \in \mathcal{P}: a = m_P(A)} \mu(P)$, for every $(a, A) \in \mathcal{O}$. In other words, the choice probability of a given alternative within a menu is given by the sum of the probability masses associated to the preferences where the alternative is maximal within the menu.

The literature has often considered these models as complex to work with, and offered models in restricted domains that facilitate their use in applications. Here, we focus on the single-crossing random utility models (SCRUMs), which are RUMs over a set of preferences satisfying the single-crossing condition.[12] Formally, SCRUMs consider probability distributions μ on a given ordered collection of preferences $\mathcal{P}' = \{P_1, P_2, \ldots, P_T\}$, satisfying the single-crossing condition $P_j \cap P_1 \subseteq P_i \cap P_1$ if and only if $j \geq i$. That is, the preference over a pair of alternatives x and y reverses once at most in the ordered collection of preferences. We denote the set of SCRUM stochastic choice functions by $\mathtt{SC}$. Proposition 5 characterizes the maximal separations for SCRUMs.

Proposition 5. *Let* $\lambda_1 = \min_{A \in \mathcal{D}} \rho(m_{P_1}(A), A)$ *and* $\delta_{\mu_1} = \delta_{P_1}$, *and for every* $i \in \{2, \ldots, T\}$ *define recursively*

(1) $\lambda_i = \min_{A \in \mathcal{D}} \left\{ \rho(m_{P_i}(A), A) + \max_{j: j<i, m_{P_j}(A) \neq m_{P_i}(A)} \lambda_j \right\}$,

(2) $\delta_{\mu_i} = (1 - \frac{\lambda_{i-1}}{\lambda_i})\delta_{P_i} + \frac{\lambda_{i-1}}{\lambda_i}\delta_{\mu_{i-1}}$.

Then, $\langle \lambda_T, \delta_{\mu_T}, \frac{\rho - \lambda_T \delta_{\mu_T}}{1 - \lambda_T} \rangle$ *is a maximal separation for SCRUM.*

[12] See Apesteguia, Ballester and Lu (2017) for a study of this model. Other RUMs using restricted domains are Gul and Pesendorfer (2006) and Lu and Saito (2017).

Proposition 5 provides a smooth recursive method with which to obtain a maximal separation. It basically computes the maximal fraction of data, λ_i, that can be explained by SCRUMs using preferences up to P_i. Trivially, the maximal fraction of data explained by P_1 is $\min_{A\in\mathcal{D}} \rho(m_{P_1}(A), A)$. Now consider any other preference $P_i \in \mathcal{P}'$ and assume that every preference P_j, $j < i$, has been analyzed. With the extra preference P_i, and for a given menu A, we can rationalize data $\rho(m_{P_i}(A), A)$ together with any other data $\rho(x, A)$, $x \neq m_{P_i}(A)$, that is rationalized by preferences preceding P_i. This can be achieved by considering the appropriate linear combination of the constructed SCRUM that uses preferences up to P_{i-1} with preference P_i.

5. An empirical application

Here we use an experimental dataset to operationalize the maximal separation results obtained in the previous section.[13] There were nine equiprobable monetary lotteries, described in Table 1. Each of the 87 participants faced 108 different menus of lotteries, including all 36 binary menus and a random sample of larger menus.[14] There were two treatments. Treatment NTL was a standard implementation, with no time limit on the choice. In treatment TL, subjects had to select a lottery within a limited time. At the end of the experiment, one of the menus was chosen at random and the subject was paid according to his or her choice from that menu.[15]

Table 1. Lotteries

$l_1 = (17)$	$l_4 = (30, 10)$	$l_7 = (40, 12, 5)$
$l_2 = (50, 0)$	$l_5 = (20, 15)$	$l_8 = (30, 12, 10)$
$l_3 = (40, 5)$	$l_6 = (50, 12, 0)$	$l_9 = (20, 12, 15)$

To ensure a sufficiently large number of data points per menu, we focus on the choices made in the binary menus, which, when aggregating both treatments, gives a

[13]We collected the experimental data together with Syngjoo Choi at UCL in March 2013, within the context of another research project. This is the first completed paper to use this dataset. We are very grateful to Syngjoo for kindly allowing us to use this dataset.

[14]There were menus of 2, 3 and 5 alternatives, presented one at a time, in a randomized order. No participant was presented more than once with the same menu of alternatives. The location of the lotteries on the screen was randomized, as was the location of the monetary prizes within a lottery.

[15]Specifically, subjects had 5, 7 and 9 seconds for the menus of 2, 3, and 5 alternatives, respectively.

total of 87 data points per menu.[16] Table 2 reports the choice probabilities in each of the binary menus. It also reports the optimal and the residual stochastic choice functions identified in the maximal separation results, using the models described in the previous section. In SCRUM we use the CRRA expected utility representation, which is by far the most widely used utility representation for risk preferences.[17] There are several lessons to be learned from the table.

First note that the maximal fractions of the data explained by the respective models increase from the deterministic choice model, to the tremble model, to the Luce model and, finally, to the SCRUM-CRRA model. It is worth noting that the deterministic model already explains about half of the data, i.e., 0.51.[18] The identified optimal instance is the one associated with the preference $l_1Pl_5Pl_4Pl_8Pl_7Pl_9Pl_3Pl_6Pl_2$. The top alternative, lottery l_1, is the safest, since it gives £17 with probability one. The next is lottery l_5, which has the second lowest variance at the expense of a very low expected return. Lottery l_2, the one with the highest expected value and highest variance, is regarded as the worst alternative. Hence, the deterministic model pictures a population that is essentially highly risk-averse. The model reaches its explanatory limits with the critical observation $(l_8, \{l_7, l_8\})$ where, by Proposition 1, the ratio of observed to predicted probability is minimal. Specifically, the observed choice probability is 0.51 while the deterministic prediction is 1. The ratio of these two values gives the fraction of data explained by the model, 0.51.

The tremble model identifies exactly the same preference as the deterministic model, while increasing the maximal fraction of the data explained from 0.51 to 0.68. This is the result of using a relatively large tremble probability, $\gamma = 0.51$. The tremble model

[16]Due to the time limit in one of the treatments, the number is slightly lower for some menus. Specifically, there are 18 menus with 87 data points, 12 with 86, 3 with 85 and 3 with 84.

[17]The CRRA Bernoulli function is $\frac{x^{1-r}}{1-r}$, whenever $r \neq 1$, and $\log x$ otherwise, with x representing money. We have also studied the cases of CARA expected utility, and mean-variance utility, and obtained similar results, which are available upon request. Note that SCRUM with CRRA is but a generalization of the random parameter model that we use in Apesteguia and Ballester (2018), in the sense that the former does not impose any probability distribution over the set of preferences.

[18]In order to put this result into perspective, consider Crawford and Pendakur's (2002) study of Danish household survey data on the purchase of six different types of milk. They find that a single preference relation is sufficient to rationalize 64% of the data. The Houtman-Maks index gives a consistency level of 66%. In Appendix D we review this index, arguing that it is slightly more flexible than the application of the maximal separation technique to the deterministic model, which explains the higher consistency found in the data.

TABLE 2. Data and Maximal Separations

(a, A)	ρ	DET $\delta^*_{\texttt{DET}}$	DET $\epsilon^*_{\texttt{DET}}$	TREMBLE $\delta^*_{\texttt{Tremble}}$	TREMBLE $\epsilon^*_{\texttt{Tremble}}$	LUCE $\delta^*_{\texttt{Luce}}$	LUCE $\epsilon^*_{\texttt{Luce}}$	SCRUM-CRRA $\delta^*_{\texttt{SC-CRRA}}$	SCRUM-CRRA $\epsilon^*_{\texttt{SC-CRRA}}$
$(l_1, \{l_1, l_2\})$	0.75	1.00	0.49	0.74	0.75	0.91	0.30	0.74	0.77
$(l_1, \{l_1, l_3\})$	0.60	1.00	0.19	0.74	0.29	0.71	0.28	0.55	0.78
$(l_2, \{l_2, l_3\})$	0.33	0.00	0.67	0.26	0.50	0.20	0.69	0.24	0.66
$(l_1, \{l_1, l_4\})$	0.53	1.00	0.05	0.74	0.07	0.62	0.27	0.47	0.75
$(l_2, \{l_2, l_4\})$	0.28	0.00	0.56	0.26	0.32	0.15	0.64	0.24	0.40
$(l_3, \{l_3, l_4\})$	0.43	0.00	0.86	0.26	0.78	0.40	0.50	0.42	0.46
$(l_1, \{l_1, l_5\})$	0.58	1.00	0.16	0.74	0.24	0.46	0.92	**0.47**	**1.00**
$(l_2, \{l_2, l_5\})$	0.25	0.00	0.51	0.26	0.25	0.08	0.73	0.26	0.23
$(l_3, \{l_3, l_5\})$	0.45	0.00	0.92	0.26	0.87	**0.26**	**1.00**	0.45	0.46
$(l_4, \{l_4, l_5\})$	0.49	0.00	0.99	0.26	0.98	0.34	0.89	0.53	0.33
$(l_1, \{l_1, l_6\})$	0.72	1.00	0.44	0.74	0.68	0.87	0.31	0.76	0.60
$(l_2, \{l_2, l_6\})$	0.44	0.00	0.89	0.26	0.84	0.42	0.51	0.42	0.53
$(l_3, \{l_3, l_6\})$	0.80	1.00	0.60	0.74	0.93	**0.74**	**1.00**	0.79	0.84
$(l_4, \{l_4, l_6\})$	0.76	1.00	0.51	0.74	0.79	0.81	0.62	0.76	0.76
$(l_5, \{l_5, l_6\})$	0.75	1.00	0.49	0.74	0.75	0.89	0.35	0.76	0.71
$(l_1, \{l_1, l_7\})$	0.63	1.00	0.25	0.74	0.38	0.77	0.23	0.74	0.22
$(l_2, \{l_2, l_7\})$	0.24	0.00	0.49	0.26	0.22	0.26	0.21	0.26	0.19
$(l_3, \{l_3, l_7\})$	0.48	0.00	0.96	0.26	0.94	0.57	0.20	0.53	0.27
$(l_4, \{l_4, l_7\})$	0.62	1.00	0.24	0.74	0.37	0.67	0.49	0.76	0.14
$(l_5, \{l_5, l_7\})$	0.63	1.00	0.26	0.74	0.40	0.79	0.18	0.76	0.18
$(l_6, \{l_6, l_7\})$	0.27	0.00	0.54	0.26	0.29	0.33	0.10	0.24	0.36
$(l_1, \{l_1, l_8\})$	0.64	1.00	0.27	0.74	0.42	0.67	0.57	0.76	0.21
$(l_2, \{l_2, l_8\})$	0.22	0.00	0.45	0.26	0.15	0.17	0.36	0.26	0.09
$(l_3, \{l_3, l_8\})$	0.36	0.00	0.73	0.26	0.58	0.45	0.12	0.45	0.03
$(l_4, \{l_4, l_8\})$	0.56	1.00	0.12	0.74	0.18	0.55	0.60	0.56	0.56
$(l_5, \{l_5, l_8\})$	0.62	1.00	0.23	0.74	0.36	0.70	0.40	0.76	0.13
$(l_6, \{l_6, l_8\})$	0.20	0.00	0.40	0.26	0.07	0.23	0.12	0.24	0.04
$(l_7, \{l_7, l_8\})$	0.49	**0.00**	**1.00**	**0.26**	**1.00**	0.37	0.83	0.42	0.77
$(l_1, \{l_1, l_9\})$	0.76	1.00	0.51	0.74	0.78	0.74	0.81	0.79	0.62
$(l_2, \{l_2, l_9\})$	0.28	0.00	0.56	0.26	0.32	0.23	0.42	0.28	0.28
$(l_3, \{l_3, l_9\})$	0.39	0.00	0.79	0.26	0.68	**0.53**	**0.00**	0.45	0.17
$(l_4, \{l_4, l_9\})$	0.55	1.00	0.08	0.74	0.13	0.63	0.32	0.53	0.60
$(l_5, \{l_5, l_9\})$	0.83	1.00	0.65	**0.74**	**1.00**	**0.76**	**1.00**	1.00	0.20
$(l_6, \{l_6, l_9\})$	0.22	0.00	0.44	0.26	0.14	0.29	0.02	0.26	0.08
$(l_7, \{l_7, l_9\})$	0.56	1.00	0.12	0.74	0.18	0.46	0.87	0.45	0.96
$(l_8, \{l_8, l_9\})$	0.64	1.00	0.26	0.74	0.41	0.58	0.78	**0.53**	**1.00**
	λ^*_Δ	0.51		0.68		0.74		0.78	

Note: (a, A) denotes the observation referring to alternative a from menu A, ρ the observed percentage of choosing lottery a from menu A, and $\langle \lambda^*_\Delta, \delta^*_\Delta, \epsilon^*_\Delta \rangle$ the maximal separation of ρ for model $\Delta \in \{\texttt{DET}, \texttt{Tremble}, \texttt{Luce}, \texttt{SC-CRRA}\}$. Data entries in bold refer to the menus containing the critical observations in the respective model.

is characterized by critical observations $(l_8, \{l_7, l_8\})$ and $(l_9, \{l_5, l_9\})$. As in the deterministic case, choice data is scarce for l_8 versus l_7, but the problem is less severe thanks to the presence of a tremble, due to which, the individual is predicted to choose l_8 only with probability 0.74, thereby reducing the ratio of observed to predicted probabilities to 0.68. This ratio cannot be improved beyond this point. Although increasing the tremble probability would increase this ratio, it would also decrease the ratio of the other critical observation, $(l_9, \{l_5, l_9\})$, which has the same value of 0.68. To see this, notice the choice prediction for alternative l_9, being worse than alternative l_5, corresponds entirely to the tremble probability, and hence, an increase in tremble would increase the predicted probability and thus decrease the ratio.

The Luce model is able to explain close to three quarters of the data. The optimal utility values for the lotteries are $u = (0.22, 0.02, 0.09, 0.13, 0.25, 0.03, 0.07, 0.11, 0.08)$, which again suggest a highly risk averse population. The alternative with the highest Luce utility value is lottery l_5, followed by lottery l_1, while lottery l_2, which is the riskiest, has the lowest Luce utility value. That is, although u does not represent $P_{\texttt{DET}}$ exactly, it represents a preference very close to it. Interestingly, we see that the Luce model can accommodate a larger fraction of the data by allowing randomness to depend on the cardinal evaluation of alternatives. The model is hard pressed to explain observations $(l_5, \{l_3, l_5\})$, $(l_6, \{l_3, l_6\})$, $(l_3, \{l_3, l_9\})$ and $(l_9, \{l_5, l_9\})$, that represent the type of cyclical structure described in Proposition 4. In each of these observations, the ratio of observed to predicted probabilities is equal to 0.74. Increasing any of these ratios would require decreasing the utility of one alternative in $\{l_3, l_5, l_6, l_9\}$, but only, of course, at the expense of the ratio of another of these critical observations.

Finally, `SC-CRRA` explains as much as nearly 80% of the data. In so doing, it assigns positive masses to 10 of the 30 possible CRRA preferences, with the largest probability mass, 0.44, associated with the most risk averse CRRA preference, i.e., preference $l_1 P l_5 P l_9 P l_8 P l_4 P l_7 P l_3 P l_6 P l_2$, which is again very close to $P_{\texttt{DET}}$. Since each preference compatible with CRRA corresponds to an interval of risk aversion levels, we can completely describe the optimal `SC-CRRA` instance by reporting the values of the cumulative distribution function at the upper bounds of these intervals. These are $F(-4.15) = 0.205$, $F(-0.31) = 0.241$, $F(-0.08) = 0.242$, $F(0.34) = 0.258$, $F(0.41) = 0.276$, $F(0.44) = 0.416$, $F(0.61) = 0.453$, $F(1) = 0.533$, $F(4) = 0.563$ and $F(\infty) = 1$. Notice that, in addition to explaining a large fraction of the data, `SC-CRRA` is also rich enough to show that a quarter of the population is risk loving,

$F(-0.08) = 0.242$. The limits of SC-CRRA in explaining the data are reached at observations $(l_5, \{l_1, l_5\})$ and $(l_9, \{l_8, l_9\})$. On the one hand, lottery l_5 is preferred over lottery l_1 by all CRRA levels with a risk aversion level below 2, which has an accumulated mass of 0.533. Given the observed choices, this leads to a critical ratio for observation $(l_5, \{l_1, l_5\})$ of 0.78. Improving this ratio would necessarily require us to assign a higher weight to levels of risk aversion above 2. However, this would immediately conflict with the ratio of l_9 to l_8, since l_9 is ranked above l_8 at all levels of risk aversion above 1. As the ratio of observed to predicted data for $(l_9, \{l_8, l_9\})$ also has the critical value of 0.78, no improvement can take place.[19]

To conclude the discussion of Table 2, we would like to emphasize that the four models are very consistent in the qualitative judgment of the population. All four models take the population of subjects to be highly risk averse. Then, we see that, by introducing different sources of randomness, it is possible to explain larger fractions of the data, and that the precise source of randomness affects the fraction of the data explained.

6. Other goodness of fit measures

The maximal separation exercise identifies a best instance of the model $\delta^* \in \Delta$ and an expression of residual behavior $\epsilon^* \in \texttt{SCF}$ such that, combined at rates λ^* and $1-\lambda^*$, generate data ρ. The value λ^* is a tight upper bound for the fraction of data that can be explained by the model. Thus, the exercise provides a goodness of fit measure of model Δ to data ρ. There are other well-known measures in the literature that partially share the structure of the maximal separation measure, in the sense that they also identify one instance of the model that maximizes a notion of closeness to the data.[20] For the sake of comparison, we adopt the standard language of minimization of loss functions and talk of lack of fit all along the section.[21]

[19]In Appendix C we use this dataset to analyze the maximal separation using the random expected utility of Gul and Pesendorfer (2006).

[20]We say that the other measures partially share the structure of maximal separations because they do not identify (minimal) expressions of residual behavior. This component, that we believe may be crucial in the understanding of actual behavior and revision of theoretical models, is unique to maximal separations.

[21]By lack of fit, sometimes also known as badness of fit, we mean the mirror notion of goodness of fit; basically, how poorly a model fits the data.

Formally, a loss function is a map $L : \Delta \times \rho \to \mathbb{R}_+$ that measures the deviation of every instance $\delta \in \Delta$ with respect to data ρ. The lack of fit and the best instance of the model follow immediately from the minimization of the loss function among the different instances of the model.[22] Now, in a maximal separation, the minimal fraction of data unexplained, $1-\lambda^*$, represents a measure of the lack of fit that can be written as the minimization of a loss function. From Proposition 1 we know that $1-\lambda^* = 1-\max_{\delta\in\Delta} \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)} = \min_{\delta\in\Delta}[\max\limits_{(a,A)\in\mathcal{O}}(1-\frac{\rho(a,A)}{\delta(a,A)})]$ and hence, we can write the maximal separation loss function as $L_{MS}(\delta,\rho) = \max\limits_{(a,A)\in\mathcal{O}}[1-\frac{\rho(a,A)}{\delta(a,A)}]$.

Two other important goodness of fit measures are maximum likelihood and least squares. The maximum likelihood exercise involves the minimization of the Kullback-Leibler divergence from δ to ρ, and this can be written as the minimization of the loss function $L_{ML}(\delta,\rho) = \sum_{(a,A)\in\mathcal{O}} \rho(a,A)\log\frac{\rho(a,A)}{\delta(a,A)}$.[23] Similarly, least squares involves the minimization of the quadratic loss function $L_{LS}(\delta,\rho) = \sum_{(a,A)\in\mathcal{O}}(\delta(a,A)-\rho(a,A))^2$.

Inspecting the loss functions, it is immediately clear that the maximal separation measure is different to those defined by maximum likelihood and least squares. Crucially, while the maximal separation is concerned with the largest deviation between the data and the specified model, maximum likelihood and least squares aggregate the deviations across the different observations. This has two implications. Firstly, there should be datasets and models where maximal separation identifies different best instances of the model. Secondly, we should expect maximal separation to provide more accurate over-estimations for those observations for which the frequency of observed choice is low, while the other measures would perform better on average. In what follows we use our experimental dataset to empirically illustrate these two points.

Table 3 illustrates the first point. It reports the instances of the models identified by the maximal separation and the maximum likelihood techniques over the entire

[22] Notice that, when the model Δ is the deterministic rational model of choice, lack of fit merely corresponds to a notion of irrationality of the data. As mentioned in Section 2, most measures of irrationality, including Afriat, Varian, Houtman-Maks and the Swaps Index adopt this minimization structure. For the specific case of the deterministic model, Appendix D formally compares the maximal separation approach with other rationality measures.

[23] The Kullback-Leibler divergence can be interpreted as the amount of information lost due to the use of δ instead of ρ

Table 3. Maximal Separation and Maximum Likelihood

	Deterministic
MS	$P = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]$
ML	$P = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]$
	Tremble
MS	$P = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]; \quad \gamma = 0.51$
ML	$P = [l_1, l_5, l_4, l_8, l_7, l_9, l_3, l_6, l_2]; \quad \gamma = 0.68$
	Luce
MS	$u = (0.22, 0.02, 0.09, 0.13, 0.25, 0.03, 0.07, 0.11, 0.08)$
ML	$u = (0.18, 0.04, 0.1, 0.14, 0.17, 0.04, 0.11, 0.13, 0.09)$
	SCRUM-CRRA
MS	$F(-4.15) = 0.205, F(-0.31) = 0.241, F(-0.08) = 0.242, F(0.34) = 0.258, F(0.41) = 0.276$
	$F(0.44) = 0.416, F(0.61) = 0.453, F(1) = 0.533, F(4) = 0.563, F(\infty) = 1$
ML	$F(-4.15) = 0.22, F(-0.31) = 0.287, F(0.44) = 0.442$
	$F(1) = 0.506, F(4) = 0.563, F(\infty) = 1$

Note: MS and ML denote maximal separation and maximum likelihood, respectively. P denotes the preference identified in the corresponding case, where the ranking declines from left to right, γ is the tremble probability in `Tremble`, u is the Luce utility vector associated with `Luce`, where the i-th entry in u corresponds to the utility value of lottery l_i, and finally $F(r)$ denotes the cumulative probability masses associated with the upper bounds of the intervals of the relative risk aversion coefficients r consistent with those CRRA preference relations that have a strictly positive mass in the corresponding estimation procedure.

dataset.[24] With respect to the deterministic model, no difference whatsoever is observed, as exactly the same preference relation is estimated. This ordinal equivalence is preserved in the case of the tremble model, although our technique predicts a substantially smaller trembling coefficient, $0.51 < 0.68$. The intuition for this difference is straightforward. Recall that, as we mentioned above, $(l_9, \{l_5, l_9\})$ is a critical observation in the maximal separation exercise for `Tremble`. The observed probability in this observation is small, 0.17, and the identified instance of the model for our technique predicts, due to the trembling parameter, a rather large relative frequency of 0.26. However, the maximum likelihood exercise is not severely affected by this local

[24]In the ML calculations we impose a lower bound in the theoretical predictions, in order to ensure strictly positive likelihoods. Least squares give practically identical results to maximum likelihood, and hence it is omitted.

consideration and makes the estimation only by averaging over all the observations. Consequently, the estimation exercise in maximum likelihood is willing to sacrifice the prediction quality of this extreme observation in order to favor the prediction over other moderate ones. This is done by increasing substantially the trembling parameter and consequently the prediction in this particular observation $(l_9, \{l_5, l_9\})$, reaching a disproportionate value of 0.34, two times the observed value. A similar reasoning applies to the comparison of the cases of `Luce` and `SC-CRRA`.

In order to illustrate the comparison of the different approaches, we now perform an out-of-sample exercise. This, in turn, allows us to evaluate the second conjecture stated above.[25] We take all the binary data except for one binary set, estimate the instances of the models by maximal separation and maximum likelihood using these data, and use the estimated instances to predict the behavior in the omitted binary set. We perform this procedure on 36 binary sets. For each binary set, there are two cases: one in which both maximal separation and maximum likelihood over-estimate the choice probability of the same alternative in the binary menu, and another in which they over-estimate the choice probability of different alternatives. By focusing on the first case, comparison of the predictive powers of maximal separation and maximum likelihood becomes straightforward; one of the methods is unambiguously more accurate than the other.[26] We therefore focus our comparison on these menus, since the conclusions may otherwise depend on the particular distance function employed. Table 4 reports the results.[27] As announced above, the analysis of the loss functions involved in the two techniques suggested a very intuitive conjecture. Namely, that the maximal separation technique is very cautious and can therefore be expected to perform better in observations with low choice probabilities. This conjecture is largely confirmed in

[25]Effectively, this second conjecture could also be evaluated using the in-sample estimations. These results, which closely mimic the pattern obtained with the out-of-sample exercise, are reported in the online appendix. In addition, the out-of-sample exercise is complemented in Appendix C by using the non-binary part of the dataset, while, in Appendix B, we elaborate further on the intuition for this conjecture by adopting a more theoretical approach, using a particular data-generating process and the tremble model.

[26]Notice that, in binary menus, if one alternative is over-estimated, the other is under-estimated and, for both observations, there is one method that is more accurate than the other. Thus, discussing the results for, say, the over-estimated alternatives implies no loss of generality.

[27]We do not report the results of the deterministic method, since, in this case, the maximal separation and maximum likelihood predictions are exactly the same.

Table 4. Forecasting Results of Maximal Separation and Maximum Likelihood

Tremble				Luce				SCRUM-CRRA			
(a, A)	ρ	MS	ML	(a, A)	ρ	MS	ML	(a, A)	ρ	MS	ML
$(l_9, \{l_5, l_9\})$	0.17	**0.28**	0.35	$(l_9, \{l_5, l_9\})$	0.17	**0.39**	0.41	$(l_6, \{l_3, l_6\})$	0.20	**0.21**	0.27
$(l_6, \{l_3, l_6\})$	0.20	**0.26**	0.35	$(l_6, \{l_3, l_6\})$	0.20	**0.27**	0.32	$(l_6, \{l_6, l_8\})$	0.20	**0.26**	0.29
$(l_6, \{l_6, l_8\})$	0.20	**0.26**	0.35	$(l_6, \{l_6, l_8\})$	0.20	**0.23**	0.25	$(l_6, \{l_6, l_9\})$	0.22	**0.26**	0.29
$(l_6, \{l_6, l_9\})$	0.22	**0.26**	0.34	$(l_6, \{l_6, l_9\})$	0.22	**0.29**	0.34	$(l_2, \{l_2, l_8\})$	0.22	**0.26**	0.29
$(l_2, \{l_2, l_8\})$	0.22	**0.26**	0.34	$(l_2, \{l_2, l_7\})$	0.24	**0.26**	0.29	$(l_6, \{l_4, l_6\})$	0.24	**0.24**	0.29
$(l_6, \{l_4, l_6\})$	0.24	**0.26**	0.34	$(l_9, \{l_1, l_9\})$	0.24	**0.26**	0.36	$(l_2, \{l_2, l_7\})$	0.24	**0.26**	0.29
$(l_2, \{l_2, l_7\})$	0.24	**0.26**	0.34	$(l_3, \{l_3, l_8\})$	0.36	0.49	*0.46*	$(l_2, \{l_1, l_2\})$	0.25	**0.26**	0.29
$(l_9, \{l_1, l_9\})$	0.24	**0.26**	0.34	$(l_9, \{l_8, l_9\})$	0.36	**0.42**	0.43	$(l_2, \{l_2, l_5\})$	0.25	**0.26**	0.29
$(l_2, \{l_1, l_2\})$	0.25	**0.26**	0.34	$(l_3, \{l_3, l_9\})$	0.39	0.63	*0.56*	$(l_3, \{l_3, l_8\})$	0.36	0.47	*0.45*
$(l_2, \{l_2, l_5\})$	0.25	**0.26**	0.34	$(l_5, \{l_1, l_5\})$	0.42	0.58	*0.50*	$(l_9, \{l_8, l_9\})$	0.36	0.55	*0.52*
$(l_6, \{l_5, l_6\})$	0.25	**0.26**	0.34	$(l_9, \{l_7, l_9\})$	0.44	0.54	*0.46*	$(l_3, \{l_3, l_9\})$	0.39	**0.45**	0.48
$(l_3, \{l_3, l_9\})$	0.39	0.74	*0.66*	$(l_8, \{l_4, l_8\})$	0.44	**0.45**	0.49	$(l_3, \{l_1, l_3\})$	0.40	0.45	*0.45*
$(l_2, \{l_2, l_6\})$	0.44	0.74	*0.66*	$(l_8, \{l_7, l_8\})$	0.51	0.59	*0.55*	$(l_5, \{l_1, l_5\})$	0.42	0.56	*0.56*
$(l_4, \{l_4, l_5\})$	0.49	0.74	*0.66*	$(l_5, \{l_4, l_5\})$	0.51	0.68	*0.54*	$(l_9, \{l_7, l_9\})$	0.44	0.58	*0.58*
$(l_7, \{l_7, l_8\})$	0.49	0.75	*0.66*	$(l_1, \{l_1, l_4\})$	0.53	0.63	*0.56*	$(l_9, \{l_4, l_9\})$	0.45	**0.47**	0.50
$(l_7, \{l_3, l_7\})$	0.52	0.74	*0.66*	$(l_5, \{l_3, l_5\})$	0.55	0.78	*0.65*	$(l_4, \{l_1, l_4\})$	0.47	0.53	*0.51*
$(l_1, \{l_1, l_4\})$	0.53	0.74	*0.66*	$(l_4, \{l_4, l_9\})$	0.55	0.64	*0.63*	$(l_3, \{l_3, l_7\})$	0.48	0.53	*0.51*
$(l_5, \{l_3, l_5\})$	0.55	0.74	*0.66*	$(l_4, \{l_3, l_4\})$	0.57	0.60	*0.60*	$(l_4, \{l_4, l_5\})$	0.49	0.53	*0.51*
$(l_4, \{l_4, l_9\})$	0.55	0.74	*0.66*	$(l_1, \{l_1, l_3\})$	0.60	0.71	*0.66*	$(l_8, \{l_7, l_8\})$	0.51	0.61	*0.57*
$(l_7, \{l_7, l_9\})$	0.56	0.74	*0.66*	$(l_3, \{l_2, l_3\})$	0.67	0.80	*0.71*	$(l_5, \{l_3, l_5\})$	0.55	**0.55**	0.56
$(l_4, \{l_3, l_4\})$	0.57	0.74	*0.66*	$(l_4, \{l_2, l_4\})$	0.72	0.85	*0.78*	$(l_6, \{l_2, l_6\})$	0.56	0.58	*0.56*
$(l_1, \{l_1, l_3\})$	0.60	0.74	*0.66*	$(l_1, \{l_1, l_6\})$	0.72	0.87	*0.83*	$(l_5, \{l_5, l_8\})$	0.62	0.76	*0.72*
$(l_5, \{l_5, l_8\})$	0.62	0.74	*0.66*	$(l_1, \{l_1, l_2\})$	0.75	0.91	*0.82*	$(l_4, \{l_4, l_7\})$	0.62	0.79	*0.75*
$(l_4, \{l_4, l_7\})$	0.62	0.74	*0.66*	$(l_5, \{l_2, l_5\})$	0.75	0.92	*0.80*	$(l_1, \{l_1, l_7\})$	0.63	0.76	*0.72*
$(l_1, \{l_1, l_7\})$	0.63	0.74	*0.66*	$(l_5, \{l_5, l_6\})$	0.75	0.89	*0.81*	$(l_5, \{l_5, l_7\})$	0.63	0.76	*0.72*
$(l_5, \{l_5, l_7\})$	0.63	0.74	*0.66*	$(l_4, \{l_4, l_6\})$	0.76	0.83	*0.78*	$(l_1, \{l_1, l_8\})$	0.64	0.76	*0.72*
$(l_8, \{l_8, l_9\})$	0.64	0.74	*0.66*					$(l_3, \{l_2, l_3\})$	0.67	0.76	*0.72*
$(l_1, \{l_1, l_8\})$	0.64	0.74	*0.66*					$(l_1, \{l_1, l_9\})$	0.76	**0.84**	0.90
$(l_8, \{l_3, l_8\})$	0.64	0.74	*0.66*					$(l_5, \{l_5, l_9\})$	0.83	1.00	1.00

Note: (a, A) denotes the observation referring to alternative a from menu A such that a is the lottery where the predictions of both maximal separation (MS) and maximum likelihood (ML) are above the observed choice data ρ. Those observations for which one of the predictions of MS or ML is above the observed choice data and the other below are not reported in the table. Then, for each one of the models, the binary menus of lotteries are ordered from lower to higher observed choice probabilities. Bold entries refer to the cases where MS is closer to the data and italicized entries refers to those cases where ML is closer to the data.

our analysis. In all three models, the over-estimation of small probabilities is less problematic for the maximal separation technique, while maximum likelihood deals better with the over-estimation of large probabilities. We conclude, therefore, that if one is interested in forecasting exercises, these results suggest that, to obtain a clear picture of the overall situation, it may be useful to apply both estimation techniques: maximal separation and maximum likelihood.

7. Discussion

We close this paper by commenting on three issues surrounding the notion of maximal separations. We begin by discussing how to select one of the available existing models by assigning a cost to the parsimony of each model. We then comment on the possibility of assuming that the model Δ is not only closed, but convex. Finally, we discuss the possibility of restricting the space of residual stochastic choice functions, and comment on possible interpretations of residual behavior.

7.1. **Model selection.** The analyst may be interested in comparing the explanatory performance of various models. Clearly, the fraction of data explained in a maximal separation constitutes an absolute measure of performance. As usual, the absolute measure of performance is in tension with the idea of over-fitting, i.e., larger models are explanatorily superior simply because of their size. As a direct example of this tension, notice that whenever $\Delta \subseteq \Delta'$, the maximal fraction of data explained by model Δ' is, independently of ρ, larger than or equal to the maximal fraction of data explained by model Δ. The natural reaction to this is to consider a penalization of model Δ that is monotonically dependent upon the size of the model.[28]

Notice that the set of all stochastic choice functions can be built as a product of $|\mathcal{D}|$ simplices. In other words, the set of all stochastic choice functions can be seen as a subset of $[0,1]^{|\mathcal{O}|-|\mathcal{D}|}$. Since all relevant stochastic models have a strictly lower dimensionality, they all have zero Lebesgue measure in the subspace of all stochastic choice functions. Hence, any measure based on the Lebesgue volume of these models would regard all models alike in terms of their size, and would differentiate them only in terms of the fraction of the data they rationalize.[29]

[28]Another approach would entail comparing the completeness of the different models, that is, the amount of predictive variation rationalized by the model. See Fudenberg, Kleinberg, Liang and Mullainathan (2019) for a recent formal treatment of the notion of completeness.

[29]Another normalization that would not discriminate beyond absolute performance is $\frac{\lambda^*-\lambda^{\min}}{\lambda^{\max}-\lambda^{\min}}$, where $\lambda^{\max}$ and $\lambda^{\min}$ are a models' maximum and minimum performance values when studying all possible datasets. Clearly, $\lambda^{\max}=1$ for all the models, and it can be easily shown that $\lambda^{\min}=0$ for all the models discussed in this paper.

An alternative approach, in the spirit of the Akaike information criterion, would be to consider a cost that depends on the largest value of n such that the model Δ has non-zero measure in the space $[0,1]^n$. This line of thought essentially counts the number of parameters in the model Δ. Considering the models previously analyzed, the deterministic choice model corresponds to a finite subset of possible datasets, and hence, does not have a strictly positive dimension. The tremble model involves one tremble parameter and hence, has dimension 1. The Luce model involves one utility value for each alternative and, normalizing the sum of utility values, involves as many parameters as the number of alternatives minus 1. The single-crossing random utility model involves a probability measure over a subset of T preferences. If all menus are available, the dimension of this model is $T-1$.

7.2. **Convex models.** We have assumed model Δ to be closed, a basic property which guarantees the existence of maximal separations. An obvious further property to be considered is convexity, especially in relation to mixture models. These models are common when dealing with heterogeneity at the population level, and they can also be used to discuss intra-personal heterogeneity. In a mixture model, the researcher convexifies a set of instances of a base model, allowing different subpopulations to be explained by different instances of the model. Notice that our methodology directly enables this type of analysis, since one can simply consider the desired, convexified, Δ model as the object of analysis. As an example of this approach, see the analysis of the single-crossing random utility model in section 4.4, which can be understood as the convex hull of a subset of instances of the deterministic model.

The convexity of Δ may have useful implications. Given data ρ and model Δ, consider two separations $\langle \lambda, \delta, \epsilon \rangle$ and $\langle \lambda', \delta', \epsilon' \rangle$, and let $\alpha \in [0,1]$. Clearly, $\alpha\lambda + (1-\alpha)\lambda' \in [0,1]$ and $\alpha\epsilon + (1-\alpha)\epsilon' \in \texttt{SCF}$, due to the convexity of $[0,1]$ and $\texttt{SCF}$. Whenever model Δ is convex, we also obtain that $\alpha\delta + (1-\alpha)\delta' \in \Delta$, and hence, $\alpha\langle \lambda, \delta, \epsilon \rangle + (1-\alpha)\langle \lambda', \delta', \epsilon' \rangle$ is also a separation, showing the convexity of the set of all separations. This transforms the search for maximal separations into a convex optimization problem.

It is important to note, however, that convex models of choice are the exception rather than the norm. It is immediately obvious, for example, that the deterministic model is not convex. A mixture of two deterministic choice functions rationalized by two different preferences will clearly lead to a stochastic choice function that cannot be rationalized by any other preference. In a similar vein, it is well-known that the

Luce model represents another case of a non-convex model (see Gul, Natenzon and Pesendorfer, 2014). Hence, the assumption of convexity, while not required for our results, would come with some loss of generality.

7.3. **Residual behavior.** In our approach to finding the maximal fraction of the data consistent with the model, we have given the best possible chance to the model by leaving the space of possible residual behaviors completely unstructured. That is, we have assumed that residual behavior ϵ can be selected from the whole set of stochastic choice functions, $\texttt{SCF}$. Consequently, as the proof of Proposition 1 shows, a necessary condition for a separation to be maximal is that residual behavior lies exactly on the frontier of $\texttt{SCF}$. In other words, the residual behavior in a maximal separation imposes that the choice probability for some observations, which we describe as critical, is zero.

On occasions, one may be interested in separations involving less extreme residual behaviors. That is, one may entertain the possibility of imposing on the space of allowable residual behaviors a particular minimal structure beyond that of a stochastic choice function. A possible objective might be to consider the case in which residual behavior is in some way similar in nature to the model of reference Δ, while allowing for more flexibility.

A set of minimal assumptions is sufficient to guarantee that the logic behind our methodology is applicable for considering restricted spaces of residual behavior, $\texttt{RB} \subseteq \texttt{SCF}$. In particular, we only need to consider that: (i) the space of residual behaviors is a relaxation of the model, i.e., $\Delta \subseteq \texttt{RB}$, (ii) the data belong to the space of residual behaviors, i.e., $\rho \in \texttt{RB}$, and (iii) the space of residual behaviors has some technical properties similar to those of the space $\texttt{SCF}$, such as being closed and convex. Under these conditions, one can reformulate the concept of separation, requiring that $\langle \lambda, \delta, \epsilon \rangle \in [0, 1] \times \Delta \times \texttt{RB}$. The logic of Proposition 1 remains valid and a necessary condition for a separation to be maximal will be that residual behavior lies on the frontier of $\texttt{RB}$.

We conclude by highlighting some final comments on the interpretation of residual behavior. We distinguish three cases. First, consider the situation in which the residual has the consistency properties that are typical of a noisy structure. To illustrate, consider that the data ρ are generated exactly by the tremble model using a preference P and tremble γ, but the analyst initially approaches the data from the perspective of the deterministic model. The maximal separation will identify the true preference P and the residual will have a very transparent structure: the optimal alternative in menu A according to P is chosen with zero probability, while any other alternative is chosen with probability $\frac{1}{|A|-1}$. Clearly, the structure of ϵ is very informative about the existing noise, and the analyst may wish to reconsider the candidate model for incorporating such noise into the deterministic model to effectively use the tremble model.

Secondly, suppose that the residual has consistency properties typical of a competing instance of the model or of a competing model. To illustrate, consider that the data ρ are generated by a mixture of preferences P (in larger proportion) and P', but the analyst initially approaches the data from the perspective of the deterministic model. The maximal separation will identify preference P and the residual will have a very clear structure: that of preference P'. Clearly, the structure of ϵ is very informative about the existing heterogeneity, and the analyst may wish to reconsider the candidate model for incorporating such heterogeneity into a mixture model. Similar reasoning can be applied when the residual resembles not an instance of the model Δ but an instance of some other reasonable model Δ'.

Finally, suppose that the residual appears rather inconsistent to the analyst. Here, a potentially fruitful option would be to apply the maximal separation approach on ϵ using some reasonable model of choice, to assess the possibility of making any sense out of the apparently chaotic behavior ϵ. That is, try to ascertain whether ϵ itself can be understood to a significant extent as the combination of some choice model, and as another expression of residual behavior.

Appendix A. Proofs

Proof of Proposition 1: Consider first the case where $\rho \in \Delta$. Then, $\langle 1, \rho, \rho \rangle$ is clearly a maximal separation. Moreover, given that $\min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)} = 1$ if and only if $\rho = \delta$, the result follows.

Let us now consider the case of $\rho \notin \Delta$. We start by claiming that, for a given $\delta \in \Delta$, there exist $\lambda \in [0,1)$ and $\epsilon \in \mathtt{SCF}$ such that $\langle \lambda, \delta, \epsilon \rangle$ is a separation if and only if $\lambda \leq \min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$. To prove the 'only if' part, assume that $\langle \lambda, \delta, \epsilon \rangle$ is a separation. Then, it must be the case that $\rho = \lambda\delta + (1-\lambda)\epsilon$, or equivalently, $\frac{\rho-\lambda\delta}{1-\lambda} = \epsilon \geq 0$. This implies that $\rho - \lambda\delta \geq 0$ and, ultimately, that $\lambda \leq \frac{\rho}{\delta}$. Hence, it must be that $\lambda \leq \min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$, as desired.[30] To prove the 'if' part, suppose that $\lambda \leq \min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$.

[30]Notice that, in dividing by δ, we are using the above-mentioned convention.

We now prove that $\langle \lambda, \delta, \epsilon = \frac{\rho - \lambda\delta}{1-\lambda} \rangle$ is a separation of data. Since by assumption $\delta \in \Delta$ and the construction guarantees that $\rho = \lambda\delta + (1-\lambda)\epsilon$, we are only required to prove that $\epsilon \in \mathtt{SCF}$. We begin by checking that $\epsilon(a, A) \geq 0$ holds for every $(a, A) \in \mathcal{O}$. To see this, suppose by contradiction that this is not true. Then, there would exist $(b, B) \in \mathcal{O}$ such that $\frac{\rho(b,B) - \lambda\delta(b,B)}{1-\lambda} < 0$. This would imply that $\rho(b, B) - \lambda\delta(b, B) < 0$ and hence, that $\delta(b, B) > 0$, with $\frac{\rho(b,B)}{\delta(b,B)} < \lambda \leq \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$, which is a contradiction. Finally, it is also the case that $\sum\limits_{a\in A} \epsilon(a, A) = \sum\limits_{a\in A} \frac{\rho(a,A)-\lambda\delta(a,A)}{1-\lambda} = \frac{1-\lambda}{1-\lambda} = 1$ for every $A \in \mathcal{D}$. Therefore $\epsilon \in \mathtt{SCF}$ and the claim is proved.

Now, the former claim shows that the maximal fraction that can be explained with model $\{\delta\}$ is $\min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta(a,A)}$. This argument immediately implies the desired results on Δ, provided that maximal separations exist.

We now show the existence of maximal separations. Given the domain, any separation $\langle \lambda, \delta, \epsilon \rangle$ of ρ is a vector in $\mathbb{R}^n$, with $n = 2|\mathcal{O}| + 1$. We first prove that the set of separations is a closed subset of $\mathbb{R}^n$. Consider a sequence of separations $\langle \lambda_t, \delta_t, \epsilon_t \rangle_{t=1}^{\infty}$ and suppose that this sequence converges in $\mathbb{R}^n$. Given the finite-dimensionality and the fact that Δ and $\mathtt{SCF}$ are closed, we clearly have that $\lim_t \lambda_t \in [0, 1]$, $\lim_t \delta_t \in \Delta$ and $\lim_t \epsilon_t \in \mathtt{SCF}$ and it is evident that $\langle \lim_t \lambda_t, \lim_t \delta_t, \lim_t \epsilon_t \rangle$ is a separation of ρ. This proves that the set of separations is closed and, being a subset of $[0, 1]^n$, it is also bounded and hence, compact. Since the maximal fraction of data explained can be thought as the result of maximizing, over the set of separations, the projection map assigning the first component of the separation, i.e., value λ, existence is guaranteed.$\blacksquare$

Proof of Proposition 2: Let $\{\lambda_S\}_{S:\mathcal{D}|_S \neq \emptyset}$ and $P \in \mathcal{P}$ satisfy (1) and (2). For every S such that $\mathcal{D}|_S \neq \emptyset$, denote by $\mathtt{DET}_{\mathcal{D}|_S}$ the deterministic rational stochastic choice functions defined over the subdomain $\mathcal{D}|_S$. Similarly, denote by $\rho|_S$ the restriction of ρ to $\mathcal{D}|_S$. We start by proving, recursively, that the maximal fraction of data $\rho|_S$ explained by model $\mathtt{DET}_{\mathcal{D}|_S}$ is equal to λ_S. Consider any subset S for which $\mathcal{D}|_S = \{S\}$. In this case, Proposition 1 guarantees that the maximal fraction of data $\rho|_S$ explained by model $\mathtt{DET}_{\mathcal{D}|_S}$ is $\max\limits_{\delta\in\mathtt{DET}_{\mathcal{D}|_S}} \min\limits_{(a,A)\in\mathcal{O}|_S} \frac{\rho|_S(a,A)}{\delta(a,A)} = \max\limits_{P\in\mathcal{P}} \min\limits_{(a,A)\in\mathcal{O}|_S} \frac{\rho(a,A)}{\delta_P(a,A)} = \max\limits_{P\in\mathcal{P}} \min\limits_{a\in S} \frac{\rho(a,S)}{\delta_P(a,S)} = \max\limits_{P\in\mathcal{P}} \frac{\rho(m_P(S),S)}{\delta_P(m_P(S),S)} = \max\limits_{P\in\mathcal{P}} \rho(m_P(S), S) = \max\limits_{a\in S} \rho(a, S) = \max\limits_{a\in S} \min\limits_{(a,A)\in\mathcal{O}|_S} \rho(a, A) = \lambda_S$. Now suppose that $\mathcal{D}|_S \neq \{S\}$ and that the result has been proved for any strict subset of S with non-empty subdomain. For any $a \in S$, denote by $\mathcal{P}_{aS}$ the set of preferences that rank a above any other alternative in S, i.e., $\mathcal{P}_{aS} = \{P \in \mathcal{P} : a = m_P(S)\}$, and by $\mathtt{aS}$ the subset of $\mathtt{DET}_{\mathcal{D}|_S}$ generated by preferences in $\mathcal{P}_{aS}$. Trivially, $\mathtt{DET}_{\mathcal{D}|_S} = \bigcup_{a\in S} \mathtt{aS} = \bigcup_{a\in S} \bigcup_{P\in\mathcal{P}_{aS}} \{\delta_P\}$. Since the only observations for which δ_P has a non-null value are those that are in the form $(m_P(A), A)$, Proposition 1 guarantees that the maximal fraction of data $\rho|_S$ explained by model $\mathtt{DET}_{\mathcal{D}|_S}$ is $\max_{a\in S} \max_{P\in\mathcal{P}_{aS}} \min_{A\in\mathcal{D}|_S} \rho(m_P(A), A)$. Since $P \in \mathcal{P}_{aS}$, we obtain that $m_P(A) = a$ whenever $a \in A$ and hence, the latter value is equal to $\max_{a\in S} \max_{P\in\mathcal{P}_{aS}} \min\Big\{\{\rho(a,A)\}_{(a,A)\in\mathcal{O}|_S}, \{\rho(m_P(B),B)\}_{B\in\mathcal{D}|_{S\setminus\{a\}}}\Big\}$. This can be expressed as $\max_{a\in S}\min\Big\{\{\rho(a,A)\}_{(a,A)\in\mathcal{O}|_S}, \max_{P\in\mathcal{P}_{aS}} \min_{B\in\mathcal{D}|_{S\setminus\{a\}}} \rho(m_P(B),B)\Big\}$ or, equivalently, as $\max_{a\in S}\min\Big\{\{\rho(a,A)\}_{(a,A)\in\mathcal{O}|_S}, \min_{B\in\mathcal{D}|_{S\setminus\{a\}}} \max_{P\in\mathcal{P}_{aS}} \min_{C\in\mathcal{D}|_B} \rho(m_P(C),C)\Big\}$. Given that $a \notin B$, it is clearly the case that $\max_{P\in\mathcal{P}_{aS}} \min_{C\in\mathcal{D}|_B} \rho(m_P(C),C) = \max_{P\in\mathcal{P}} \min_{C\in\mathcal{D}|_B} \rho(m_P(C),C)$ and, by Proposition 1 and the structure of deterministic stochastic choice functions, the latter is the maximal fraction of data $\rho|_B$ explained by model $\mathtt{DET}_{\mathcal{D}|_B}$, which is equal to λ_B by hypothesis. Hence, the maximal fraction of data $\rho|_S$ explained by model $\mathtt{DET}_{\mathcal{D}|_S}$ must be also equal to λ_S, as desired. As a corollary, we have that for the maximal fraction of the data explained by the deterministic model is λ_X and from the construction, the claim follows. ■

Proof of Proposition 3: Since the proof has the same structure as the proof of Proposition 2, we skip some of the steps and use the same notation as before. We start by (recursively) proving that the maximal fraction of data $\rho|_S$ explained by the collection of stochastic choice functions in $\mathtt{Tremble}_{\mathcal{D}|_S}$ with a fixed degree of tremble γ, which we denote by $\mathtt{Tremble}_{\mathcal{D}|_S}(\gamma)$, is equal to $\lambda_S(\gamma)$. We start with any subset S for which $\mathcal{D}|_S = \{S\}$. The maximal fraction of data $\rho|_S$ explained by $\mathtt{Tremble}_{\mathcal{D}|_S}(\gamma)$ is $\max_{\delta\in\mathtt{Tremble}_{\mathcal{D}|_S}(\gamma)} \min_{(a,A)\in\mathcal{O}|_S} \frac{\rho|_S(a,A)}{\delta(a,A)} = \max_{P\in\mathcal{P}}\min\big\{\frac{\rho(m_P(S),S)}{\delta_{[P,\gamma]}(m_P(S),S)}, \{\frac{\rho(b,S)}{\delta_{[P,\gamma]}(b,S)}\}_{b\in S\setminus\{m_P(S)\}}\big\} = \max_{P\in\mathcal{P}}\min\big\{\frac{|S|\rho(m_P(S),S)}{(1-\gamma)|S|+\gamma}, \{\frac{|S|\rho(b,S)}{\gamma}\}_{b\in S\setminus\{m_P(S)\}}\big\} = \max_{a\in S}\min\big\{\frac{|S|\rho(a,S)}{(1-\gamma)|S|+\gamma}, \{\frac{|S|\rho(b,S)}{\gamma}\}_{b\in S\setminus\{a\}}\big\} = \max_{a\in S}\min\Big\{\{\frac{|A|\rho(a,A)}{(1-\gamma)|A|+\gamma}\}_{(a,A)\in\mathcal{O}|_S}, \{\frac{|A|\rho(b,A)}{\gamma}\}_{\substack{(b,A)\in\mathcal{O}|_S\\ b\neq a}}\big\} = \lambda_S(\gamma)$. Whenever $\mathcal{D}|_S \neq \{S\}$, we can write the maximal fraction of data $\rho|_S$ explained by model $\mathtt{Tremble}_{\mathcal{D}|_S}(\gamma)$ as $\max_{a\in S} \max_{P\in\mathcal{P}_{aS}} \min\big\{\{\frac{|A|\rho(m_P(A),A)}{(1-\gamma)|A|+\gamma}\}_{A\in\mathcal{D}|_S}, \{\frac{|A|\rho(b,A)}{\gamma}\}_{(b,A)\in\mathcal{O}|_S, b\neq m_P(A)}\big\}$. Notice that we can decompose $\{\frac{|A|\rho(m_P(A),A)}{(1-\gamma)|A|+\gamma}\}_{A\in\mathcal{D}|_S}$ into $\{\frac{|A|\rho(a,A)}{(1-\gamma)|A|+\gamma}\}_{(a,A)\in\mathcal{O}|_S}$ and $\{\frac{|B|\rho(m_P(B),B)}{(1-\gamma)|B|+\gamma}\}_{B\in\mathcal{D}|_{S\setminus\{a\}}}$. Similarly, we can decompose $\{\frac{|A|\rho(b,A)}{\gamma}\}_{(b,A)\in\mathcal{O}|_S, b\neq m_P(A)}$ into components $\{\frac{|A|\rho(b,A)}{\gamma}\}_{\substack{(b,A)\in\mathcal{O}|_S\\ b\neq a\in A}}$ and $\{\frac{|B|\rho(b,B)}{\gamma}\}_{\substack{B\in\mathcal{D}|_{S\setminus\{a\}}\\ b\neq m_P(B)}}$. By the same reasoning as in the proof of Proposition 2, consideration of both $\{\frac{|B|\rho(m_P(B),B)}{(1-\gamma)|B|+\gamma}\}_{B\in\mathcal{D}|_{S\setminus\{a\}}}$ and $\{\frac{|B|\rho(b,B)}{\gamma}\}_{\substack{B\in\mathcal{D}|_{S\setminus\{a\}}\\ b\neq m_P(B)}}$ yields the value $\{\lambda_B(\gamma)\}_{B\in\mathcal{D}|_{S\setminus\{a\}}}$. This proves the claim. From Proposition 2, the maximal separations for model $\texttt{Tremble}_{\mathcal{D}}(\gamma)$ explain a fraction $\lambda_X(\gamma)$ of the data. Since $\texttt{Tremble}_{\mathcal{D}} = \cup_\gamma \texttt{Tremble}_{\mathcal{D}}(\gamma)$, one simply needs to consider the value γ^* maximizing $\lambda_X(\gamma)$ and the result follows immediately from Proposition 1. ■

Proof of Proposition 4: To prove the 'if' part let $\delta_L \in \texttt{Luce}$ and suppose that there exists $\{(a_i, A_i)\}_{i=1}^I \subseteq \mathcal{O}_{\delta_L}$ such that $\bigcup_{i=1}^I\{a_i\} = \bigcup_{i=1}^I A_i$. From Proposition 1, the maximal fraction that can be explained by model $\{\delta_L\}$ is $\min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_L(a,A)}$. Assume, by way of contradiction, that δ_L is not part of a maximal separation for the Luce model. Therefore, there exists $\langle\lambda^*, \delta_L^*, \epsilon^*\rangle$ such that, for every $i \in \{1, 2, \dots, I\}$, $\frac{\rho(a_i,A_i)}{\delta_L(a_i,A_i)} = \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_L(a,A)} < \lambda^* = \min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_L^*(a,A)} \leq \frac{\rho(a_i,A_i)}{\delta_L^*(a_i,A_i)}$. For every $i \in \{1, 2, \dots, I\}$, we have that $\rho(a_i, A_i) > 0$ and hence, since the ρ/δ_L ratio is minimized at $\mathcal{O}_{\delta_L}$, it must be that $\delta_L(a_i, A_i) > 0$, making $\frac{\rho(a_i,A_i)}{\delta_L(a_i,A_i)} < \frac{\rho(a_i,A_i)}{\delta_L^*(a_i,A_i)}$ equivalent to $\delta_L^*(a_i, A_i) < \delta_L(a_i, A_i)$. Let $\{\delta'_{v_n}\}_{n=1}^\infty$ and $\{\delta_{u_n}\}_{n=1}^\infty$ be two sequences of strictly positive Luce stochastic choice functions that converge to δ_L^* and δ_L, respectively. Select an m sufficiently large that $\delta_L^*(a_i, A_i) < \delta_{u_m}(a_i, A_i)$ holds for every $i \in \{1, 2, \dots, I\}$. Given m, now select an m' sufficiently large that, for every $i \in \{1, 2, \dots, I\}$, $\delta'_{v_{m'}}(a_i, A_i) < \delta_{u_m}(a_i, A_i)$ holds. We then have that $\frac{1}{\sum_{x\in A_i}\frac{v_{m'}(x)}{v_{m'}(a_i)}} = \frac{v_{m'}(a_i)}{\sum_{x\in A_i} v_{m'}(x)} = \delta'_{v_{m'}}(a_i, A_i) < \delta_{u_m}(a_i, A_i) = \frac{u_m(a_i)}{\sum_{x\in A_i} u_m(x)} = \frac{1}{\sum_{x\in A_i}\frac{u_m(x)}{u_m(a_i)}}$, thus guaranteeing, for every $i \in \{1, 2, \dots, I\}$, the existence of one alternative $\bar{x}_i \in A_i \setminus \{a_i\}$ such that $\frac{v_{m'}(a_i)}{v_{m'}(\bar{x}_i)} < \frac{u_m(a_i)}{u_m(\bar{x}_i)}$. Given that $\bigcup_{i=1}^I\{a_i\} = \bigcup_{i=1}^I A_i$, there exists a subcollection $\{a_{i_h}\}_{h=1}^H$ of $\{a_i\}_{i=1}^I$ with the following properties: (i) $a_{i_{h+1}} \in A_{i_h}$, with $h = 1, \dots, H-1$, and $a_{i_1} \in A_{i_H}$, and (ii) $\frac{v_{m'}(a_{i_h})}{v_{m'}(a_{i_{h+1}})} < \frac{u_m(a_{i_h})}{u_m(a_{i_{h+1}})}$ with $h = 1, \dots, H-1$ and $\frac{v_{m'}(a_{i_H})}{v_{m'}(a_{i_1})} < \frac{u_m(a_{i_H})}{u_m(a_{i_1})}$. Obviously, $1 = \frac{v_{m'}(a_{i_H})}{v_{m'}(a_{i_1})}\Pi_{h=1}^{H-1}\frac{v_{m'}(a_{i_h})}{v_{m'}(a_{i_{h+1}})} < \frac{u_m(a_{i_H})}{u_m(a_{i_1})}\Pi_{h=1}^{H-1}\frac{u_m(a_{i_h})}{u_m(a_{i_{h+1}})} = 1$, which is a contradiction. This concludes the 'if' part of the proof.

To prove the 'only if' part, suppose that δ_L belongs to a maximal separation for the Luce model. Let $[x]$ be the set of all alternatives $x' \in X$ for which there exists a sequence of observations $\{(b_j, B_j)\}_{j=1}^J$, with: (i) $x = b_1$ and $x' \equiv b_{J+1} \in B_J$, and (ii) for every $j \in \{1, 2, \dots, J\}$, $\delta_L(b_j, B_j) > 0$ and $\delta_L(b_{j+1}, B_j) > 0$. If there is no alternative for which such a sequence exists, let $[x] = \{x\}$. Clearly, $[\cdot]$ defines equivalence classes on X. Whenever there exists $A \in \mathcal{D}$ with $\{x, y\} \subseteq A$ and $\delta_L(x, A) > \delta_L(y, A) = 0$, we write $[x] \succ [y]$. We claim that $\succ$ is an acyclic relation on the set of equivalence classes. To see this, assume, by contradiction, that there is a cycle of pairs $\{a_q, b_q\}$, menus $A_q \supseteq \{a_q, b_q\}$, and equivalence classes $[x_q]$, $q \in \{1, 2, \ldots, Q\}$, such that: (i) $\delta_L(a_q, A_q) > \delta_L(b_q, A_q) = 0$ for every $q \in \{1, 2, \ldots, Q\}$, (ii) $a_q \in [x_q]$ for every $q \in \{1, 2, \ldots, Q\}$, and (iii) $b_q \in [x_{q+1}]$ for every $q \in \{1, 2, \ldots, Q-1\}$ and $b_Q \in [x_1]$. We can then consider a sequence of stochastic choice functions $\{\delta_{u_n}\}_{n=1}^{\infty}$ that converges to δ_L. Since b_q and a_{q+1} belong to the same equivalence class $[x_{q+1}]$, either $b_q = a_{q+1}$ or there exists a sequence of observations $\{(d_j, D_j)\}_{j=1}^{J}$ with: (i) $b_q = d_1$ and $a_{q+1} = d_{J+1} \in D_J$, and (ii) for every $j \in \{1, 2, \ldots, J\}$, $\delta_L(d_j, D_j) > 0$ and $\delta_L(d_{j+1}, D_j) > 0$ (and the same holds for a_Q and b_1). Define the strictly positive constant $K_q = 1$ whenever $b_q = a_{q+1}$, and $K_q = \frac{1}{2}\Pi_{j=1}^{J} \frac{\delta_L(d_j, D_j)}{\delta_L(d_{j+1}, D_j)}$ otherwise (with a similar definition for K_Q relating a_Q and b_1). If $b_q = a_{q+1}$, then trivially $u_n(b_q) = u_n(a_{q+1})$ for every n. Otherwise, for an n sufficiently large in the sequence $\{u_n\}_{n=1}^{\infty}$, we have that $\frac{u_n(b_q)}{u_n(a_{q+1})} = \Pi_{j=1}^{J} \frac{u_n(d_j)}{u_n(d_{j+1})} = \Pi_{j=1}^{J} \frac{\delta_{u_m}(d_j, D_j)}{\delta_{u_m}(d_{j+1}, D_j)} \geq K_q$. Hence, in any case, $\frac{u_n(b_q)}{K_q} \geq u_n(a_{q+1})$ holds for any sufficiently large n (and the same holds for b_Q and a_1). Also, since $\delta_L(a_q, A_q) > \delta_L(b_q, A_q) = 0$ for every $q \in \{1, 2, \ldots, Q\}$, we can find an n sufficiently large that $u_n(a_q) > \frac{u_n(b_q)}{K_q}$. Hence, we can find an m that is sufficiently large that $u_m(a_1) > \frac{u_m(b_1)}{K_1} \geq u_m(a_2) > \frac{u_m(b_2)}{K_2} \geq \cdots \geq u_m(a_Q) > \frac{u_m(b_Q)}{K_Q} \geq u_m(a_1)$. This is a contradiction which proves the acyclicity of $\succ$. We can then denote the equivalence classes as $\{[x_e]\}_{e=1}^{E}$, where $[x_e] \succ [x_{e'}]$ implies that $e < e'$. For an equivalence class $[x_e]$, define the vector $u_{[x_e]} \in \mathcal{U}$ such that $u_{[x_e]}(y) = 0$ if $y \notin [x_e]$ and, $\frac{u_{[x_e]}(y)}{u_{[x_e]}(y')} = \frac{\delta_L(y, A)}{\delta_L(y', A)}$ whenever $y, y' \in [x_e]$, $\delta_L(y, A) > 0$ and $\delta_L(y', A) > 0$. This is clearly well-defined due to the structure of Luce stochastic choice functions. Now consider the sequence of Luce stochastic choice functions $\{\delta_{v_n}\}_{n=1}^{\infty}$ given by $v_n = (1 - \sum_{e=2}^{E} (\frac{1}{2^e})^n) u_{[x_1]} + \sum_{e=2}^{E} (\frac{1}{2^e})^n u_{[x_e]}$, which clearly converges to δ_L. Consider the following three collections of observations $\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_3$. $\mathcal{O}_1$ is composed of all observations $(a, A) \in \mathcal{O}$ such that $A \subseteq [a]$. $\mathcal{O}_2$ is composed of all observations $(a, A) \in \mathcal{O} \setminus \mathcal{O}_1$, such that $b \in A$, $a \in [a_i]$ and $b \in [a_j]$ imply $i \geq j$. $\mathcal{O}_3$ is composed of observations in $\mathcal{O} \setminus (\mathcal{O}_1 \cup \mathcal{O}_2)$. Notice that, for an n sufficiently large, for every $(a, A) \in \mathcal{O}_1$ we have that $\frac{\rho(a,A)}{\delta_{v_n}(a,A)} = \frac{\rho(a,A)}{\delta_L(a,A)}$ and for every $(a, A) \in \mathcal{O}_2$ we have that $\frac{\rho(a,A)}{\delta_{v_n}(a,A)} > \frac{\rho(a,A)}{\delta_L(a,A)}$. Also, for an n sufficiently large, $(\frac{1}{2})^n < \min\{\rho(a, A) : a \in A \in \mathcal{D}\}$, and hence $(a, A) \in \mathcal{O}_3$ implies that $\frac{\rho(a,A)}{\delta_{v_n}(a,A)} \geq \frac{\rho(a,A)}{(\frac{1}{2})^m} > 1$. In this case, we can fix an m sufficiently large that, from Proposition 1, $\min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_{v_m}(a,A)} = \min\limits_{(a,A)\in\mathcal{O}_1\cup\mathcal{O}_2} \frac{\rho(a,A)}{\delta_{v_m}(a,A)} \geq$

$\min_{(a,A)\in\mathcal{O}_1\cup\mathcal{O}_2} \frac{\rho(a,A)}{\delta_L(a,A)} \geq \min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_L(a,A)}$.[31] Indeed, since δ_L belongs to a maximal separation of the model of Luce, it must be that $\min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_{v_m}(a,A)} = \min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_L(a,A)}$, and hence $\mathcal{O}_1$ is non-empty, with $\mathcal{O}_{\delta_{v_m}} \subseteq \mathcal{O}_{\delta_L} \subseteq \mathcal{O}_1$.

Assume, by way of contradiction, that there is no subcollection $\{(a_i, A_i)\}_{i=1}^I \subseteq \mathcal{O}_{\delta_L}$ such that $\bigcup_{i=1}^I \{a_i\} = \bigcup_{i=1}^I A_i$. Then, for every subcollection $\{(a_i, A_i)\}_{i=1}^I \subseteq \mathcal{O}_{\delta_{v_m}}$ it must also be that $\bigcup_{i=1}^I \{a_i\} \neq \bigcup_{i=1}^I A_i$. Hence, there must exist at least one alternative x such that $x \neq a$ for every $(a, A) \in \mathcal{O}_{\delta_{v_m}}$ and $x \in A$ for some $(a, A) \in \mathcal{O}_{\delta_{v_m}}$. Consider the segment $\alpha\mathbf{1}_x + (1 - \alpha)v_m$, with $\alpha \in [0, 1]$. Select the maximal separation in this segment, which can be identified as follows. Partition the set of observations into two classes $\mathcal{O}' = \{(a, A) \in \mathcal{O}, a \neq x \in A\}$ and $\mathcal{O}'' = \mathcal{O} \setminus \mathcal{O}'$ and then select the Luce utilities defined by the unique value $\bar{\alpha} \in [0, 1]$ that solves $\min_{(a,A)\in\mathcal{O}'} \frac{\rho(a,A)}{\delta_{\alpha\mathbf{1}_x+(1-\alpha)v_m}(a,A)} = \min_{(a,A)\in\mathcal{O}''} \frac{\rho(a,A)}{\delta_{\alpha\mathbf{1}_x+(1-\alpha)v_m}(a,A)}$. Notice that, given the structure of the Luce model, the left-hand ratio increases with α, continuously and strictly, approaching infinity. Similarly, the right-hand ratio weakly decreases with α continuously. Notice also that, for $\alpha = 0$, the left-hand ratio is strictly below the right-hand ratio. This is because there exists at least one observation on the left-hand side that belongs to $\mathcal{O}_{\delta_{v_m}}$. Thus, $\bar{\alpha}$ must exist and Proposition 1 guarantees that this provides the maximal separation in the segment. Then, consider the vector of Luce utilities $v = \bar{\alpha}\mathbf{1}_x + (1 - \bar{\alpha})v_m$. If alternative x is present in all the menus in $\mathcal{O}_{\delta_{v_m}}$, then $\min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_v(a,A)} > \min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_{v_m}(a,A)} = \min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_L(a,A)}$, thus contradicting the maximality of δ_L. If x is not present in some menu of $\mathcal{O}_{\delta_{v_m}}$, it must be the case that $\mathcal{O}_{\delta_v} \subsetneq \mathcal{O}_{\delta_{v_m}}$ and $\min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_v(a,A)} = \min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_L(a,A)}$. Given the finiteness of the data, we can repeat the same exercise for δ_v and, eventually, contradict the optimality of δ_L. This concludes the proof. ■

Proof of Proposition 5: We start by proving that λ_T is lower or equal than the maximal fraction of the data that can be explained by SCRUM. The construction guarantees that $1 \geq \lambda_T \geq \lambda_{T-1} \geq \cdots \geq \lambda_1 \geq 0$. Whenever $\lambda_T = 0$, the result is immediate. Assume that $\lambda_T \in (0, 1)$. We prove that there exists a separation of ρ of the form $\langle \lambda_T, \delta_{\mu_T}, \frac{\rho-\lambda_T\delta_{\mu_T}}{1-\lambda_T} \rangle$. Since the construction guarantees that $\delta_{\mu_T} \in \mathtt{SC}$, we only need to prove that $\epsilon = \frac{\rho-\lambda_T\delta_{\mu_T}}{1-\lambda_T} \in \mathtt{SCF}$. To see this, consider $(a, A) \in \mathcal{O}$ and denote by $\underline{i}$ and $\bar{i}$ the integers of the first and last preferences in $\mathcal{P}'$, such that a is the maximal

[31] This shows, in addition, that there is always a strictly positive instance of Luce that is maximal.

element in A. The construction guarantees that $\rho(a,A) \geq \lambda_{\bar{i}} - \lambda_{\underline{i}-1} = \lambda_T \frac{\lambda_{\bar{i}}-\lambda_{\underline{i}-1}}{\lambda_T}$. Now, the recursive equations can be written as $\mu_T(P_i) = \frac{\lambda_{P_i}-\lambda_{P_{i-1}}}{\lambda_T}$ for every $i \in \{1,2,\ldots,T\}$, with $\lambda_0 = 0$ and hence, $\rho(a,A) \geq \lambda_T \sum_{i=\underline{i}}^{\bar{i}} \mu_T(P_i) = \lambda_T \delta_{\mu_T}(a,A)$. This implies that $\epsilon(a,A) \geq 0$. Notice also that $\sum_{a\in A} \epsilon(a,A) = \sum_{a\in A} \frac{\rho(a,A)-\lambda_T\delta_{\mu_T}(a,A)}{1-\lambda_T} = \frac{1-\lambda_T}{1-\lambda_T} = 1$, thus proving that $\epsilon \in \mathtt{SCF}$. This shows the claim and, hence, the desired inequality. Finally, suppose that $\lambda_T = 1$. In this case, by noticing again that the construction guarantees that $\rho = \delta_{\mu_T} \in \mathtt{SC}$, the desired inequality follows.

We now show that λ_T is greater or equal than the maximal fraction of the data that can be explained by SCRUM. To show this let $\langle \lambda, \delta_\mu, \epsilon \rangle$ be a separation for SCRUM. We need to show that $\lambda_T \geq \lambda$. We proceed recursively to show that $\lambda_i \geq \sum_{j=1}^{i} \lambda\mu(P_j)$ holds, and hence, $\lambda_T \geq \sum_{j=1}^{T} \lambda\mu(P_j) = \lambda$, as desired. Let $i = 1$ and A' be a menu solving $\min\limits_{A\in\mathcal{D}} \rho(m_{P_1}(A), A)$. Hence, $\lambda_1 - \lambda\mu(P_1) = \rho(m_{P_1}(A'), A') - \lambda\mu(P_1) \geq \rho(m_{P_1}(A'), A') - \lambda \sum_{j:m_{P_j}(A')=m_{P_1}(A')} \mu(P_j)$. By the definition of SCRUMs, the last expression can be written as $\rho(m_{P_1}(A'), A') - \lambda\delta_\mu(m_{P_1}(A'), A')$, or equivalently as $(1-\lambda)\epsilon(m_{P_1}(A'), A')$. Since $\epsilon \in \mathtt{SCF}$, the latter expression must be positive, thus proving the desired result. Suppose that the inequality is true for every P_j with $j < i$. We now prove this for P_i. Let $\bar{A}$ be a menu solving $\min_{A\in\mathcal{D}}[\rho(m_{P_i}(A), A) + \max_{j:j\leq i, m_{P_j}(A)\neq m_{P_i}(A)} \lambda_j]$. Then, we have $\rho(m_{P_i}(\bar{A}), \bar{A}) = \lambda\delta_\mu(m_{P_i}(\bar{A}), \bar{A}) + (1-\lambda)\epsilon(m_{P_i}(\bar{A}), \bar{A}) \geq \lambda\delta_\mu(m_{P_i}(\bar{A}), \bar{A}) = \lambda \sum_{P:m_P(\bar{A})=m_{P_i}(\bar{A})} \mu(P)$. If it is the case that $\{P : m_P(\bar{A}) = m_{P_i}(\bar{A})\} \supseteq \{P_1, P_2, \ldots, P_i\}$, then clearly $\lambda_i = \rho(m_{P_i}(\bar{A}), \bar{A}) \geq \lambda \sum_{P:m_P(\bar{A})=m_{P_i}(\bar{A})} \mu(P) = \sum_{j=1}^{i} \lambda\mu(P_j)$ and we have concluded the induction argument. Otherwise, the single-crossing condition guarantees that there exists $\bar{j} \in \{1,\ldots,i-1\}$ such that $\{P : m_P(\bar{A}) = m_{P_i}(\bar{A})\} \supseteq \{P_{\bar{j}+1}, P_{\bar{j}+2}, \ldots, P_i\}$ and $\rho(m_{P_i}(\bar{A}), \bar{A}) \geq \sum_{j=\bar{j}+1}^{i} \lambda\mu(P_j)$. In this case, the induction hypothesis also guarantees that $\lambda_{\bar{j}} \geq \sum_{j=1}^{\bar{j}} \lambda\mu(P_j)$. By combining these two inequalities, we are able to conclude that $\lambda_i \geq \sum_{j=1}^{i} \lambda\mu(P_j)$ and the induction step is complete. This implies, in particular, that $\lambda \leq \lambda_T$.

By combining the above two claims, we have shown that $\langle \lambda_T, \delta_{\mu_T}, \frac{\rho-\lambda_T\delta_{\mu_T}}{1-\lambda_T} \rangle$ is a maximal separation for SCRUM, which concludes the proof. ■

Appendix B. Examples

We first propose a simple example of a stochastic choice function, and derive the maximal separations for all the models studied in Section 4, we then use another example to show that the maximal separations for the deterministic model and the tremble model do not necessarily identify the same preference relations, and, finally, we propose a particular data-generating process and use the tremble model to illustrate our conjecture on the differences between maximal separation and maximum likelihood in over-estimating choice probabilities.

Table 5 reports a stochastic choice function ρ defined on every non-singleton subset of $X = \{x, y, z\}$, i.e., $\mathcal{D} = \{\{x, y, z\}, \{x, y\}, \{x, z\}, \{y, z\}\}$. Note that this stochastic choice function involves behavior rather unstructured, in the sense that it does not satisfy weak stochastic transitivity.

TABLE 5. A stochastic choice function ρ

	x	y	z
$\{x, y, z\}$	0.15	0.6	0.25
$\{x, y\}$	0.25	0.75	
$\{x, z\}$	0.7		0.3
$\{y, z\}$		0.4	0.6

We start with the deterministic model. We can first calculate the maximal fraction for every set for which $\mathcal{D}|_S = \{S\}$, i.e., the binary sets:

$$\lambda_{\{x,y\}} = \max\{\rho(x, \{x, y\}), \rho(y, \{x, y\})\} = 0.75,$$

$$\lambda_{\{x,z\}} = \max\{\rho(x, \{x, z\}), \rho(z, \{x, z\})\} = 0.7, \text{ and}$$

$$\lambda_{\{y,z\}} = \max\{\rho(y, \{y, z\}), \rho(z, \{y, z\})\} = 0.6.$$

We can then proceed to assign a value to menu X, for which we first analyze the alternatives in X one-by-one. For alternative x, we compute the minimum of $\Big\{\{\rho(x, \{x, y\}), \rho(x, \{x, z\}), \rho(x, X)\}, \lambda_{\{y,z\}}\Big\} = \rho(x, \{x, y, z\}) = 0.15$. For alternative y, the minimum of $\Big\{\{\rho(y, \{x, y\}), \rho(y, \{y, z\}), \rho(y, X)\}, \lambda_{\{x,z\}}\Big\} = \rho(y, \{y, z\}) = 0.4$ is the relevant value. Finally, for alternative z we are required to compute the minimum of $\Big\{\{\rho(z, \{x, z\}), \rho(z, \{y, z\}), \rho(z, X)\}, \lambda_{\{x,y\}}\Big\} = \rho(z, \{x, y, z\}) = 0.25$. Thus, we get

$$\lambda_X = \max\{0.15, 0.4, 0.25\} = 0.4.$$

Notice that the last value is obtained with alternative y. In subset $X \setminus \{y\}$, the alternative determining the value $\lambda_{\{x,z\}}$ is x. Hence, the second part of Proposition 2 guarantees that δ_P with $yPxPz$ conforms to a maximal separation of ρ. From $\lambda_X = 0.4$, one can immediately obtain the corresponding residual behavior as $\epsilon = \frac{\rho - 0.4\delta_P}{0.6}$, i.e.,

$\epsilon(x,X) = \frac{1}{4}, \epsilon(y,X) = \frac{1}{3}, \epsilon(x,\{x,y\}) = \frac{5}{12}, \epsilon(x,\{x,z\}) = \frac{1}{2}$, and $\epsilon(y,\{y,z\}) = 0$. To close the discussion of this example, notice from the residual behavior that the frontier of `SCF` is reached at $(y,\{y,z\})$. This is precisely the observation where the identified instance δ_P fails most seriously. In turn it determines the maximal fraction of data explained by `DET`, i.e., $\frac{\rho(y,\{y,z\})}{\delta_P(y,\{y,z\})} = \frac{0.4}{1} = 0.4$.

We now illustrate the treatment of the tremble model. Replicating the steps taken in the analysis of `DET`, we conclude that $yPxPz$ is the optimal preference relation for every given value of γ.[32] In order to find the optimal value of γ, note that there are only two possible critical observations, depending on the value of γ. When γ is low, we know from the study of the deterministic case that the critical observation is $(y,\{y,z\})$, with a ratio ρ to δ equal to $\frac{0.4}{1-\gamma+\frac{\gamma}{2}}$. When γ is high the critical observation is $(x,\{x,y,z\})$, with a ratio ρ to δ equal to $\frac{0.15}{\frac{\gamma}{3}}$. By noticing that the first ratio is increasing and starts at a value below the second ratio, which is decreasing, it follows that the maximal fraction of data explained by the optimal tremble can be found by equating these two ratios, which yields $\gamma^* = 0.72$. Hence, the maximal fraction of data explained is 0.625, obtained with the trembling stochastic choice function $\delta_{[P,0.72]}$ and residual behavior $\epsilon = \frac{\rho - 0.625\delta_{[P,0.72]}}{0.375}$, i.e., $\epsilon(x,X) = 0, \epsilon(y,X) = \frac{11}{15}, \epsilon(x,\{x,y\}) = \frac{1}{15}, \epsilon(x,\{x,z\}) = \frac{4}{5}$, and $\epsilon(y,\{y,z\}) = 0$. To conclude, notice that the ability of the tremble model to explain the data reaches its limits as a result of the tension created by the two critical observations, $(y,\{y,z\})$ and $(x,\{x,y,z\})$.

As for the model of Luce, consider the Luce utilities $u = (\frac{1}{3},\frac{1}{3},\frac{1}{3})$. The value $\min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_u(a,A)} = 0.45$ is obtained only for observation $(x,\{x,y,z\})$. Since $\{x,y,z\} \setminus \{x\} = \{y,z\}$ is non-empty, we can select one of the alternatives in $\{y,z\}$, say y, and move the utility values within the segment $\alpha(0,1,0) + (1-\alpha)u = (\frac{1-\alpha}{3}, \frac{1+2\alpha}{3}, \frac{1-\alpha}{3})$. In order to select the appropriate value of α, we consider the observations (a,A) with $a \neq y \in A$ and the observations (y,A). Among the former, the minimal ratio of the data to the Luce probabilities is obtained for $(x,\{x,y,z\})$, with value $\frac{0.45}{1-\alpha}$. In the latter, the minimal ratio is reached at $(y,\{y,z\})$, with value $\frac{0.4(2+\alpha)}{1+2\alpha}$. Equation $\frac{0.45}{1-\alpha} = \frac{0.4(2+\alpha)}{1+2\alpha}$ yields $\bar{\alpha} = \frac{1}{4}$, which leads to $v = (\frac{1}{4},\frac{1}{2},\frac{1}{4})$. The value $\min\limits_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta_v(a,A)} = 0.6$ is obtained for pairs $\{(x,\{x,y,z\}),(z,\{x,z\}),(y,\{y,z\})\}$. Notice that the critical observations of δ_v have the cyclical structure described by Proposition 4, i.e., $\{x,y,z\} \cup \{x,z\} \cup \{y,z\} = \{x\} \cup \{z\} \cup \{y\}$ and, as a result, the fraction

[32] In an example below we show that this is not necessarily the case in general.

of data explained by the model of Luce cannot be increased further. We have then found the maximal separation $\langle \lambda^*, \delta^*_L, \epsilon^* \rangle$, with $\delta^*_L = \delta_v$, $\lambda^* = \min_{(a,A)\in\mathcal{O}} \frac{\rho(a,A)}{\delta^*_L(a,A)} = 0.6$, and $\epsilon^* = \frac{\rho - \lambda^*\delta^*_L}{1-\lambda^*}$, that is $\epsilon^*(x,X) = 0, \epsilon^*(y,X) = \frac{3}{4}, \epsilon^*(x,\{x,y\}) = \frac{1}{8}, \epsilon^*(x,\{x,z\}) = 1$, and $\epsilon^*(y,\{y,z\}) = 0$.

We now illustrate how Proposition 5 works in the example of Table 5, with the set of single-crossing preferences zP_1yP_1x, yP_2zP_2x, yP_3xP_3z and xP_4yP_4z. We start with P_1. The maximal fraction of data explained by P_1 is $\lambda_1 = \min_{A\subseteq X} \rho(m_{P_1}(A), A) = \min\{\rho(z,X), \rho(y,\{x,y\}), \rho(z,\{x,z\}), \rho(z,\{y,z\})\} = \min\{0.25, 0.75, 0.3, 0.6\} = 0.25$, where trivially $\mu_1(P_1) = 1$. We then consider preference P_2, where we have that $\lambda_2 = \min\{\rho(y,X)+\lambda_1, \rho(y,\{x,y\}), \rho(z,\{x,z\}), \rho(y,\{y,z\})+\lambda_1\} = \min\{0.6+0.25, 0.75, 0.3, 0.4+0.25\} = 0.3$ with $\mu_2(P_1) = \frac{\lambda_1}{\lambda_2} = \frac{5}{6}$ and $\mu_2(P_2) = \frac{1}{6}$. For preference P_3, we have that $\lambda_3 = \min\{\rho(y,X) + \lambda_1, \rho(y,\{x,y\}), \rho(x,\{x,z\}) + \lambda_2, \rho(y,\{y,z\}) + \lambda_1\} = \min\{0.6 + 0.25, 0.75, 0.7 + 0.3, 0.4 + 0.25\} = 0.65$, with $\mu_3(P_1) = \frac{\lambda_2}{\lambda_3}\mu_2(P_1) = \frac{5}{13}$, $\mu_3(P_2) = \frac{\lambda_2}{\lambda_3}\mu_2(P_2) = \frac{1}{13}$ and $\mu_2(P_3) = \frac{7}{13}$. Finally, we have that $\lambda_4 = \min\{\rho(x,X)+\lambda_3, \rho(x,\{x,y\})+\lambda_3, \rho(x,\{x,z\}) + \lambda_2, \rho(y,\{y,z\}) + \lambda_1\} = \min\{0.15 + 0.65, 0.25 + 0.65, 0.7 + 0.3, 0.4 + 0.25\} = 0.65$ and hence $\mu_4 = \mu_3$. Thus, we conclude that the maximal fraction of the data that can be explained by SCRUM is 0.65, with maximal SCRUM δ_{μ_4} and residual behavior $\epsilon(x,X) = \frac{3}{7}, \epsilon(y,X) = \frac{4}{7}, \epsilon(x,\{x,y\}) = \frac{5}{7}, \epsilon(x,\{x,z\}) = 1$, and $\epsilon(y,\{y,z\}) = 0$, with critical observations $(x,X), (z,\{x,z\})$ and $(y,\{y,z\})$. Note that the example illustrates that using a superset of preferences does not necessarily lead to a strict improvement in the goodness of fit.

We now provide an example to illustrate that the deterministic model and the tremble model do not necessarily identify the same preference relations.

TABLE 6. $P_{\texttt{DET}}$ and $P_{\texttt{Tremble}}$

	x	y	z
$\{x,y,z\}$	0.39	0.55	0.06
$\{x,y\}$	0.6	0.4	
$\{x,z\}$	0.95		0.05
$\{y,z\}$		0.95	0.05

Repeating the above logic, it is easy to see that the optimal preference relation for the deterministic model is $yP_{\texttt{DET}}xP_{\texttt{DET}}z$, while the one for the tremble model is

$xP_{\texttt{Tremble}}yP_{\texttt{Tremble}}z$ with a tremble of 30/137.

Finally, in Section 6 we conjectured that, given the nature of maximal separations, we can expect these to perform better in the over-estimation of low observed choice frequencies, while maximum likelihood may perform better on average. We then saw this conjecture reflected in the data. Here, we use a simple example involving a particular data-generating process and the tremble model to illustrate the content and intuition of our conjecture more formally, and leave its general development for future research.

Suppose that the individual has a preference P, and consider a set of binary menus $m_i = \{x_i, y_i\}$ where $x_i P y_i$. The data-generating process involves the maximization of P except for a small menu-dependent error ϵ_i of choosing alternative y_i.

The log-likelihood of the data with respect to the tremble model is $\sum_i (1-\epsilon_i)\log(1-\frac{\gamma}{2}) + \sum_i \epsilon_i \log\frac{\gamma}{2}$, and its maximization leads to $\frac{1-\frac{\gamma}{2}}{\frac{\gamma}{2}} = \frac{\sum_i(1-\epsilon_i)}{\sum_i \epsilon_i} = \frac{1-\bar{\epsilon}}{\bar{\epsilon}}$, where $\bar{\epsilon}$ is the average observed mistake. That is, maximum log-likelihood averages out the mistakes observed across different menus, suggesting the tremble $\gamma_{ML} = 2\bar{\epsilon}$ and consequently, a choice probability of the inferior alternative equal to $\bar{\epsilon}$. Now consider the maximal separation of data. For a given tremble γ, the only potentially critical observations are those in which the mistake is greatest or least, that is either $\max_i \epsilon_i$ or $\min_i \epsilon_i$. In the first case, the superior alternative has been chosen with probability $1-\max_i \epsilon_i$ and the estimated tremble model will, by maximal separation, over-estimate this probability. Obviously, the superior alternative in any other menu will be less over-estimated and cannot be critical. In the second case, likewise, the inferior alternative has been chosen with probability $\min_i \epsilon_i$ and its maximal separation specification will over-estimate it to a greater degree than any other inferior alternative in the remaining menus. In order to find the maximal separation, we need to equalize these two observations, that is $\frac{1-\frac{\gamma}{2}}{\frac{\gamma}{2}} = \frac{1-\max_i \epsilon_i}{\min_i \epsilon_i}$.

For most data-generating processes, e.g. any symmetric distribution of mistake probabilities, the following condition holds: $\frac{1-\max_i \epsilon_i}{\min_i \epsilon_i} > \frac{1-\bar{\epsilon}}{\bar{\epsilon}}$. Whenever this happens, the estimation of maximal separation will provide an estimated tremble $\gamma_{MS} < \gamma_{ML}$, and will therefore better accommodate the most extreme observations. In terms of out-of-sample predictions, the same logic applies. Consider a new menu with a mistake probability equal to ϵ. If $\gamma_{MS} < \gamma_{ML}$, there are three cases of interest: (i) $\epsilon < \frac{\gamma_{MS}}{2}$, (ii) $\frac{\gamma_{MS}}{2} < \epsilon < \frac{\gamma_{ML}}{2}$ and (iii) $\epsilon > \frac{\gamma_{ML}}{2}$. In the first and third cases, the estimations fail in a similar fashion. That is, they both over-estimate the choice probability of the inferior alternative (in case (i)) or the choice probability of the superior alternative (in case (iii)). Clearly, maximal separation does a better job in the first case, where the data are scarce (the relevant alternative is inferior), while maximum likelihood does a better job in the latter cases, and also on average.

Appendix C. Empirical application: further considerations

In this section we report on the application of the maximal separation approach to random expected utility, and the out-of-sample results involving the 3- and 5-option menus.

The random expected utility (REU) of Gul and Pesendorfer (2006) is a key reference in the stochastic treatment of risk preferences. Here we discuss how to use Proposition 1 in order to obtain its maximal separation using our experimental dataset.

For the sake of consistency throughout the analysis in this paper, we impose the requirement that all the relevant expected utility preferences be linear orders. Secondly, given that we are working with binary menus, we can understand each particular instance of the REU as a probability distribution over the set of all preferences satisfying the standard properties of independence and first order stochastic dominance. Notice that, in our setting: (i) independence requires that l_iPl_j if and only if $l_{i+4}Pl_{j+4}$ for $i, j \in \{2, 3, 4, 5\}$, and (ii) first order stochastic dominance requires that l_5Pl_9. Thus, an instance of REU is merely a probability distribution over the set of linear orders satisfying these conditions.

We can then use Proposition 1 to explain how the maximal separation of the data for REU can be obtained. Consider, first, a case of independence, say, l_4Pl_5 if and only if l_8Pl_9. This leads to the linearity property of REU where $\delta(l_4, \{l_4, l_5\}) = \delta(l_8, \{l_8, l_9\})$. However, since $\rho(l_4, \{l_4, l_5\}) = 0.49$ and $\rho(l_8, \{l_8, l_9\}) = 0.64$, this is not observed in the data. Finding the instance of REU that is closest to these data implies finding a value $0.49 < x < 0.64$ such that $\frac{0.49}{x} = \frac{1-0.64}{1-x}$, which leads to $x = 0.576$. Then, setting $\delta(l_4, \{l_4, l_5\}) = \delta(l_8, \{l_8, l_9\}) = 0.576$ gives a ρ/δ ratio of 0.85, which means that the maximal separation can explain no more than 85% of the data. One can check that the other violations of independence are less severe, and hence the bound imposed by independence is 0.85. Now consider the implications of stochastic dominance. This requires that, for every instance of REU, it must be that $\delta(l_5, \{l_5, l_9\}) = 1$. However, we observe that $\rho(l_5, \{l_5, l_9\}) = 0.83$, thus yielding a ρ/δ ratio of 0.83. It turns out, therefore, that this ratio determines the goodness of fit measure of REU in our dataset. Clearly, the fraction of the data explained increases with respect to that explained when using CRRA expected utilities, since the latter involve only a subset of expected utilities.

Since REU has no uniqueness in a finite domain, one can find multiple instances of the model for which 83% of the data are explained. We now construct one such instance. Start with the set of lotteries $\{l_3, l_5, l_7, l_8, l_9\}$ and select the following four linear orders over it: (i) $l_8 P_1 l_3 P_1 l_7 P_1 l_5 P_1 l_9$, (ii) $l_5 P_2 l_8 P_2 l_9 P_2 l_7 P_2 l_3$, (iii) $l_3 P_3 l_5 P_3 l_7 P_3 l_9 P_3 l_8$ and (iv) $l_5 P_4 l_9 P_4 l_7 P_4 l_3 P_4 l_8$. Notice that they all place l_5 at the top of l_9, and hence any RUM using them will satisfy stochastic dominance. Notice, also, that the independence relationship involving the lotteries l_3, l_5, l_7 and l_9 is always respected. Assign to the four linear orders the probabilities pq, $p(1-q)$, $(1-p)q$ and $(1-p)(1-q)$, respectively. For each of these four preferences, consider two linear orders, the one that places l_1 at the top and the one that places l_1 at the bottom, and assign to each of them the conditional probabilities r and $1-r$, respectively. For each of these eight preferences, we now place l_4 either at the top or at the bottom, while respecting independence. That is, for the preferences constructed on the basis of P_1 and P_2, independence requires that l_4 must be above l_3 and l_5 and hence, we place it at the top. Similarly, for the preferences constructed on the basis of P_3 and P_4, we place l_4 at the bottom. Finally, for each of these 8 preferences, create 4 preferences that place l_2 and l_6 at the top, in both orders, and l_2 and l_6 at the bottom, in both orders. Notice that this respects independence for any pair associated with l_2 and l_6. Assign to them the conditional probabilities $ts, t(1-s), (1-t)s, (1-t)(1-s)$. A direct application of Proposition 1 allows us to find values of these parameters p, q, r, s, t which yield the maximal REU separation value 0.83, using the 32 expected utility linear orders described. For instance, $p = 0.565$, $q = 0.473$, $r = 0.705$, $s = 0.2$, $t = 0.5$. The nature of the residual stochastic choice function follows directly from this construction and Proposition 1.

Our experimental dataset involved the choices from 2-, 3- and 5-option menus. In the main text, we have focused on the binary menus, since we have a relatively large number of data points for each binary menu; that is, about 87 choices for each of the 36. In contrast, each participant was confronted with 36 out of the possible 84 menus of 3 lotteries and 36 out of the possible 126 menus of 5 lotteries, all randomly selected without replacement. This gives an average of 37 (25) observations in the

3- and 5-option menus, respectively, which means markedly fewer data points per menu. In this appendix, we use the data pertaining to the 3- and 5-option menus to perform another out-of-sample exercise. We take the estimated models for maximal separation and maximum likelihood using the binary data reported in Table 3 and follow the methodology adopted in Section 6, evaluating the predictions of these models and techniques using the observations from the non-binary menus. As in the main text, we focus on those observations in which both maximal separation and maximum likelihood over-estimate the observed choice frequencies, and evaluate the probability of the maximal separation prediction being closer to the data than the maximum likelihood prediction. Reporting the results for each observation, as in Table 4, is unfeasible, since there are now far too many observations.[33]

Table 7. Summary Statistics for the Forecasting Results for the 3- and 5-Option Menus

	Tremble	Luce	SCRUM-CRRA
First quintile	100%	88%	59%
Second quintile	97%	78%	68%
Third quintile	46%	67%	80%
Fourth quintile	0%	34%	82%
Fifth quintile	0%	14%	49%
Average	49%	56%	68%

Table 7 reports some summary statistics. Focusing on those observations for which both maximal separation and maximum likelihood over-estimate the observed choice frequencies, and ordering the observations from lower to higher observed choice frequencies, the table reports, by quintiles and on average, the frequency with which maximal separation is closer than maximum likelihood to the data. We see that, in general, maximal separation is better than maximum likelihood at over-estimating low observed choice frequencies. This is particularly true in the case of Tremble and Luce, but also in SCRUM-CRRA when comparing the first quintile against the fifth one. We also see that, on average, maximal separation does a remarkably good job: it is closer to the observed choices than maximum likelihood in 49%, 56%, and 68% of all the over-estimated cases. This may have to do with the fact that, in larger menus, the

[33]We provide all the results in the online appendix.

observed choice probabilities are generally smaller, and thus better accommodated by maximal separation. However, given the small number of observations for the 3- and 5-option menus, these conclusions should be taken with a grain of salt.

Appendix D. Inconsistency indices

Starting with Afriat (1973), there is a literature on measuring deviations of actual behavior with respect to the standard, deterministic, rational choice model. Formally, an inconsistency index can be defined as a mapping $I : \texttt{SCF} \to \mathbb{R}$ describing the inconsistency of a dataset $\rho \in \texttt{SCF}$ with the standard deterministic model, that is when the reference model is set as $\Delta = \texttt{DET}$. Most of the existing inconsistency indices are obtained throughout the minimization of a loss function.[34] We can then analyze the inconsistency index emerging from the maximal separation technique. Using the loss function discussed in Section 6, and the insights obtained in Section 4.1, we have

$$I_{MS} = 1 - \lambda^* = \min_P \max_A \sum_{\substack{a \in A: \\ \delta_P(a,A)=0}} \rho(a, A).$$

It is important to note that the nature of this index is unique in this literature. To illustrate this more clearly, we now compare it with the well-known inconsistency index of Houtman and Maks (1985), which represents the closest index to I_{MS}. The Houtman and Maks index measures inconsistency by the minimal amount of data that needs to be removed in order to make the remainder of the data rationalizable by the standard choice model. The key difference is that the Houtman-Maks index enables different proportions of data to be removed from different menus of alternatives. Hence, using our notation, we can write the Houtman-Maks index as

$$I_{HM} = \min_P \sum_A \sum_{\substack{a \in A: \\ \delta_P(a,A)=0}} \rho(a, A).$$

These formulations provide a transparent comparison between the two approaches. Both methods remove data minimally until the surviving data is rationalizable. In the case of a maximal separation, since data must be removed at the same rate across all menus, the index focuses on the most problematic menu. In the case of Houtman and Maks, different proportions of data can be removed from different menus, therefore an aggregation across menus takes place.

[34] See Apesteguia and Ballester (2015) for a characterization of this class and for a review of the literature.

Table 8. I_{MS} versus I_{HM}

	x	y	z
$\{x, y, z\}$	0.25	0.3	0.4
$\{x, y\}$	0.8	0.2	
$\{x, z\}$	0.4		0.6
$\{y, z\}$		0.7	0.3

Table 8 reports an example of a choice function ρ with three alternatives and with data on all the relevant menus of alternatives. Taken from the perspective of I_{MS}, the data show the optimal preference to be $zPxPy$, while from the perspective of I_{HM} it is $xP'yP'z$.

References

[1] Afriat, S.N. (1973). "On a System of Inequalities in Demand Analysis: An Extension of the Classical Method," *International Economic Reviw* 14:460–72.

[2] Agranov, M. and P. Ortoleva (2017). "Stochastic Choice and Preferences for Randomization," *Journal of Political Economy*, 125.1:40–68.

[3] Ahn, D.S. and T. Sarver (2013). "Preference for Flexibility and Random Choice," *Econometrica*, 81(1):341–361.

[4] Apesteguia, J. and M.A. Ballester (2010). "The Computational Complexity of Rationalizing Behavior," *Journal of Mathematical Economics*, 46.3:356-363.

[5] Apesteguia, J. and M.A. Ballester (2015). "A Measure of Rationality and Welfare," *Journal of Political Economy*, 123.6:1278–1310.

[6] Apesteguia, J. and M.A. Ballester (2018). "Monotone Stochastic Choice Models: The Case of Risk and Time Preferences," *Journal of Political Economy*, 126.1:74–106.

[7] Apesteguia, J., M.A. Ballester and J. Lu (2017). "Single-Crossing Random Utility Models," *Econometrica*, 85.2:661–674.

[8] Barseghyan, L., F. Molinari, T. O'Donoghue and J.C. Teitelbaum (2016), "Estimating Risk Preferences in the Field," *Journal of Economic Literature*, forthcoming

[9] Block, H.D. and J. Marschak (1960). "Random Orderings and Stochastic Theories of Response," in I. Olkin et al., eds., Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, 97–132. Standford, Calif.: Standford University Press.

[10] Böckenholt (2006). "Thurstonian-Based Analyses: Past, Present, and Future Utilities," *Psychometrica*, 71.4:615–629.

[11] Brady, R.L. and J. Rehbeck (2016). "Menu-Dependent Stochastic Feasibility," *Econometrica*, 84.3:1203–1223.

[12] Caplin, A., M. Dean and D. Martin (2011). “Search and Satisficing,” *American Economic Review*, 101: 2899–2922.

[13] Caplin, A. and M. Dean (2016). “Revealed Preference, Rational Inattention, and Costly Information Acquisition,” *American Economic Review*, forthcoming.

[14] Crawford, I., K. Pendakur (2013). “How many types are there?” *Economic Journal*, 123.567:77–95.

[15] Dean, M. and D. Martin (2016). “Measuring Rationality with the Minimum Cost of Revealed Preference Violations,” *Review of Economics and Statistics*, 98.3: 524–534.

[16] Dayton, C.M. (2003). “Applications and Computational Strategies for the Two-Point Mixture Index,” *British Journal of Mathematical and Statistical Psychology*, 56:1–13.

[17] Dickhaut, J., A. Rustichini and V. Smith (2009). “A Neuroeconomic Theory of the Decision Process,” *Proceedings of the National Academy of Sciences*, 106.52:22146–22150.

[18] Echenique, F., S. Lee, and M. Shum (2011). “The Money Pump as a Measure of Revealed Preference Violations,” *Journal of Political Economy*, 119.6:1201–1223.

[19] Echenique, F. and K. Saito (2018). “General Luce Model,” *Economic Theory*, forthcoming.

[20] Famulari, M. (1995). “A Household-Based, Nonparametric Test of Demand Theory,” *Review of Economics and Statistics* 77:372–383.

[21] Fudenberg, D., R. Iijima and T. Strzalecki (2015). “Stochastic Choice and Revealed Perturbed Utility,” *Econometrica*, 83.6:2371–2409.

[22] Fudenberg, D., J. Kleinberg, A. Liang and S. Mullainathan (2019). “Measuring the Completeness of Theories,” mimeo.

[23] Fudenberg, D. and T. Strzalecki (2015). “Dynamic Logit with Choice Aversion,” *Econometrica*, 83.2:651–691.

[24] Gul, F. and W. Pesendorfer (2006). “Random Expected Utility,” *Econometrica*, 74.1:121–146.

[25] Gul, F., P. Natenzon, and W. Pesendorfer (2014). “Random Choice as Behavioral Optimization,” *Econometrica*, 82.5:1873–1912.

[26] Halevy, Y., D. Persitz and L. Zrill (2018). “Parametric Recoverability of Preferences,” *Journal of Political Economy*, 126.4:1558–1593.

[27] Harless, D.H. and C.F. Camerer (1994). “The Predictive Utility of Generalized Expected Utility Theories,” *Econometrica*, 62.6:1251–1289.

[28] Horan, S. (2019). “Threshold Luce Rules,” mimeo.

[29] Houtman, M. and J.A.H. Maks (1985). “Determining All Maximal Data Subsets Consistent with Revealed Preference,” *Kwantitatieve Methoden*, 19:89–104.

[30] Kalai, G., A. Rubinstein and R. Spiegler (2002). “Rationalizing Choice Functions by Multiple Rationales,” *Econometrica*, 70:2481–2488.

[31] Liang, A. (2019). “Inference of Preference Heterogeneity from Choice Data,” *Journal of Economic Theory*, 179:275–311

[32] Liu, J. and B.G. Lindsay (2009). “Building and Using Semiparametric Tolerance Regions for Parametric Multinomial Models,” *Annals of Statistics*, 37.6A:3644–3659.

[33] Lu, J. and K. Saito (2017). “Random Intertemporal Choice,” mimeo.

[34] Luce, R.D. (1959). "Individual Choice Behavior: A Theoretical Analysis," New York, New York: Wiley.

[35] Manzini, P. and M. Mariotti (2014). "Stochastic Choice and Consideration Sets," *Econometrica*, 82(3):1153–1176.

[36] Matejka, P. and A. McKay (2015). "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, 105(1):272–98.

[37] Mosteller, F. and P. Nogee (1951). "An Experimental Measurement of Utility," *Journal of Political Economy*, 59:371–404.

[38] Natenzon, P. (2017). "Random Choice and Learning," *Journal of Political Economy*, forthcoming.

[39] Rudas T., C.C. Clogg, B.G. Lindsay (1994). "A New Index of Fit Based on Mixture Methods for the Analysis of Contingency Tables," *Journal of the Royal Statistical Society. Series B*, 56.4:623–639.

[40] Swofford, J.L. and G.A. Whitney (1987). "Nonparametric Test of Utility Maximization and Weak Separability for Consumption, Leisure and Money," *Review of Economic and Statistics*, 69:458–464.

[41] Webb, R. (2019). "The (Neural) Dynamics of Stochastic Choice," *Management Science*, 65.1:230–255.