Which Performs Best? Comparing Discrete Choice Models*

Geoffroy de Clippel and Kareen Rozen

October 2022

Abstract

Fundamental questions regarding the individual-level performance of discrete-choice models remain unsettled. We propose a method of fit-optimized Bayesian model comparison that accounts for, and disentangles, issues of goodness of fit, sampling variation and model size. We apply the approach to a well-known replication of Tversky (1969)'s experiment on intransitivity, studying repeated choices from binary menus of lotteries. We find that parsimonious specifications of classic models such as rationality and the Luce rule, which both have a single underlying preference, perform surprisingly well at the subject level. We further evaluate the tradeoffs when hoping to explain many subjects with a common model.

1 Introduction

Choice theory has developed tremendously since its earliest contributions. Many modifications have been proposed of the simple, benchmark model of rationality, which requires the decision maker (DM) to maximize a preference ordering over each choice problem. Some, like expected utility, further restrict the class of admissible preferences. Others accommodate more choice patterns, as part of an effort, surveyed in de Clippel and Rozen (2022), to incorporate bounded rationality and lessons from behavioral economics.

The development of new models of discrete choice has generally occurred at a faster pace than the testing of existing ones, and many fundamental questions remain unsettled regarding their performance, especially at the individual level.[1] How does the basic, rational

*Part of this research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University. We thank seminar participants at Brown University, Johns Hopkins, Princeton University, the University of Essex and the University of Queensland for helpful comments.

[1]Oftentimes, datasets with multiple choices from menus are derived by combining the choices of different individuals; such aggregate data can make it difficult to judge the performance of theories from the perspective of individual decision making.

choice model fare? Is its explanatory power, if any, diminished by restricting to common parametric specifications of utility? Should we instead be more permissive? Do deterministic theories of bounded rationality dominate rationality?[2] Do stochastic choice theories perform best?[3] And ultimately, how do we select among competing theories? The present paper strives to shed some light on the above questions.

Our analysis takes into account three important factors. As will be seen in our discussion of the related literature, prior approaches for evaluating or comparing models have accounted for only one or two of these factors. As far as we are aware, the approach we propose here is unique in that it accounts for, and disentangles, all three.

First, one would not expect a model of human behavior to make perfect predictions. Rather, the goal is most often to highlight key features of a phenomenon and explain choice patterns *reasonably well*. Restricting attention to perfect consistency with a model may thus be a lost opportunity, in that it obscures the model's general goodness of fit. This point is not only relevant for deterministic theories, which are fragile to noise, but also for the many theories of stochastic choice that have zero measure: consistency fails almost surely, however small the departure from the model.

Second, if we do expect some randomness in behavior (whatever its source), then the choice frequencies we observe reflect sampling variation. They need not approximate the anticipated behavior under the model, unless we have impractically many observations per menu. For a quick illustration, imagine the DM's true choice probability of picking a out of $\{a, b\}$ is 80%, and we observe 20 decisions from that menu (as will be the case in the data we analyze). There is roughly a 21% chance that she will choose a exactly 80% of the time (i.e., 16 times out of 20); and if we were to also observe 20 decisions from a second menu $\{c, d\}$ from which the DM true probability of picking c is 80%, then there would be

[2]For instance, Manzini and Mariotti (2007)'s Shortlisting model and Masatlioglu, Nakajima and Ozbay (2012)'s Limited Attention model are among several works that share the feature with rationality that choices will be deterministic.

[3]Seminal models of stochastic choice include Luce (1959) and the broader class of random utility models (RUM). Though multiple interpretations exist, Luce's model can be viewed as capturing the possibility a rational DM misperceives her utility and thus makes the wrong choice. RUM may be seen as capturing a fickle DM, whose preference changes with her state of mind. Alternatively, the DM's preference may be seen as changing with contextual details that the modeler does not observe, or does not know how they impact preference. For another, related interpretation, Gul, Natenzon and Pesendorfer (2014) observe that any RUM stochastic choice function is an attribute rule, or a limit of attribute rules. Of course, RUM also has an interpretation as describing the choice behavior of a collection of potentially heterogenous but rational individuals. Many other theories of stochastic choice have been studied. Some impose further structure; see for instance Gul and Pesendorfer (2006)'s random expected utility model, and Apesteguia, Ballester and Lu (2017)'s single-crossing random utility model. Others accommodate larger sets of stochastic choice functions; see for instance Dardanoni, Manzini, Mariotti, and Tyson (2020)'s combination of cognitive capacity constraints and stochastic preferences, or Cattaneo, Ma, Masatlioglu and Suleymanov (2020)'s random attention model.

only a 4% chance that the DM's observed frequencies match her true choice probabilities for both menus. As more menus are observed, the probability of 'hitting' the right frequencies becomes negligible. Sampling variation thus adds a layer of complication to assessing a model's goodness of fit, as the fit captures the disparity between true choice probabilities and the model. Moreover, if we acknowledge that a DM might not follow a deterministic model perfectly, then sampling variation is not just an issue for stochastic choice theories: it must be tackled for deterministic theories as well.

The third factor is whether it is surprising for a behavior to be explained by a theory (taking into account the questions being asked of the DM), given some level of goodness of fit. For instance, when one theory is contained in another, it may be more surprising to be near behaviors predicted by the smaller theory than the larger one. More starkly, some theories, including Luce (1959) and any deterministic theory, make knife-edge predictions. It is perhaps more impressive to be near those behaviors than it is to be perfectly consistent with a theory such as RUM, which permits a positive measure of stochastic choice functions. This relates to the issue of power that is discussed at length when testing rationality and refinements of it in consumer theory.

In the spirit of Becker (1962) and Bronars (1987), we take a uniform prior over behaviors as a benchmark when assessing how successful a model is at explaining choices. We assess a model's goodness of fit by minimizing what we call the *margin of error*; that is, by measuring how close the DM's true choice probabilities are to the stochastic choice functions arising under the model.[4] Under the assumption that a DM's observed choice frequencies perfectly reflect her true choice probabilities, one could assess the fit a model achieves to those observed frequencies. For each model, one may then compare the distribution of fits across subjects to the distribution arising under the random-choice benchmark.[5] Sampling variation means that fits measured this way are bound to be inaccurate for many subjects. This issue should be attenuated when looking at the distribution of all subjects' fits *as a whole*, providing a first look at the data (see Section 4). However, a full analysis of model selection at the individual level must also incorporate the presence of sampling variation.

Bayesian model selection is an important tool that accounts for both sampling variation and model size (the second and third factors above, respectively). The evidence for a model is examined through the lens of the Bayes factor, which measures how much more likely it

[4]Goodness of fit was introduced in the stochastic-choice literature by Apesteguia and Ballester (2021), but without considering power and sampling variation. Conceptually different, their measure assesses the maximal weight one can place on some behavior arising under the model when explaining choice probabilities. In the appendix, we replicate our analysis using their measure to assess fit. Qualitative results remain unchanged.

[5]Such comparisons are reminiscent of techniques deployed in the consumer theory literature using Afriat's *critical cost efficiency index* (CCEI); see the related literature section.

is for the observed choices to arise under one model or another. To better understand the meaning of the Bayes factor, imagine first comparing the model of interest to an alternative, unrestricted one (remember also that the prior distribution here is the uniform benchmark). As detailed in Section 2.2, the Bayes factor for the model of interest boils down to a likelihood ratio, comparing the probability of the model under the *posterior* distribution of behaviors derived from the data (in the numerator) to the corresponding probability under the prior distribution (in the denominator). A large Bayes factor means that the subject's choices tilt the prior in a direction that makes it more likely to generate behavior that 'hits the model.' When comparing two (nontrivial) models of interest, the Bayes factor describing the strength of evidence for one model over the other is simply the ratio of the two models' Bayes factors, each computed against the unrestricted alternative. Notice that the Bayes factor may in general be any nonnegative number. Typically, one says that the strength of evidence for a model over some alternative is 'substantial' when the Bayes factor is between $\sqrt{10}$ and 10, 'strong' when it is between 10 and 100, and 'decisive' when it is above 100 (Jeffreys 1961; Kass and Raftery 1995).

Our approach acknowledges that a model need not perfectly accommodate true choice probabilities–even if sampling variation were not an issue. In other words, 'hitting the model' is too restrictive a notion, unless we are more generous in our definition of a model. Hence the 'models' we consider when computing Bayes factors are expanded versions of corresponding baseline models: these comprise those choice probabilities that meet a threshold level of goodness of fit to the baseline model itself. The Bayes factor for an expanded model is then the likelihood ratio of achieving at least the given threshold of fit to the baseline model, under the posterior versus the prior distributions. The baseline models we consider are rationality, deterministic models of bounded rationality, expected utility with a CRRA specification, Luce (1959), RUM, Apesteguia, Ballester and Lu (2017)'s SCRUM with a CRRA specification, and a variant of Luce (1959) restricting to underlying CRRA preferences. We consider the individual-level dataset collected by Regenwetter et al. (2011) that replicates Tversky (1969)'s influential and controversial experiment of repeated choices from binary menus of lotteries.[6] We use Monte Carlo simulation to compute the relevant Bayes factors.

We apply Bayesian model comparison at the individual level, to identify both the optimal fit for each baseline model (the fit that maximizes the Bayes factor among all expansions of the given baseline model), as well as the best model overall (once optimal fit is taken into

[6]Originally, Tversky recruited 18 subjects; but he then used screening questions to select 8 subjects who seemed the most likely to be irrational. As discussed later below, Tversky's conclusion that these subjects violate weak stochastic transitivity were called into question on the basis of potentially arising from sampling variation. Regenwetter et al. (2011) replicates the experiment with a full set of 18 subjects and adjusting prizes for inflation.

account) for a subject. We also consider a more parsimonious exercise, which continues to take advantage of the individual-level data: can a single baseline model and a single margin of error be used to capture the behavior of many, if not all, subjects?

Some lessons emerge from the analysis. We provide here a quick preview of our results, before discussing related literature below. Though subjects' choice frequencies are often far from being deterministic, rationality as a baseline model has substantial explanatory power. But restricting attention to CRRA preferences has even greater appeal. For instance, whenever the optimal fit using rationality as baseline generates a Bayes factor larger than 10, the optimal fit using only CRRA preferences is identical, and generates a Bayes factor that is at least ten times larger. To be clear, this does not invalidate rationality. Rather, it suggests for this experiment that when a subject generates strong evidence for rationality, an even more stringent specification is reasonable.

But optimal fit under these baseline models, and the associated Bayes factors, is low for multiple subjects. Which model helps improve this? Moving to deterministic bounded rationality theories that accommodate all pairwise choice patterns (as do Limited Attention and Shortlisting) does not help. For instance, whenever rationality as baseline generates a Bayes factor smaller than 10 (which is the case for half the subjects), the optimal Bayes factor when using the expanded model is smaller for all but one subject. By contrast, standard stochastic choice models do succeed at bringing optimal fits and associated Bayes factors substantially higher for multiple subjects. Taking all models into account, rationality with CRRA preferences and the Luce model with CRRA preferences are most often the preferred baseline models. It is also worth noting that RUM, a specification the literature has favored for describing individual choices in this experiment, almost never emerges as a preferred baseline model. Again, this does not mean that RUM is refuted. Instead, the data suggests that more discerning models are warranted, at least in this experiment.

An intuitive tradeoff arises when using a single model to describe the choices of many: further enlarging the baseline model by decreasing fit is helpful for accommodating additional subjects, but lowers the significance of the explanation found for the subjects who were already accommodated with larger fit. The total impact on the aggregate Bayes factor is thus a priori ambiguous. A single-peaked relationship is observed, however, with the largest factor reached at 8 subjects with a fit of roughly 91% for rationality with CRRA preferences, 13 subjects with a fit of roughly 98% for the Luce model with CRRA preferences, and 16 subjects with a fit of roughly 99% for RUM.

Related literature

Assessing the performance of choice models at explaining actual choices, and comparing their relative performance, is a time-honored problem. Our work here contributes in several ways. We use a common methodology to assess and compare some central choice models in economics: rationality, deterministic bounded rationality, Luce, RUM, and their CRRA expected-utility counterparts. A characteristic feature of our analysis is the recognition that models of human behavior (a) are valuable even if capturing choices imperfectly,[7] and (b) may vary in their permissiveness on the tested menus, which should be factored in when assessing performance.[8]

Some important contributions on the topic come from the mathematical psychology literature. Psychologists since Tversky (1969) have been interested in understanding whether people act transitively. Iverson and Falmagne (1985) call into question Tversky's observation that choice frequencies violate weak stochastic transitivity; they point out that the observed violations may be attributable to sampling variation, and find that WST cannot be rejected at a 5% significance level for all but one of Tversky's 8 subjects. Regenwetter et al. (2011) focuses on RUM as a way to establish transitivity, interpreting its success in their dataset as a form of rationality: subjects maximize a preference ordering, albeit an unstable one. Cavagnaro and Davis-Stober (2014) were the first in this literature to use Bayes factors for model selection, asking which axiom of transitivity (weak, moderate or strong) or the RUM best captures individual choices.

We pursue a Bayesian model selection exercise, using several key economic models of choice, and expanding the set of SCFs they capture based on realized fit. From the perspective of economists, it would be useful to know whether more parsimonious models of choice provide a good fit to the data; how well common parametric specifications perform; whether recent deterministic models of bounded rationality outperform the rational benchmark; and whether stochastic choice models are superior. Our approach builds on some different strands of the economics literature, and helps to bridge them in the process. We discuss these strands below.

The first is a literature investigating how to incorporate stochasticity into deterministic models of choice over lotteries like expected utility and rank-dependent utility. Relevant works include Harless and Camerer (1994), Hey and Orme (1994), Hey (1995), Harrison and Rutström (2009), and Conte, Hey and Moffatt (2011). These papers vary in whether

[7]Even with infinitely many observations, would we expect choice probabilities to precisely match those of a restrictive model, such as that of Luce for instance?

[8]For instance, it is relevant to know how much more permissive RUM is than SCRUM. Similarly, being close to Luce may be more remarkable than being perfectly consistent with RUM.

they fit a model to pooled (aggregate) data, fit individual models, or fit a mixture model. The works in this vein emphasize parsimony, interpreted as a restriction on the number of parameters one uses when specifying both a choice model and a noise process, and look for the combination that maximizes the likelihood of the data. Maximum likelihood estimation takes into account sampling variation (though is known to be sensitive to such variation when samples are small), but does not effectively consider model size. For an extreme and artificial example, one can construct a single-parameter model that spans all choice functions. Even for well-known models, the number of parameters they have may be of little help for assessing their flexibility, as Fudenberg, Gao and Liang (2021) point out. We account for the richness of the behaviors the (expanded) model actually accommodates through Bayesian model comparison. This allows us to study how remarkable it is to be near one model or another given the dataset collected. Moreover, rather than committing to a structured noise process, we consider expansions of both deterministic and stochastic models based on measures of fit, and envision the decision maker as potentially using such a 'nearby' stochastic choice function.

Balakrishnan, Ok and Ortoleva (2022) is also interested in whether some form of rationality underlies stochastic choice data, though taking a different approach than the above works. They define a method for refining the support of observed choices into a choice correspondence, and propose a statistical test for whether the resulting correspondence satisfies the weak axiom of revealed preference. Balakrishnan et al. also make use of Regenwetter et al. (2011)'s dataset and join the growing consensus that Tversky underestimated the extent of rationality in subjects. In this paper, we consider models predicting stochastic choice functions (including deterministic ones), but do not consider choice correspondences. Our main focus is on the comparative performance of several classic models.

The second relevant strand of the economics literature pertains to the testing of rationality and its refinements in consumer theory. Representative works rely on a goodness-of-fit measure, most often Afriat (1973)'s CCEI; see Fisman Kariv and Markovits (2007), Choi, Kariv, Müller and Silverman (2014), Halevy, Persitz and Zrill (2018), and Polisson, Quah and Renou among others. Building on Bronars (1987), the typical approach compares the goodness-of-fit distribution of observed choices to the goodness-of-fit distribution one would get when drawing choices uniformly from budget frontiers. These papers focus on deterministic models. Though they account for goodness of fit through the CCEI, they do not consider the issue of sampling variation. Our paper imports ideas from this literature, and applies two innovations: a Bayesian model-selection approach and goodness-of-fit measures adapted to the current, discrete-choice setting. As seen in Section 3, our use of a uniform prior over stochastic choice functions renders our approach easy to compare and contrast

with that of Bronars'. Using Bayesian model selection to determine goodness of fit has the benefit of disentangling two separate issues: imperfectly following a benchmark model and sampling variation (which arises even if one perfectly follows a stochastic benchmark model).

We consider natural measures in lieu of the CCEI, which is not applicable to this setting. Our approach may be used in conjunction with different possible measures of goodness of fit. One of the measures we apply was proposed by Apesteguia and Ballester (2021), which they use to assess the fit of four different models (rationality, rationality with uniform noise, Luce and SCRUM) to pooled choice data over binary menus. Our work pushes their approach in several new directions. First, we are interested in individual-level data, as much information may be lost in the aggregation process. For instance, Luce may capture individual choices, and yet fail at the aggregate level due to preference heterogeneity. In that case, RUM may perform best in the aggregate. But RUM at the aggregate level allows for many possible behaviors at the individual level, having very different implications from a welfare perspective.[9] Is the subject endowed with a single preference ordering, multiple ones, or none at all? By incorporating RUM and the CRRA counterparts of different models into the analysis, we may understand, for instance, whether an individual follows RUM (as suggested by Regenwetter et al., 2011), Apesteguia, Ballester and Lu (2017)'s more parsimonious SCRUM, or even uses a single CRRA preference. Moreover, they apply their measure to observed frequencies, while we attempt to disentangle sampling variation from fit. They also do not delve into questions of power. Our analysis considers whether the increase in fit from using a more permissive model is meaningful, in the sense of reflecting regularities in the data instead of richness of the model, or a simple lack of power in the tested dataset.

The final related strand of the economics literature studies bounded rationality in choice theory. The past fifteen years have seen a revival of interest in this topic, and our paper applies the ideas discussed above to some of those new models. We test, for instance, whether models such as Manzini and Mariotti (2007)'s shortlisting and Masatlioglu, Nakajima and Ozbay (2012)'s limited attention, which allow for irrational deterministic choice patterns, perform better than rationality itself. Having data only over binary menus is certainly limiting, though, as these models (and many others) place no restriction on deterministic choice patterns.[10] The analysis remains important though, as it reveals whether poor fits under the rational choice benchmark are due primarily to the presence of complex choice

[9]Individuals might also be rational, actually use RUM, or even use particular *irrational* deterministic functions that happen to aggregate as a RUM stochastic choice function.

[10]Note, however, that behavior over binary menus can have different implications on behavior in larger menus depending on the model. For instance, rationality on binary menus implies rationality on all menus under shortlisting, but not under limited attention.

patterns that are not too far from being deterministic, or the intrinsic randomness of subjects' choices. We find strong evidence for the latter over the former in the data we analyze. By contrast to rationality, which has been extensively tested in consumer theory, few works test the validity of deterministic theories of bounded rationality. Manzini and Mariotti (2006), which checks whether observed behavior directly verifies some key axioms, is a rare example.

2 Preliminaries

2.1 Theoretical framework

Let X be the finite set of all options DMs may encounter. A nonempty subset of X is called a *menu*. A *stochastic choice function* (SCF) describes for each menu S a probability distribution in $\Delta(S)$. The set of all stochastic choice functions is denoted $\mathcal{SCF}$. A *choice function* is the special case where the distribution for each menu is degenerate, thereby identifying a unique selection.

A model $\mathcal{M}$ spells out how the DM makes her choice, which narrows down the possible SCFs to some subset. To simplify notation, we use $\mathcal{M}$ to denote the set of SCFs described by the model, and use $m \in \mathcal{M}$ to denote an arbitrary SCF in the model. For example, under *Rationality*, the DM maximizes a preference ordering; so an SCF m is rational if, and only if, there exists an ordering $\succ$ on X such that for all menus S, $m_S(\cdot)$ puts probability one on the maximizer of $\succ$ within S. The model of expected utility restricts the possible preference orderings that may be applied; oftentimes, the modeler may consider only those arising from parametric classes of utility functions, such as constant relative risk aversion (a Bernoulli utility of the form $u(y) = y^{1-\rho}/(1-\rho)$). There are many interesting and important models beyond these workhorse models of economics; we detail further below in Section 2.5 the main models we consider in our application to the data.

In the data we analyze, DMs' choices were collected over binary menus. Hence, though it generalizes to any domain, we describe our methodology below for that special case. The dataset comprises independent choice observations, with n repeated observations from each menu in some collection $\mathcal{D}$.[11] The DM's observed choice frequency of choosing x from a menu $\{x, y\} \in \mathcal{D}$ is denoted $\phi_{obs}(x|\{x, y\}) \in \{0, 1/n, 2/n, \ldots, 1\}$. The DM's underlying behavior, which we cannot perfectly observe unless $n = \infty$,[12] is given by ϕ, which describes a choice distribution in $\Delta(S)$ for each $S \in \mathcal{D}$. We say ϕ is *consistent* with the model $\mathcal{M}$ if there is $m \in \mathcal{M}$ such that $\phi(\cdot|S) = m(\cdot|S)$ for all $S \in \mathcal{D}$.

[11] Other datasets may involve a different number of repetitions for different menus. There is no conceptual change in what follows, only more cumbersome notation.

[12] In that case, ϕ coincides with ϕ_{obs} on $\mathcal{D}$.

2.2 Goodness of fit

Being consistent with a model $\mathcal{M}$ is an all-or-nothing proposition. Yet a model in economics is rarely expected to provide a perfect description of a DM's behavior. Just to introduce the measures of goodness of fit we study, in this subsection we abstract from the problem of sampling variation and imagine we know $\phi(\cdot|S)$ for each menu $S \in \mathcal{D}$; the next subsection relaxes this.

There are many ways to gauge how close the data is to a model $\mathcal{M}$, and we consider the robustness of our results to a couple of conceptually distinct measures. The approach we focus on in the main text is a simple margin of error. Fixing a menu $\{x, y\}$ and an SCF m arising under $\mathcal{M}$, the distance between ϕ and m is $|\phi(x|\{x, y\}) - m(x|\{x, y\})|$ $(= |\phi(y|\{x, y\}) - m(y|\{x, y\})|)$. Zero distance amounts to a perfect fit, while a distance of 1 is worst. Hence we measure m's fit for ϕ given $\{x, y\}$ by $1 - |\phi(x|\{x, y\}) - m(x|\{x, y\})|$. But observations span multiple menus, and $\mathcal{M}$ comprises multiple SCFs. Here we simply follow the method underlying Afriat's CCEI to obtain a full-fledged goodness of fit measure.[13]

Definition 1 (Margin-of-Error Approach)**.** *Under the margin-of-error approach, the goodness of fit $\mathcal{M}$ provides for ϕ is measured by:*

$$GF(\phi, \mathcal{M}) = 1 - \inf_{m \in \mathcal{M}} \max_{\{x,y\} \in \mathcal{D}} |\phi(x|\{x, y\}) - m(x|\{x, y\})|.$$

In this approach, achieving a goodness-of-fit γ means there is an SCF in the model whose probabilities fall within distance $1 - GF(\phi, \mathcal{M})$ of observed choice frequencies in *all* observed menus. Let $\mathcal{M}^\gamma$ be the set of SCFs ϕ such that $GF(\phi, \mathcal{M}) \geq \gamma$. The expanded model $\mathcal{M}^\gamma$ becomes increasingly restrictive as γ increases, and is simply $\mathcal{M}$ when γ equals one.

The approach we apply in this paper could also be used with other measures of goodness of fit. For instance, Apesteguia and Ballester (2021) propose a conceptually different measure that captures the DM's *reliance* on the model; that is, how often it is used to make decisions.

Definition 2 (Model-Reliance Approach)**.** *Using Apesteguia and Ballester (2021)'s notion*

[13] Considering linear budget sets in consumer theory, define an inconsistency index $I(u, B)$ for $\phi(B)$ being selected from a budget set B under the utility function u as the minimum percentage of lost income necessary to guarantee $\phi(B)$ is u-optimal in the shrunk budget set. With fit as $1 - I(u, B)$, Halevy, Persitz and Zrill (2018) point out that Afriat (1973)'s CCEI extends this measure to account for all tested budget sets in $\mathcal{D}$ and all utility functions in a reference set $\mathcal{U}$: $CCEI = 1 - \min_{u \in \mathcal{U}} \max_{B \in \mathcal{D}} I(u, B)$. Of course, $I(u, B)$ is specific to rational models in consumer theory, and does not apply to our discrete-choice framework which accommodates both deterministic and stochastic behavior that may be rational or not. But the point is that once a measure of fit has been adopted for a specific menu and a specific occurence of the model, a same method is applied to expand it into a full-fledged goodness-of-fit measure.

of maximal separation, the goodness of fit $\mathcal{M}$ provides for ϕ is measured by:

$$GF_{mr}(\phi, \mathcal{M}) = sup\{\gamma' \in [0,1] \mid \exists m \in \mathcal{M} \text{ and } \sigma \in \mathcal{SCF} \text{ with } \phi = \gamma' m + (1-\gamma')\sigma \text{ on } \mathcal{D}\}.$$

One may analogously define the expanded model $\mathcal{M}^{\gamma}_{mr}$ to be the set of SCFs ϕ such that $GF_{mr}(\phi, \mathcal{M}) \geq \gamma$. We consider properties of the model-reliance approach in Section 7.1, showing that for stochastic choice theories, it provides a different ranking of closeness than the margin-of-error approach. Nonetheless, we see that our conclusions based on the margin-of-error approach remain robust when repeating the exercise using model reliance instead.

Regardless of one's preferred measure of goodness of fit, one must still address a fundamental problem: how to evaluate fit when the DM's true SCF ϕ is imperfectly observed due to having only a finite sample. The approach we lay out next accounts for this issue.

2.3 Bayesian model comparison

Our approach in this paper builds on Bayesian model comparison. Observed choice frequencies in the data only imperfectly reveal the true choice probabilities. Formally, with a prior over stochastic choice functions in mind, the data ϕ_{obs} yields a posterior distribution $\text{Prob}(\cdot|\phi_{obs})$ for the underlying ϕ. We consider a uniform prior over $\mathcal{SCF}$. This provides a natural, random choice benchmark, and, as will be seen, allows for a clean comparison with Selten's measure (Selten and Krischker 1983, Selten 1991) as well as approaches in consumer theory building on Bronars (1987).

The *Bayes factor* quantifies the evidence in favor of one model over another. Formally, it describes the relative likelihood of observing the data ϕ_{obs} under a model $\mathcal{M}$ versus an alternative model $\mathcal{M}_a$:

$$BF(\mathcal{M}, \mathcal{M}_a) = \frac{\text{Prob}(\phi_{obs}|\mathcal{M})}{\text{Prob}(\phi_{obs}|\mathcal{M}_a)}, \tag{1}$$

where $\text{Prob}(\phi_{obs}|\mathcal{M}') = \int_{m \in \mathcal{M}'} \text{Prob}(\phi_{obs}|m)\text{Prob}(m|\mathcal{M}')dm$ is obtained by integrating over SCFs within the model.[14] Applying Bayes' rule to (1) shows that

$$BF(\mathcal{M}, \mathcal{M}_a) = \frac{\text{Prob}(\mathcal{M}|\phi_{obs})}{\text{Prob}(\mathcal{M})} \Big/ \frac{\text{Prob}(\mathcal{M}_a|\phi_{obs})}{\text{Prob}(\mathcal{M}_a)}, \tag{2}$$

where $\text{Prob}(\mathcal{M}'|\phi_{obs}) = \int_{m \in \mathcal{M}'} \text{Prob}(m|\phi_{obs})dm$. The Bayes factor takes into account that observed frequencies imperfectly reveal choice probabilities, while also including an adjustment for model permissiveness, as captured by $\text{Prob}(\mathcal{M})$ and $\text{Prob}(\mathcal{M}_a)$. In general, the

[14] Given the uniform prior, $\text{Prob}(m|\mathcal{M}')$ is simply one divided by the measure of $\mathcal{M}'$.

Bayes factor may be any nonnegative number. As seen from (1), a large Bayes factor means it is much more likely to observe the data under $\mathcal{M}$; and as seen from (2), it also means that the subject's choices tilt the prior in a direction that makes it more likely to generate behavior that 'hits' the model $\mathcal{M}$. Based on Jeffreys (1961) and Kass and Raftery (1995), the strength of evidence for $\mathcal{M}$ against $\mathcal{M}_a$ is considered 'not worth more than a bare mention' when the Bayes factor is between 1 and 3, 'substantial' when it is between 3 and 10, 'strong' when it is between 10 and 100, and 'decisive' when it is above 100.

Notice from the expressions above that any two models $\mathcal{M}$, $\mathcal{M}'$ may be compared by the ratio of their Bayes factors using any common comparison model $\mathcal{M}_a$:

$$BF(\mathcal{M}, \mathcal{M}') = BF(\mathcal{M}, \mathcal{M}_a)/BF(\mathcal{M}', \mathcal{M}_a).$$

Moreover, for a given comparison model $\mathcal{M}_a$, the strongest model $\mathcal{M}$ simply maximizes the ratio of its posterior probability to its prior probability. For convenience, we take the comparison model $\mathcal{M}_a$ to be the unrestricted model $\mathcal{SCF}$. In this case, $BF(\mathcal{M}, \mathcal{SCF})$ is simply $\text{Prob}(\mathcal{M}|\phi_{obs})/\text{Prob}(\mathcal{M})$ for any model $\mathcal{M}$, and denoted $BF_{\mathcal{M}}$ for short. We will simply refer to $BF_{\mathcal{M}}$ as the Bayes factor for the model $\mathcal{M}$. This number describes the strength of evidence for the model $\mathcal{M}$ in the data, compared to an unrestricted, random choice benchmark in the spirit of Becker (1962) and Bronars (1987). When explicitly comparing two restrictive models of interest $\mathcal{M}$ and $\mathcal{M}'$, we will refer to $BF(\mathcal{M}, \mathcal{M}')$ (which is the ratio of their Bayes factors) as the Bayes factor for $\mathcal{M}$ over $\mathcal{M}'$.

The denominator of $BF_{\mathcal{M}} = \text{Prob}(\mathcal{M}|\phi_{obs})/\text{Prob}(\mathcal{M})$ uses the prior. Given our choice of a uniform prior, the denominator thus corresponds to Selten's notion of a theory's *area*. The numerator, however, relies on the posterior. Helpfully, the posterior is a well-known distribution. If our prior distribution for $\phi(x|\{x, y\})$ is uniform over $[0, 1]$, then the posterior distribution for $\phi(x|\{x, y\})$ is a beta distribution with parameters $\alpha = n\phi_{obs}(x|\{x, y\}) + 1$ and $\beta = n - n\phi_{obs}(x|\{x, y\}) + 1$, where $n\phi_{obs}(x|\{x, y\})$ is the number of times x is selected out of n observations from $\{x, y\}$.[15] As n goes to infinity, ϕ_{obs} converges to ϕ and the numerator $\text{Prob}(\mathcal{M}|\phi_{obs})$ will be either zero or one, depending on whether the observed frequencies belong to the model $\mathcal{M}$; in that case, the Bayes factor will favor the most restrictive model containing the observed frequencies.[16] The difficulty is that n is relatively small in actual datasets, so the numerator plays an active role in practice.

[15] More generally, with a beta(α, β) prior for 'success', and having observed d successes out of n trials, the posterior distribution is beta$(d + \alpha, n - d + \beta)$; see, for instance, Greene (2018, p. 701). A uniform distribution over $[0, 1]$ corresponds to a beta$(1, 1)$ distribution.

[16] One could think of Selten's *hit rate* as the fraction of individuals with a 1 (observed frequencies that belong to the model), which would make sense if the underlying model is deterministic, or it is stochastic *and* the data contains infinitely many observations.

Our analysis proceeds through two layers of model selection; we refer to it as *fit-optimized Bayesian model selection*. To fix ideas, take the margin-of-error approach throughout this paragraph (we thus suppress the margin-of-error identifier in the notation here).[17] The first layer of model selection is in the determination of goodness of fit for a fixed baseline model $\mathcal{M}$. Remember that $\mathcal{M}^\gamma$ is an expanded version of $\mathcal{M}$ (describing the set of all SCFs which have at least a γ-level of goodness of fit to $\mathcal{M}$) and is thus a model in its own right.[18] We denote by $\gamma_{\mathcal{M}}$ the level γ that maximizes $BF_{\mathcal{M}^\gamma}$. The best-performing expansion of the baseline model $\mathcal{M}$ is then $\mathcal{M}^{\gamma_{\mathcal{M}}}$. Comparing across any two different baseline models $\mathcal{M}$ and $\bar{\mathcal{M}}$ then corresponds to a second-layer of Bayesian model selection, based on evaluating $BF(\mathcal{M}^{\gamma_{\mathcal{M}}}, \bar{\mathcal{M}}^{\gamma_{\bar{\mathcal{M}}}})$. The approach thus requires finding $\text{Prob}(\mathcal{M}^\gamma|\phi_{obs})/\text{Prob}(\mathcal{M}^\gamma)$ for different models and levels γ. We assess these quantities through Monte-Carlo simulation, except in cases where a closed form solution is readily found analytically. Our computational procedures are discussed in Section 5 and further detailed in Appendix A. The dataset to which we apply them is discussed next.

2.4 Description of dataset

In a seminal paper, Tversky (1969) designed five binary lotteries, having exactly one nonzero cash prize each; moving across these lotteries, the winning prize increases but the winning probability decreases in small steps of 1/24 (overall, the expected payoff increases with the winning probability). Tversky conjectured that the larger payoff may drive choices when comparing lotteries involving similar probabilities, while expected payoffs will be salient when probability differences are stark. People thinking this way would display intransitive choices over these alternatives. Tversky had his subjects face each of the resulting 10 binary menus twenty different times, and noted that choice frequencies typically failed weak stochastic transitivity. Some weaknesses have been noted in this nonetheless highly influential work. Iverson and Falmagne (1985) show that Tversky's conclusion of intransitivity, based on a simple examination of observed choice frequencies, would be overturned using a proper statistical analysis. Other weaknesses pertain to the experimental procedure itself. While Tversky initially recruited eighteen subjects, he selected the eight subjects described in the experiment after screening out ten who appeared less likely to display intransitivity. Moreover, each subject's data was collected over multiple sessions, spread across four weeks.

Our dataset comprises the replication of Tversky's experiment by Regenwetter et al. (2011). The lotteries themselves are adjusted only for inflation, and displayed in Table 1. Like in Tversky's original experiment, subjects face each of the 10 binary menus of lotteries

[17] An analogous construction applies under the model-reliance approach.

[18] Remember that $\mathcal{M}^1$, the subset of SCFs which perfectly fit the model, is equal to $\mathcal{M}$ itself.

Lottery	a	b	c	d	e
Payoff	\$28.00	\$26.60	\$25.20	\$23.80	\$22.40
Probability	7/24	8/24	9/24	10/24	11/24

Table 1: Inflation-adjusted versions of Tversky (1969)'s lotteries tested by Regenwetter et al. (2011). Each lottery corresponds to a column: the listed payoff is received with the probability beneath it (\$0 is received with the complementary probability).

20 times; hence choice frequencies for each menu occur in multiples of 5%. Their replication address the issues with Tversky (1969) by using a full set of eighteen randomly-selected participants who answer questions over the course of one session. The experiment is carefully designed to minimize possible correlation in choices across repetitions of a menu, including the use of distractor questions like in Tversky (1969), as well as questions from other treatments, with the sequencing constructed so that the same menu never appeared in five successive sets of four questions.[19]

2.5 Tested baseline models

Though any number of models could be tested, we find it important to restrict attention to prominent, pre-existing models and specifications that are applicable to a large domain of choice. To illustrate, systematically picking the lottery giving the largest probability of winning a prize defines a valid choice rule when selecting between binary lotteries in the experiment whose data we analyze, but does not work well as a general model of choice under risk.[20]

For their central roles in economics, Rationality (RAT) and its standard refinements for choice under risk come first to mind. Expected Utility (EU) is an obvious candidate here. But, as emphasized before, it is important to understand what models entail on the set of tested menus, before bringing them to the data. It turns out that RAT and EU

[19]One of those other treatments involves lotteries over objects (non-monetary prizes), which we do not discuss here as many of the models we study take advantage of the richer domain of monetary lotteries. Another treatment involves another set of monetary lotteries, and is often considered a different stimulus in the mathematical psychology literature. We consider this other monetary treatment in Section 7.2.

[20]A related point should be emphasized here: the conclusion of our analysis always holds within the confines of the tested menus and lotteries, and generalizing those conclusions to problems involving more complex menus and lotteries can be done safely only after collecting more data. To illustrate, note that the simple choice rule suggested in the text agrees with the maximization of a CRRA preference. It is conceivable that a subject's choices agree with the CRRA model here, because they followed this simple rule in this experiment, but would disagree with the CRRA model in a more complex situation where the rule does not provide a clear choice.

are indistinguishable for questions a la Tversky (we thus refer to these jointly as RAT).[21] A fortiori, such data won't allow performance comparisons with intermediate models (such as rank-dependent expected utility) either. On the other hand, the CRRA specification commonly used in economics (which further restricts the Bernoulli function used under expected utility) does accommodate much fewer choice patterns, and is included in our analysis.[22]

One of our primary interests was also to take stock of, and assess, models developed in the recent literature on bounded rationality. For deterministic choices, most of them suggest choice procedures that generalizes rationality to accommodate a wider range of choice patterns. Consider for instance Manzini and Mariotti (2007)'s Shortlisting, or Masatlioglu, Nakajima and Ozbay (2012)'s Limited Attention. Under Shortlisting, the DM eliminates options that are dominated according to an acyclic binary relation, and then selects an element from the shortlist by maximizing an asymmetric relation. Limited Attention posits that the DM picks the best element, according to a standard preference ordering, from the subset of options she considers; the only requirement on consideration sets is that they stay the same when removing options that are not considered. With data over binary menus (as in most[23] experiments collecting repeated individual choices), any deterministic choice function is compatible with these two prominent theories of bounded rationality (and many others).[24] Such permissiveness may leave the false impression there is nothing interesting left to test. On the contrary, one should wonder when finding a large margin of error under RAT whether it is primarily due to the prevalence of irrational choice patterns with limited noise, or to substantial stochasticity in behavior. For this reason, we include DET, the set of all deterministic functions, in our analysis.

Stochastic models also have a long tradition in economics. Sometimes, they are used as an econometric tool for adding noise to a deterministic model. By contrast, noise is accom-

[21] Indeed, because the prizes and probabilities $(m_i, p_i)_{i=1}^5$ move inversely, it is clear that one can construct a strictly increasing utility function $u(m_i) = 1/p_i$ and $u(0) = 0$ for which all the lotteries are indifferent; and this utility function can be perturbed slightly to replicate any preference ordering over the lotteries while remaining strictly increasing.

[22] CARA, another common specification, is comparable in size to CRRA, but almost indistinguishable to CARA, over the binary menus tested here (sharing 10 out of 11 preference orderings). Thus, the existing data analyzed here does not allow us to assess the relative prevalence of these different specifications, only the potential validity of using a standard parametric specification beyond expected utility.

[23] Some notable exceptions include Tversky (1972) and McCausland et al. (2020).

[24] Any choice function c over binary menus arises under Limited Attention by taking $\{c(\{x, y\})\}$ as the attention set for a binary menu $\{x, y\}$ (attention sets are left unrestricted over binary menus as they cannot be compared by inclusion). Shortlisting can also accommodate c: fix an ordering $\succ$ to serve as the second criterion, and define the first one as x above y if $c(\{x, y\}) = x$ and $y \succ x$. This argument also shows that any choice pattern over binary menus can be explained by menu-dependent preference maximization with only two selves (Kalai, Rubinstein and Spiegler, 2001).

modated in our analysis by recognizing that theories need not provide perfect fits and we are interested in testing the overall validity of the specific predictions that these stochastic models make. RUM, for instance, may be seen as capturing a fickle DM, whose preference changes with her state of mind.[25] Luce (1959)'s model can be viewed as capturing the possibility of a rational DM misperceiving her utility and thus makes the wrong choice on occasion. Our analysis naturally covers these two central models of stochastic choices. As for deterministic models, we will also include their variants when restricting attention to expected utility with CRRA: Apesteguia, Ballester and Lu (2018)'s single-crossing random utility (SCRUM) defined over CRRA preferences, as well as the Luce model with an underlying preference from the CRRA class (CLUCE).[26]

In summary, we study seven baseline models that are distinguishable in this context: CRRA, RAT, DET, C-LUCE, LUCE, SCRUM and RUM.[27]

3 Model Size

How rich are classic models of choice? This question has not been addressed much in the literature. One obvious issue is that multiple baseline models have zero measure. Assessing fit instead of perfect consistency offers a way forward, which also highlights another relevant feature when discussing richness: two baseline models of equal measure may differ in size when considering nearby SCFs. For an abstract analogy, any finite set A of probabilities in the interval $[0, 1]$ has zero measure, but the set of probabilities falling within distance $\varepsilon > 0$ of A varies in size with the composition of A. Figure 1 provides a visual assessment of model size, by presenting the CDF of realized fit for ten million SCFs drawn uniformly at random. These curves could be viewed as the analogue for discrete, stochastic choices of Bronars' curves drawn in consumer theory.

A fit of one means perfect consistency. All baseline models we consider, except for SCRUM and RUM, have zero measure. Their CDFs in Figure 1 reach the top edge of the square as fit is strictly inferior to one for all ten million SCFs. By contrast, both SCRUM and RUM have an atom at one. SCRUM consistency occurs, in theory, with probability[28]

[25]For another, related interpretation, Gul, Natenzon and Pesendorfer (2014) observe that any RUM stochastic choice function is an attribute rule, or a limit of attribute rules. Of course, RUM also has an interpretation as describing the choice behavior of a collection of potentially heterogeneous but rational individuals.

[26]As explained earlier, RAT and EU are undistinguishable with the present data. Hence so are RUM and Gul and Pesendorfer (2006)'s random expected utility.

[27]Note that CRRA is a subset of RAT, which in turn is a subset of DET. CRRA is a subset of SCRUM, which itself is a subset of RUM. CRRA is also included in the closure of C-LUCE, which is itself a subset of LUCE. Finally, RAT is contained in the closure of LUCE, which is itself contained in RUM.

[28]See Appendix A.2.

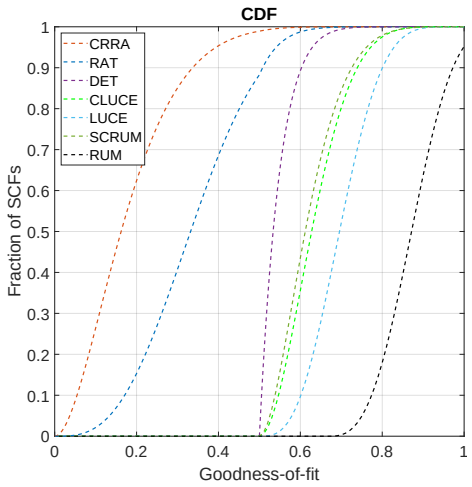


Figure 1: Model Size Using Fit

$1/(10!)$; this is on the order of 10^{-7}, and none of the 10 million randomly-selected SCFs were consistent with it. By contrast, approximately 4.79% of the randomly drawn SCFs are consistent with RUM. As this suggests, RUM is far more permissive than any of the other baseline models we test. This remains true at all levels of fit, as we see the RUM CDF is first-order stochastically superior, and very much so, to all other CDFs in Figure 1. For a stark comparison, it is harder to achieve a fit of merely 40% to the baseline model of rationality with CRRA preferences than it is to achieve perfect consistency with RUM.

To better understand this, let's pause a moment to consider how one computes the margin of error between a deterministic model and observed frequencies. For each binary menu, the choice frequency is a number between 0 and 1, while deterministic choice functions pick *either* 0 or 1. Not surprisingly, the margin of error under DET is always lower than or equal to 1/2 (as can be seen from the corresponding CDF in Figure 1): we can pick the choice function that falls for each menu on the same side as the observed frequency (that is, picking 0 for menus where the frequency is lower than 1/2, and 1 otherwise). What is more surprising, however, is to get a goodness of fit larger than 50% for a model that restricts choices across binary menus. While there are $2^{10} = 1024$ possible choice patterns under DET, only 120 are compatible with Rationality and a mere 11 are compatible with the maximization of a CRRA preference. Consider, for instance, a 50% fit with the baseline model of rationality with CRRA preferences. Although the gap between the model and the actual choice frequency can be substantial in any given menu, what is more remarkable when achieving such a fit for an SCF is the possibility of finding one of the eleven CRRA choice patterns that correctly predicts, for *all* ten menus, whether the choice frequency in that menu will be bigger or smaller than 1/2.

Luce with CRRA preferences, and Luce itself, have zero measure. Consistency with these models is thus harder to obtain, and in some sense infinitely so, than consistency with SCRUM. Yet there is a fit at which the comparison reverses. For instance, Figure 1 shows that achieving an 80%-fit (or below) to SCRUM is a bit harder than achieving that same fit to Luce with CRRA preferences. Despite having zero measure under perfect fit, Luce with CRRA preferences generates, in some sense, more dispersion in the set of consistent SCFs. This further illustrates how considering the entire CDF of model size across fits, instead of simply measuring the probability of perfect consistency, can be insightful for understanding a model's richness.

4 First Look at the Data

In this section, we compare model-by-model the CDF of fits associated to the DMs' *observed* choice frequencies (as if they perfectly reveal underlying SCFs) with the random-choice benchmark from the previous section. Of course, sampling variation implies that measuring fit this way is bound to be inaccurate for many subjects. But this issue should be attenuated when looking at the distribution *as a whole*. While the more elaborate analysis of the next section will provide insights for model selection at the individual level, the preliminary approach we pursue here provides a quick, visual way to roughly assess the models' overall performance. It also facilitates comparison with the CCEI-based approach of consumer theory.

Figures 2 and 3 pursue this exercise for the deterministic and stochastic models, respectively. In all cases, the step lines represent the 18 subjects, while the dashed curves represent the 10 million 'fictitious subjects' whose choice probabilities were drawn uniformly at random (see the previous section). In theory, the step lines could fall anywhere in comparison to the dashed curves. But, for all models other than DET, goodness-of-fit distributions based on actual choices are first-order stochastically superior to those based on random choices, and substantially so. This suggests that they indeed capture some of the driving forces that underlie the choices.

The deterministic models in the panels of Figure 2 become more permissive when moving from left to right (rationality with CRRA preferences, unrestricted rationality, and finally all deterministic choice patterns). A more permissive deterministic model will, in general, not improve the fit achieved for subjects where the less permissive model already achieves a fit above 50%.[29] The hope in choosing a more permissive deterministic model is to obtain a

[29]Suppose a deterministic m is such that $|m(x|\{x,y\}) - \phi(x|\{x,y\})| < 1/2$. Then $m(x|\{x,y\})$ is the unique minimizer of $|m'(x|\{x,y\}) - \phi(x|\{x,y\})|$ over all $m'(x|\{x,y\}) \in \{0,1\}$. If $|m(x|\{x,y\}) - \phi(x|\{x,y\})| < 1/2$

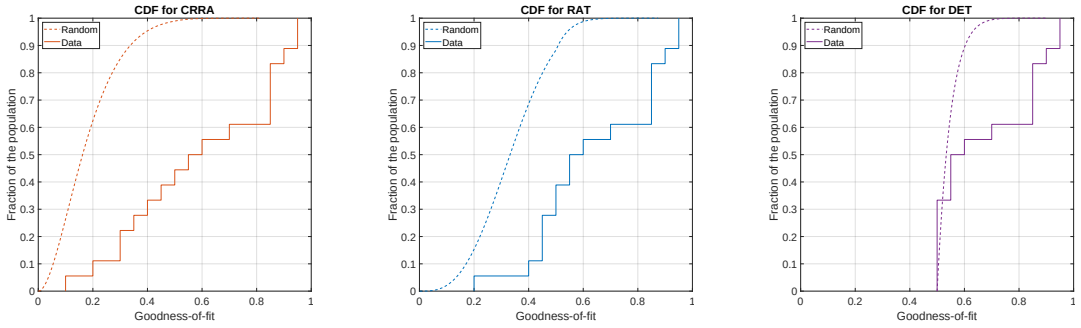


Figure 2: Realized vs. benchmark fit for the deterministic models

better fit (as high as possible over 50%) for those subjects whose choice frequencies generate a fit below 50% under the original model. To illustrate with a toy example involving only three alternatives, suppose that a subject's choice frequencies are 85% a's from $\{a, b\}$, 75% b's from $\{b, c\}$ and 95% c's from $\{a, c\}$. The margin of error under rationality is 0.75 using $c \succ a \succ b$ (leading to a goodness of fit of 25%). In this example, allowing for any deterministic function reduces the margin of error to 0.25 (leading to a goodness of fit of 75%), by selecting a from $\{a, b\}$, b from $\{b, c\}$ and c from $\{a, c\}$. This is a case where, although there certainly remains some unexplained randomness in choices, relaxing rationality to accommodate all deterministic choice patterns would substantially improve our understanding of the subject's choice frequencies.

Yet looking at Figure 2, essentially all subjects with a fit below 50% under one of the more restrictive models get pooled at the smallest possible fit of 50% under DET instead of getting a much larger index (as one might have hoped). By contrast, the random-choice benchmark curve substantially shifts to the right as the model becomes more permissive. This suggests that expanding Rationality (or CRRA) to accommodate all irrational deterministic choice patterns is unhelpful.[30]

The above analysis suggests that we must look at stochastic choice models if explanatory power is to be improved. Comparing Figures 2 and 3, we see that for all four stochastic models we consider, the realized fits are markedly higher than those under the deterministic

holds for all $\{x, y\} \in \mathcal{D}$, then m is the unique minimizer of the margin of error over all of DET.

[30]One might be skeptical of this analysis, worrying that the disappointing performance of DET is mostly due to the worst-case approach underlying our goodness-of-fit measure: a fit of 50% obtains if, for some menu, the subject picked the two options equally often. Could it be that DET's performance improves when errors in all menus are weighted in? In Appendix B.1, we consider mean-squared errors as an alternative, with fit measured as $1 - \min_{m \in \mathcal{M}} \sum_{\{x,y\} \in \mathcal{D}} \frac{(\phi(x|\{x,y\}) - m(x|\{x,y\}))^2}{\#\mathcal{D}}$. The resulting step lines and dashed curves are of course different, but the qualitative result remains unchanged: while the realized fit distribution first-order stochastically dominates, and substantially so, the random-choice fit distributions for both rationality and rationality with CRRA preferences, the unrestricted deterministic model performs poorly (the former distribution is in fact first-order stochastically inferior to the latter for almost half the subjects).

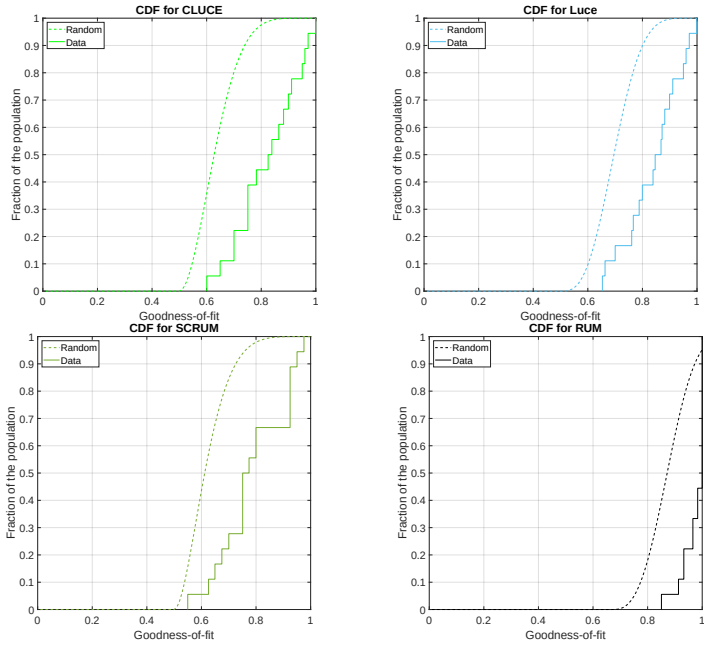


Figure 3: Realized vs. benchmark fit for the stochastic models

models. However, the fits also increase substantially under the random choice benchmark. For instance, the bottom-right panel in Figure 3 shows, as Regenwetter et al. (2011) already observed, that the choice frequencies of 10 subjects are perfectly consistent (100% fit) with RUM. It also shows that the fit remains high for many more subjects (not surprising given Regenwetter et al.'s finding that RUM cannot be rejected at the 5% level for most subjects). But the random-choice fit CDF (dashed curve) also highlights that the model is rather permissive given the collection of tested questions. This preliminary analysis does not provide a clear sense of which model performs best. The Bayes factor analysis pursued in the next section incorporates sampling variation to shed light on this question at the individual level.

5 Fit-optimized Bayesian model comparison

Recall from Section 2 that each subject's observed choice frequencies lead to a posterior distribution for her true, underlying SCF. For any model $\mathcal{M}$ and index γ, the Bayes factor $BF_{\mathcal{M}^\gamma}$ is the probability that the subject's underlying SCF belongs to $\mathcal{M}^\gamma$ under the posterior distribution, divided by the corresponding probability computed under the uniform prior distribution instead.

Aside from a few exceptions where a closed-form solution is analytically available, we use Monte Carlo simulation to compute $BF^i_{\mathcal{M}^\gamma}$ for each subject i, each baseline model $\mathcal{M}$,

and each percentage fit γ (in steps of 1%).[31] In Appendix A we detail our computational procedures, which were implemented using Matlab and benefitted from the use of Brown University's high performance computing cluster. For each M^γ, denominators and numerators are found by drawing a 'large' set of SCFs (as described in Appendix A), using the prior distribution when computing denominators and a subject's updated beta distribution when computing numerators. We then test for each of those SCFs whether it belongs to M^γ. For the deterministic models, the speed of computations permits drawing 1 billion SCFs for both numerators and denominators at all levels of fit. The computations for stochastic choice models are more time intensive. Those models are tested by drawing 10 million SCFs for numerators and denominators.[32] We increase the draw of SCFs by an order of magnitude to recompute denominators, up to a maximum of 100 billion, to ensure precision for high levels of fits when the model becomes rare (see details in Appendix A).

For each subject and model $\mathcal{M}$, we refer to the fit γ maximizing $BF_{\mathcal{M}^\gamma}$ as the subject's *optimal fit for* $\mathcal{M}$; we refer to the maximal value of $BF_{\mathcal{M}^\gamma}$ as the subject's *optimal Bayes factor for* $\mathcal{M}$. These quantities are summarized for each subject and model in Table 2. When the Bayes factor is below 1000, we report the number rounded to the nearest first decimal. To improve readability of the table, we report here only the magnitudes for larger Bayes factors, as powers of 10 (e.g., a Bayes factor for subject 8 of around $35,284$ is reported as $\sim 10^4$ in the table).[33] The largest optimal BF that can be achieved for each subject, when considering our seven baseline models, is bolded in the table. This provides a bird's eye view of model performance for each subject. A more detailed visualization is also possible. For two representative subjects discussed further below, Figure 4 provides a graph plotting, for each of the seven baseline models, the subject's realized Bayes factor as a function of fit. The corresponding figures for all subjects are found in Appendix B.

There are several takeaways. The optimal BFs and corresponding fits suggest that significant regularity is explained. Recall from Section 2.3 that the strength of a model $\mathcal{M}_1$ against $\mathcal{M}_2$ is considered 'substantial' when the Bayes factor is between $\sqrt{10}$ and 10, 'strong' when it is between 10 and 100, and 'decisive' when it is above 100. Compared to the random-choice benchmark, Table 2 shows that with the singular exception of Subject 4, we can identify for each subject a model for which there is substantial evidence (i.e., whose optimal BF is above

[31] Closed-form solutions are available for the denominators of SCRUM with $\gamma = 1$ and any deterministic model with $\gamma > 1/2$. For SCRUM with $\gamma = 1$, this value is $1/10!$, as noted earlier and shown in Appendix A.2. For deterministic models, if $\gamma > 1/2$ then there is no overlap in behaviors across deterministic choice functions when they are extended. So the area is $(1-\gamma)^{10}$ times the number of choice functions in the baseline model. That number is 11 for CRRA, $120 = 5!$ for rationality, and $1,024 = 2^{10}$ for DET.

[32] For Luce with CRRA preferences, which is the most computationally intensive model we consider, we begin by drawing just 1 million SCFs at low levels of fit, where the model is less rare.

[33] Table B.2 in the Appendix reports those numbers with greater precision.

	CRRA	RAT	DET	CLUCE	LUCE	SCRUM	RUM
Subject 1	4.1 (31%)	3.7 (49%)	1.0 (50%)	**37.9** (89%)	12.9 (87%)	5.7 (73%)	12.7 (100%)
Subject 2	**440.8** (67%)	41.5 (67%)	4.9 (67%)	123.9 (90%)	26.4 (89%)	25.9 (82%)	13.3 (100%)
Subject 3	$\mathbf{10^9}$ (99%)	10^8 (99%)	10^7 (99%)	10^5 (99%)	10^4 (99%)	10^3 (95%)	8.3 (98%)
Subject 4	1.0 (4%)	1.0 (18%)	**2.4** (64%)	1.0 (54%)	1.2 (64%)	1.0 (53%)	1.7 (88%)
Subject 5	$\mathbf{10^7}$ (93%)	10^6 (93%)	10^5 (93%)	10^3 (99%)	844.0 (99%)	10^3 (93%)	8.3 (98%)
Subject 6	3.4 (31%)	2.2 (42%)	1.0 (53%)	**21.5** (87%)	5.6 (83%)	1.9 (64%)	5.7 (100%)
Subject 7	$\mathbf{10^5}$ (84%)	10^4 (84%)	10^3 (84%)	292.4 (92%)	38.2 (91%)	590.7 (92%)	8.9 (99%)
Subject 8	$\mathbf{10^8}$ (96%)	10^7 (96%)	10^6 (96%)	10^5 (99%)	10^4 (99%)	10^3 (94%)	8.6 (98%)
Subject 9	6.2 (39%)	3.3 (47%)	1.0 (50%)	**20.7** (83%)	7.5 (83%)	9.4 (79%)	9.3 (100%)
Subject 10	$\mathbf{10^5}$ (86%)	10^4 (86%)	10^3 (86%)	10^3 (98%)	10^3 (99%)	10^3 (94%)	11.0 (100%)
Subject 11	$\mathbf{10^7}$ (94%)	10^6 (94%)	10^5 (94%)	10^5 (99%)	10^4 (99%)	10^3 (93%)	9.8 (99%)
Subject 12	1.6 (19%)	1.9 (37%)	1.0 (50%)	3.9 (78%)	3.1 (77%)	1.6 (63%)	**4.6** (98%)
Subject 13	12.4 (44%)	3.2 (46%)	1.0 (50%)	**83.5** (90%)	20.7 (90%)	11.9 (79%)	12.9 (100%)
Subject 14	$\mathbf{10^9}$ (99%)	10^8 (99%)	10^7 (99%)	10^5 (99%)	10^4 (99%)	10^3 (96%)	8.4 (98%)
Subject 15	58.4 (54%)	7.2 (53%)	1.0 (51%)	$\mathbf{10^3}$ (98%)	204.3 (98%)	57.3 (86%)	18.6 (100%)
Subject 16	$\mathbf{10^3}$ (79%)	519.3 (79%)	60.9 (79%)	4.6 (73%)	1.7 (71%)	47.1 (85%)	1.6 (87%)
Subject 17	4.8 (33%)	1.5 (31%)	1.0 (50%)	4.9 (74%)	2.2 (74%)	**6.4** (76%)	2.3 (92%)
Subject 18	35.3 (50%)	5.2 (50%)	1.0 (50%)	**990.6** (96%)	165.1 (99%)	51.5 (86%)	19.1 (100%)

Table 2: The optimal Bayes factor, with the corresponding fit underset in parenthesis, for each subject and model. For brevity, we simply state the order of magnitude for Bayes factors greater than 1000.

$\sqrt{10}$). But the Bayes factors we find are often much higher. For 11 out of the 18 subjects, the analysis identifies a model achieving an optimal BF of at least 100.

Moreover, there is often evidence in favor of one model over others. Formally, we say a baseline model is *undominated* if there is no other baseline model that achieves an optimal BF at least $\sqrt{10}$ times larger. For those 11 subjects with an optimal BF over 100, the winning model is the unique undominated model. This is particularly striking in the plots of Bayes factors, where one model often dwarfs the others; for such an example, see Subject 5 in the left panel of Figure 4. There is also a unique undominated model for two additional subjects whose optimal BF's are between 10 and 100. Thus for over two-thirds of the subjects, the

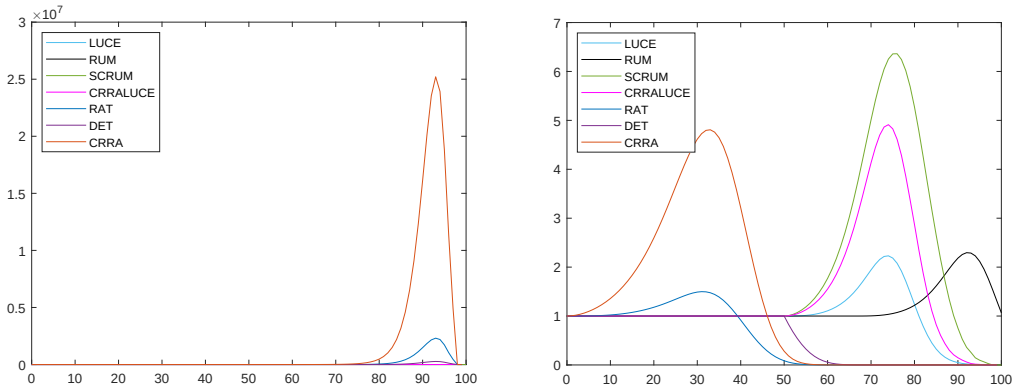


Figure 4: Example plots for Subject 5 (left) and Subject 17 (right) of the Bayes factor as a function of fit, for each of the seven baseline models. All plots in Appendix B.

analysis identifies a unique undominated model with strong evidence (or more) in its favor.

Looking down the columns of Table 2 to see which entries are bolded suggests a striking, systematic regularity. Among the seven baseline models considered, Bayesian model selection almost always narrows attention to either the deterministic CRRA model or Luce with CRRA preferences. Indeed, the deterministic CRRA model is selected for fully half of all subjects. More broadly, we see that whenever the optimal fit using rationality as baseline generates a Bayes factor larger than 10, the optimal fit using only CRRA preferences is identical, and generates a Bayes factor that is at least ten times larger. This suggests that the more parsimonious specification is preferable in those cases. On the other hand, whenever rationality as baseline generates an optimal Bayes factor smaller than 10 (which is the case for half the subjects), the optimal Bayes factor when using the more expansive deterministic model is smaller for all but one subject.

When a stochastic model is selected, Luce with CRRA preferences tends to be the only undominated model (6 out of 8 of these cases). When RUM is the winning baseline model (for Subject 12), we cannot rule out Luce with CRRA preferences. Similarly, when SCRUM is the winning baseline model (for Subject 17), both CRRA and Luce with CRRA preferences remain undominated and have substantial evidence. This equivocation is apparent in Subject 17's plots in the right panel of Figure 4.

The success of CRRA and Luce with CRRA preferences we see here does not contradict the success of RUM found in Regenwetter et al. (2011), since RUM is a more expansive model. But it does show that much more precise models can shed light on the data. A striking example is Subject 3, who may be deemed very close to RUM and yet, when more restrictive models are tested, the optimal fits we find are almost all higher than that for RUM. Overlooking this may result in information loss for welfare analysis, as in that particular case,

the models that improve upon RUM's fit all have an underlying preference.

6 One Size Fits Most

Section 5 provides a fine-grained analysis, indicating for each subject which baseline model and which fit maximizes a personalized Bayes factor. One may also be interested in comparing the ability of models to explain many participants at once, which may favor larger models as they offer more flexibility. Important information about model performance, however, may be lost by considering aggregate stochastic choice data (i.e., pooled over all subjects). Here, we retain the individual-level data and examine whether a single baseline model and a single margin of error can be used to capture the behavior of many, if not all, subjects. A benefit of this more parsimonious exercise is understanding whether some model works well for most subjects, despite heterogeneity and the possibility that some individuals are too anomalous to explain. A better understanding of model prevalence in this sense may also inform the study of more complex settings building on these choice models.[34]

Formally, we assume subjects may have different underlying SCFs and act independently. That is, one subject's observed frequencies do not shed light on the underlying SCFs of other subjects. As a start, one might want to evaluate the performance of a baseline model $\mathcal{M}$ in explaining all 18 subjects. Given independence, the joint Bayes factor $\overline{BF}_{\mathcal{M}^\gamma}$ for $\mathcal{M}^\gamma$ would be the product $\prod_{i=1}^{18} BF^i_{\mathcal{M}^\gamma}$ of all individuals' Bayes factors for $\mathcal{M}^\gamma$. More broadly, though, one may find a model impressive if it can well explain even n out of the 18 subjects. In that case, the joint Bayes factor $\overline{BF}_{\mathcal{M}^\gamma}$ of interest for $\mathcal{M}^\gamma$ would be the product $\max_{|S|=n, S\subset\{1,\dots,18\}} \prod_{i\in S} BF^i_{\mathcal{M}^\gamma}$. This is simply the product of the n largest individual Bayes factors for $\mathcal{M}^\gamma$.

For a given model, one may attempt to accommodate more subjects by enlarging the baseline model (i.e., reducing the goodness of fit). Doing so, however, reduces the Bayes factor for those subjects already accommodated at a higher level of fit. The total impact on the joint Bayes factor is thus a priori ambiguous. We perform this analysis. For each of our baseline models, Figure 5 plots the maximal achievable joint Bayes factor $\overline{BF}_{\mathcal{M}^\gamma}$ along with the fit γ that achieves it. As seen in Figure 5, optimal fit declines with the number n of subjects to be explained. On the other hand, the maximal joint Bayes factor has an inverse U-shape. The figure suggests that CRRA is a promising model for capturing around one-third of subjects. Luce (with and without CRRA preferences) and SCRUM, however, are about equally impressive in their ability to capture larger numbers of subjects (at least

[34]For instance, one may consider analogues to Eliaz (2002)'s problem of fault-tolerance implementation where at most k agents do not optimize correctly.

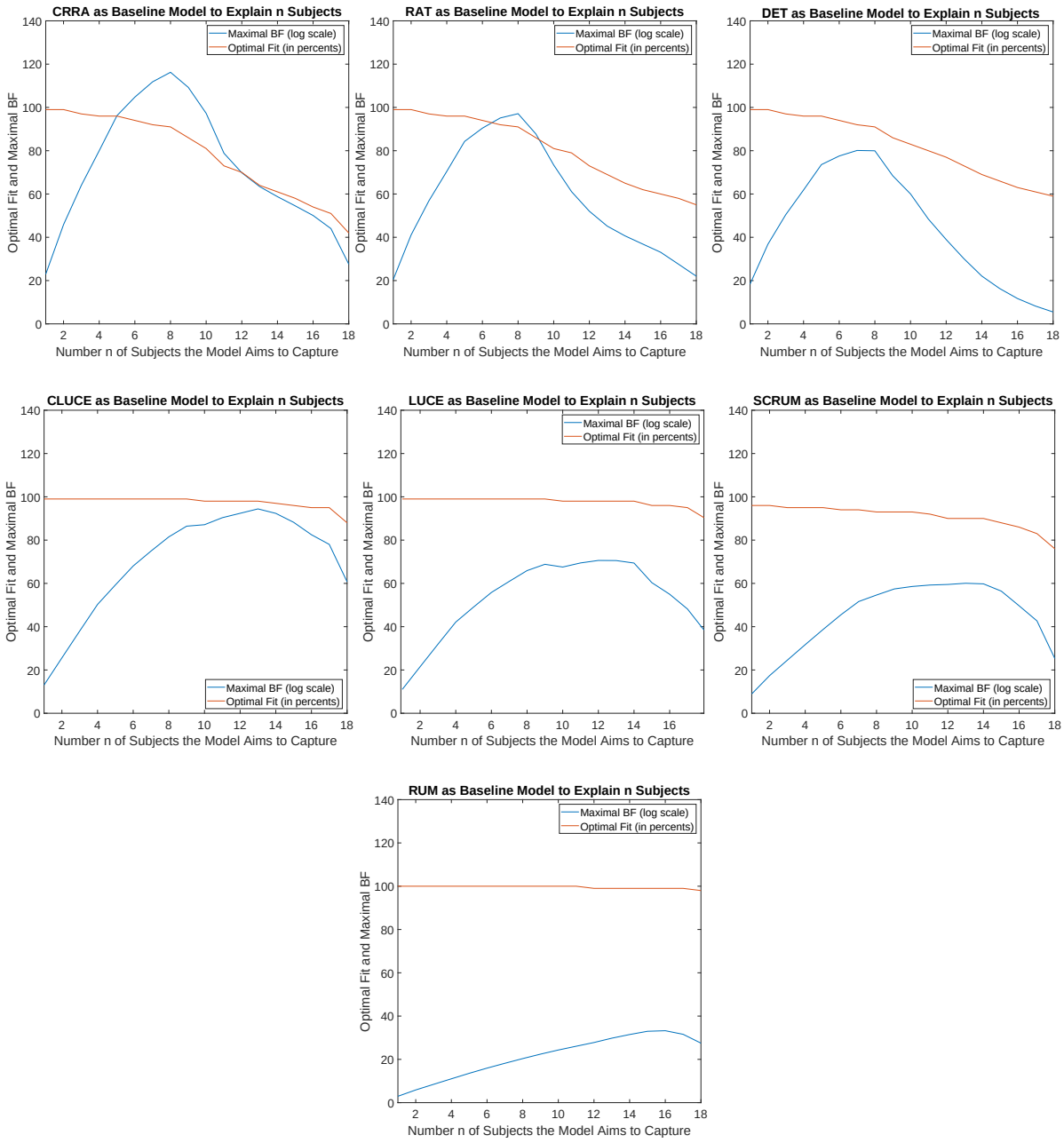


Figure 5: Using a single baseline model and margin of error to capture the behavior of many.

two-thirds). RUM captures all but a couple subjects with fits close to 100%. Perhaps surprisingly, the figure reveals that fits change relatively little when moving across some nested models (for instance, RAT versus DET, and CLUCE to LUCE). On the other hand, the significance of explanations, in terms of the joint BF, greatly decreases for a larger model. This suggest that more demanding specifications than RUM remain successful in capturing choices at large.

7 Concluding Observations

In this concluding section, we replicate our analysis (a) using a different measure of fit, and (b) over another set of related questions.

7.1 Model Reliance

Remember from Section 2.2 the goodness-of-fit measure GF_{mr} of Apesteguia and Ballester (2021), which is conceptually different from the margin of error approach applied in most of this paper (notated GF). Their Proposition 1.1 provides an alternative formula to compute GF_{mr} that we can use to facilitate comparisons with GF. In particular,

$$GF_{mr}(\phi, \mathcal{M}) = 1 - \inf_{m \in \mathcal{M}} \max_{\{x,y\} \in \mathcal{D}} I(\phi, m, \{x, y\}), \tag{3}$$

where the inconsistency between ϕ and m over $\{x, y\}$ is defined as

$$I(\phi, m, \{x, y\}) = \min\{\frac{|\phi(x|\{x, y\}) - m(x|\{x, y\})|}{m(x|\{x, y\})}, \frac{|\phi(x|\{x, y\}) - m(x|\{x, y\})|}{m(y|\{x, y\})}\}.$$

We recognize in (3) the same minmax operator as in the definition of GF (and the CCEI, see footnote 13), but applied to a different index. Though perhaps less obvious than measuring plain distances, $I(\phi, m, \{x, y\})$ admits an intuitive interpretation in terms of reliance. Indeed, it amounts to the minimum weight one needs to put on some unexplained noise *within the menu* $\{x, y\}$ to recover $\phi(x|\{x, y\})$ when otherwise using $m(x|\{x, y\})$:

$$I(\phi, m, \{x, y\}) = \min\{\alpha \in [0, 1] | \exists \nu \in [0, 1] \text{ s.t. } \phi(x|\{x, y\}) = \alpha\nu + (1 - \alpha)m(x|\{x, y\}).$$

Thus, while it may seem more demanding to decompose ϕ in its entirety into a convex combination of an SCF under the model and one capturing noise, this reinterpretation of the original definition shows the decomposition can equivalently be done menu by menu under a worst-case scenario analysis. The above rewriting using $I(\phi, m, \{x, y\})$ also makes the following observation straightforward.

Observation 1. *$GF_{mr}(\phi, \mathcal{M}) \leq GF(\phi, \mathcal{M})$, for all $(\phi, \mathcal{M})$, and the two measures coincide when $\mathcal{M}$ is deterministic.*

This means that in general, an expanded model under the model-reliance approach will be contained in the corresponding expanded model under the margin-of-error approach. The impact on the Bayes factor is thus ambiguous. Moreover, we note that in the case of

stochastic choice models, the two measures are inherently different; that is, one is not a monotone transformation of the other.

Observation 2. *For stochastic models the two approaches may offer opposite assessments of how close different pairs* (ϕ, m) *are within a menu* $\{x, y\}$.

To see this, consider the following example. Suppose $\phi_1(x|\{x,y\}) = 3/4$, $\phi_2(x|\{x,y\}) = 1/10$, $m_1(x|\{x,y\}) = 1/2$, and $m_2(x|\{x,y\}) = 3/10$. Under the margin-of-error approach, m_1 provides a 3/4 fit to ϕ_1, which is worse than the 4/5 fit m_2 provides to ϕ_2. Under the model-reliance approach, m_1 provides a 1/2 fit to ϕ_1, which is better than the 1/3 fit that m_2 provides to ϕ_2

We replicated the fit-optimized Bayesian model comparison exercise of Section 5 using the model-reliance approach. Though the deterministic model results would not change, those for the stochastic models could. While numbers vary, the qualitative results we find are strikingly similar. Indeed, the plots of optimal BF per model, provided in Appendix C, are strongly reminiscent of the plots under the margin-of-error approach, though optimal fits tend to shift slightly downward (not surprising in view of Observation 1). For all subjects, the baseline model with the highest optimal BF is unchanged from our earlier analysis; only for two subjects (2 and 6) does an extra model become undominated (Luce with CRRA preferences for subject 2, and RUM for subject 6). Graphs in the appendix also show that qualitative results for the one-size-fits-most analysis are consistent when changing the margin of error into the model-reliance measure.

7.2 Cash II

The experiment of Regenwetter et al. (2011) included another treatment tested on the same subjects: a second set of five monetary gambles listed in Table 3.[35] At first sight, these gambles are similar to those tested in the main 'Cash I' treatment (the Tversky (1969) lotteries adjusted for inflation). There is, however, a noteworthy difference: all lotteries in the 'Cash-II' treatment have an equal expected value of \$8.80. This may substantially impact behavior if subjects tend to use expected values to set their choices.

We repeat the fit-optimized Bayesian model selection exercise for the Cash-II data. Table 4 provides optimal fits and associated Bayes factors; the related graphs are provided in Section D of the Appendix. We observe that for seven subjects (2, 3, 5, 8, 10, 11 and 14), CRRA is the unique undominated model in both Cash I and Cash II. Also, Luce with CRRA

[35]Their paper also includes other questions for five gambles over non-monetary prizes. Given our interest in the performance of standard parametric specifications (like CRRA preferences or SCRUM with CRRA preferences), we focused our effort on lotteries with monetary prizes.

Lottery	a'	b'	c'	d'	e'
Payoff	\$31.43	\$27.50	\$24.44	\$22.00	\$20.00
Probability	0.28	0.32	0.36	0.40	0.44

Table 3: Another set of lotteries tested by Regenwetter et al. (2011), called Cash II. The information is presented in the same format as in Table 1.

	CRRA	RAT	DET	CLUCE	LUCE	SCRUM	RUM
Subject 1	2.4 (27%)	1.2 (23%)	1.0 (50%)	2.6 (69%)	1.5 (68%)	**3.6** (72%)	1.9 (90%)
Subject 2	$\mathbf{10^5}$ (89%)	10^4 (89%)	10^3 (89%)	10^4 (99%)	10^4 (99%)	263.4 (90%)	11.3 (99%)
Subject 3	$\mathbf{10^5}$ (88%)	10^4 (88%)	10^3 (88%)	10^5 (99%)	10^4 (99%)	862.4 (93%)	15.8 (100%)
Subject 4	**17.8** (47%)	3.9 (50%)	1 (50%)	5.5 (74%)	3.4 (77%)	16.1 (81%)	15.4 (98%)
Subject 5	$\mathbf{10^4}$ (82%)	10^3 (82%)	688.5 (82%)	429.5 (92%)	54.6 (92%)	10^3 (93%)	10.3 (100%)
Subject 6	6.0 (35%)	1.8 (36%)	1.0 (50%)	10.1 (78%)	3.7 (78%)	**46.3** (88%)	4.8 (98%)
Subject 7	3.7 (31%)	6.8 (55%)	1.0 (53%)	82.4 (96%)	**201.8** (99%)	7.1 (76%)	15.5 (100%)
Subject 8	$\mathbf{10^7}$ (94%)	10^6 (94%)	10^5 (94%)	10^5 (99%)	10^4 (99%)	10^3 (96%)	9.7 (99%)
Subject 9	10.3 (44%)	8.7 (56%)	1.1 (55%)	**875.5** (98%)	774.8 (98%)	14.6 (81%)	19.8 (100%)
Subject 10	$\mathbf{10^5}$ (88%)	10^4 (88%)	10^3 (88%)	449.6 (92%)	57.2 (92%)	730.7 (93%)	6.3 (97%)
Subject 11	$\mathbf{10^7}$ (94%)	10^6 (94%)	10^5 (94%)	10^3 (98%)	10^3 (99%)	10^3 (94%)	7.2 (98%)
Subject 12	2.8 (25%)	3.2 (46%)	1.0 (50%)	4.8 (74%)	8.2 (84%)	5.7 (74%)	**8.4** (100%)
Subject 13	6.2 (34%)	2.5 (41%)	1.0 (50%)	22.5 (84%)	18.5 (90%)	**23.4** (82%)	12.9 (100%)
Subject 14	$\mathbf{10^{11}}$ (99%)	10^{10} (99%)	10^9 (99%)	10^5 (99%)	10^4 (99%)	10^4 (96%)	7.4 (97%)
Subject 15	9.8 (42%)	4.9 (50%)	1.0 (50%)	**467.5** (95%)	117.5 (95%)	10.6 (78%)	17.0 (100%)
Subject 16	1.7 (18%)	1.2 (24%)	1.1 (54%)	1.5 (62%)	1.1 (62%)	**2.2** (66%)	1.2 (82%)
Subject 17	**706.6** (68%)	64.8 (68%)	7.6 (68%)	142.0 (92%)	19.1 (90%)	627.2 (93%)	12.1 (100%)
Subject 18	2.3 (22%)	1.4 (29%)	1.0 (50%)	**4.6** (75%)	2.9 (76%)	4.6 (72%)	3.3 (96%)

Table 4: The optimal Bayes factors for CASH II, with the corresponding fit underset in parenthesis, for each subject and model. For brevity, we simply state the order of magnitude for Bayes factors greater than 1000.

preferences is the unique undominated in both Cash I and Cash II for subject 15. Hence the same baseline model is robustly singled out for 44.4% of the subjects.

Some differences, but no strong disagreement, occur for another eight subjects when

comparing Cash I and II. For two subjects (9 and 13), Luce with CRRA preferences is undominated in both Cash I and Cash II, but now accompanied by additional undominated models in Cash II. In both Cash I and II, all models perform rather poorly at capturing Subject 12's choices (with all Bayes factors below 10). There are also five subjects (1, 4, 16, 17, 18) for whom successful models can be identified in one treatment (optimal Bayes factor above 10), but choices do not allow to clearly identify a successful model for the other set of prizes (all Bayes factors below 10).

The clearest sign of behavioral changes occur for only two subjects: 6 and 7. For the latter, an optimal BF of order 10^5 (84% fit) is achieved by CRRA in Cash I, while Luce achieves only 38.2. By contrast, an optimal BF of 201.8 is achieved by Luce (99% fit) in Cash 2, against a mere 3.7 for CRRA. For subject 6, the undominated model switches from Luce with CRRA preferences in Cash I to SCRUM in Cash II.

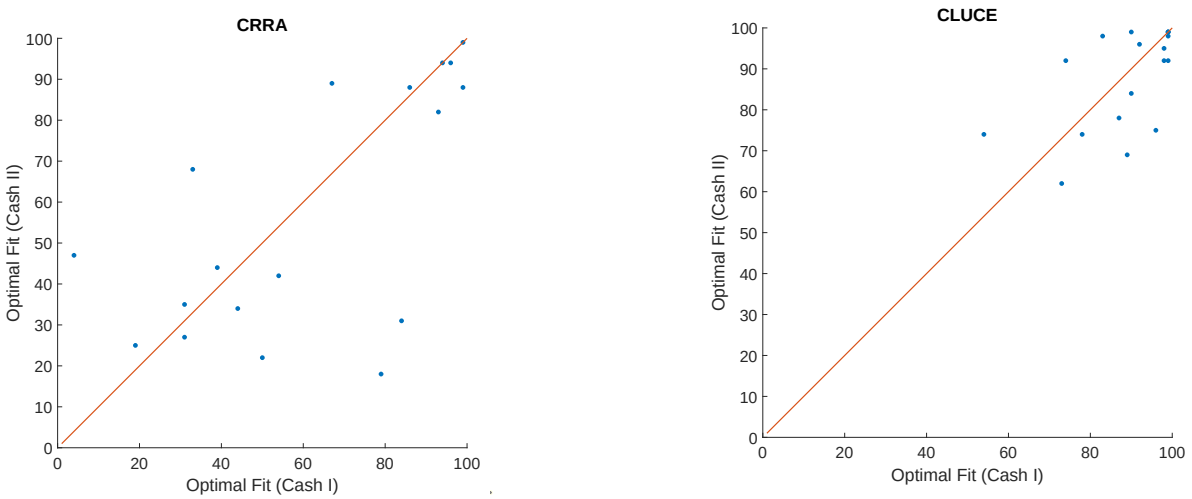


Figure 6: Optimal Fit Correlation between Cash I and Cash II

Thus we find a good degree of coherence in the analysis across different prizes, but also some evidence that the elimination of a simple choice criterion can impact behavior. To provide another illustration, we plot in Figure 6 optimal fits for a model in Cash II, against optimal fits for that same model in Cash I. For brevity, we focus on the two models that are most often undominated: CRRA and Luce with CRRA preferences. There is a substantial degree of correlation in the fits across treatment: the Pearson correlation coefficients are 0.6447 for CRRA (p-value 0.0039) and 0.5704 for Luce with CRRA preferences (p-value 0.0134). We do spot, however, some outliers as well. For instance, CRRA's optimal fit for subject 7 (discussed at the end of the previous paragraph) decreases from 84% in Cash I to 31% in Cash 2.

Graphs in the appendix also show that qualitative results for the one-size-fits-most analysis are consistent across Cash 1 and Cash 2.

Of course, one may apply our methodology to each subject's pooled set of choices from Cash I and Cash II. While this prevents analyzing the similarities and differences in behavior across the two treatments, it may provide some insight into which model(s) describe a wider variety of choices better. There are some computational challenges, however, as probabilities become extremely small. For an example, the probability that an SCF uniformly drawn at random is consistent with SCRUM (with CRRA preferences) is $1/(10!) \sim 2.76 \times 10^{-7}$ when considering Cash 1 or Cash 2 in isolation, and shrinks to $1/(20!) \sim 4.11 \times 10^{-19}$ when considering pooled choices. Only extremely large Bayes factors may then be detectable numerically.[36] Though we have a formula to compute the volume of SCRUM, no such formula exists for expanded versions of SCRUM, or for other models. There are ways to fine tune the numerical methods used to assess those numbers,[37] but deriving all Bayes factors with high precision would be very time consuming. The preliminary analysis we conducted to obtain (rough) Bayes Factors for the pooled data[38] broadly support our understanding of Cash I and Cash II. For instance, as expected, CRRA still appears as the unique winning model for the group of seven subjects (2, 3, 5, 8, 10, 11 and 14) for which this model was uniquely optimal in both Cash I and Cash 2.[39] Similarly, Luce with CRRA preferences remains the unique optimal model when pooling Cash I and Cash II. Subjects 9 and 13 had Luce with CRRA preferences as an optimal model in both Cash I and Cash II, but only uniquely so in Cash I. The model appears uniquely optimal for the pooled data when it comes to Subject 9. However, SCRUM also remains a contender for Subject 13. For three of the remaining subjects, Luce with CRRA preferences appears as a clear winner (with optimal fits above 90%). SCRUM is clearly selected for subject 17 with the pooled data, while it was only one of the optimal models in Cash II.

[36] Say one tests 100 billions SCFs and that one trusts the frequency is close enough to the true probability of being consistent with the model when one gets at least 100 hits to the model. Then the minimal Bayes factor one could identify is of the order 10^{10}.

[37] For instance, no question involves both a Cash I and a Cash II lottery. Hence, the volume of a theory like Luce with fit f when pooling the two sets of questions is equal to the square of the volume of that same theory over Cash I alone. But subtler methods must be deployed for models involving CRRA preferences: the set of preferences (21 of them) over the 20 binary problems obtained when pooling Cash I and Cash II is not the cartesian product of CRRA preferences (11 of them) for Cash I and Cash II separately.

[38] Available upon request.

[39] Here, for instance, it can be very hard to figure out the optimal fit when it is high as the models get extremely small. The lower bounds we get by computing the Bayes Factors for fits above 90% clearly suggest, however, that CRRA is uniquely selected.

References

Afriat, Sydney (1973), On a System of Inequalities in Demand Analysis: An Extension of the Classical Method, *International Economic Review*, 14(2), 460–472.

Apesteguia, Jose and Miguel Ballester (2021), Separating Predicted Randomness from Residual Behavior, *Journal of the European Economic Association*, 19(2), 1041–1076.

Apesteguia, Jose, Miguel Ballester, and Jay Lu (2017), Single-Crossing Random Utility Models, *Econometrica*, 85(2), 661–674.

Balakrishnan, Narayanaswamy, Efe Ok and Pietro Ortoleva (2022), Inferential Choice Theory, *mimeo*.

Bronars, Stephen G. (1987), The Power of Nonparametric Tests of Preference Maximization , *Econometrica*, 55(3), 693–698.

Cattaneo, Matias, Xinwei Ma, Yusufcan Masatlioglu, and Elchin Suleymanov (2020), A Random Attention Model, *Journal of Political Economy*, 128(7), 2796–2836.

Cavagnaro, Daniel and Clintin Davis-Stober (2014), Transitive in our preferences, but transitive in different ways: An analysis of choice variability, *Decision*, 1(2), 102–122.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman (2014), Who is (More) Rational?, *American Economic Review*, 104(6), 1518–1550.

Cohen, Michael, and Jean-Claude Falmagne (1990), Random Utility Representation of Binary Choice Probabilities: A New Class of Necessary Conditions, *Journal of Mathematical Psychology*, 34(1), 88–94.

Conte, Anna, John Hey and Peter Moffatt (2011). Mixture models of choice under risk, *Journal of Econometrics*, 162(1), 79–88.

Dardanoni, Valentino, Paola Manzini, Marco Mariotti and Christopher J. Tyson (2020), Inferring Cognitive Heterogeneity from Aggregate Choices, *Econometrica*, 88(3), 1269–1296.

de Clippel, Geoffroy and Kareen Rozen (2022), Bounded Rationality in Choice Theory: A Survey, *mimeo*.

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

Fudenberg, Drew, Wayne Gao and Annie Liang (2021), How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories, *mimeo*.

Greene, William H. (2018). *Econometric Analysis* (8th edition), Pearson.

Gul, Faruk, Paulo Natenzon and Wolfgang Pesendorfer (2014), Random Choice as Behavioral Optimization, *Econometrica*, 82(5), 1873–1912.

Gul, Faruk and Wolfgang Pesendorfer (2006), Random Expected Utility, *Econometrica*, 74(1), 121–146.

Halevy, Yoram, Dotan Persitz and Lanny Zrill (2018), Parametric Recoverability of Preferences, *Journal of Political Economy,* 126(4), 1558–1593.

Harless, D.W. and Colin Camerer (1994), The predictive power of Generalized Expected Utility Theories, *Econometrica*, 62, 1251–1289.

Harrison, Glenn and E. Elisabet Rutström (2009), Expected utility theory and Prospect theory: One Wedding and a Decent Funeral, *Experimental Economics*, 12(2), 133–158.

Hey, John (1995), Experimental Investigations of Errors in Decision Making Under Risk, *European Economic Review*, 39, 633–640.

Hey, John and Chris Orme (1994), Investigating Generalizations of the Expected Utility Theory Using Experimental Data, *Econometrica*, 62(6), 1291–1326.

Jeffreys, H. (1961). *Theory of Probability*, 3rd edition, Oxford, U.K.: Oxford University Press.

Iverson, G. and J.C. Falmagne (1985), Statistical Issues in Measurement, *Mathematical Social Sciences*, 10, 131–153.

Kalai, Gil, Ariel Rubinstein, and Ran Spiegler (2002), Rationalizing Choice Functions by Multiple Rationales, *Econometrica*, 70(6), 2481–2488.

Kass, Robert E., and Adrian E. Raftery (1995), Bayes Factors, *Journal of the American Statistical Association*, 90(430), 773–795.

Liang, Annie (2019), Inference of Preference Heterogeneity from Choice Data, *Journal of Economic Theory*, 179, 275–311.

Luce, R. Duncan (1959), *Individual choice behavior: A theoretical analysis.* John Wiley, New York.

Manzini, Paola, and Marco Mariotti (2006), Revealed Preferences and Boundedly Rational Choice Procedures: an Experiment, *IZA Discussion Paper No. 2341.*

Manzini, Paola, and Marco Mariotti (2007), Sequentially Rationalizable Choice, *American Economic Review*, 97(5), 1824–1839.

Manzini, Paola, and Marco Mariotti (2014), Stochastic Choice and Consideration Sets, *Econometrica*, 3, 1153–1176.

Manzini, Paola, and Marco Mariotti (2018), Dual Random Utility Maximisation, *Journal of Economic Theory*, 177, 162–182.

Masatlioglu, Yusufcan, Daisuke Nakajima and Erkut Ozbay (2012), Revealed Attention, *American Economic Review*, 102(5), 2183–2205.

McCausland, William, Clintin Davis-Stober, A.A.J. Marley, Sanghyuk Park and Nicholas Brown (2020), Testing the Random Utility Hypothesis Directly, *The Economic Journal*, 130, 183–207.

Polisson, Matthew, John K.H. Quah, and Ludovic Renou (2020), Revealed Preferences over Risk and Uncertainty, *American Economic Review*, 6(110), 1782–1820.

Regenwetter, Michel, Jason Dana and Clintin Davis-Stober (2011), Transitivity of Preferences, *Psychological Review*, 118(1), 42–56.

Selten, Richard (1991), Properties of a Predictive Measure of Success, *Mathematical Social Sciences*, 21, 153–167.

Selten, Richard and S. Krischker (1983), Comparison of two theories for characteristic function experiments. *Aspiration Levels in Bargaining and Economic Decision Making* (edited by R. Tietz), Springer, Berlin and Heidelberg, 259–264.

Tversky, Amos (1969), "Intransitivity of Preferences," *Psychological Review*, 76(1), 31–48.

Tversky, Amos (1972), "Elimination by Aspects: A Theory of Choice," *Psychological Review*, 79, 281–299.

Appendix

This Appendix is divided as follows. Appendix A details the computational procedures. Appendix B contains additional tables and graphs for Sections 4 and 5. Appendix C provides graphs for the Apesteguia and Ballester (2021) measure discussed in Section 7.1. Finally, Appendix D contains the graphs for Section 7.2 corresponding to the CASH-II treatment.

A Computational Procedures

We begin by describing some conventions used throughout. Cash I involves 10 binary menus defined over defined over the set $X = \{a, b, c, d, e\}$.[40] We follow the numbering convention that the first menu is $\{a, b\}$, the second one is $\{a, c\}$, and then $\{a, d\}$, $\{a, e\}$, $\{b, c\}$, $\{b, d\}$, $\{b, e\}$, $\{c, d\}$, $\{c, e\}$ and $\{d, e\}$. An SCF p belong to $[0, 1]^{10}$, with p_i representing the probability of picking the first element in menu $i = 1, \ldots, 10$.

A.1 Deterministic models

An SCF m is deterministic if it belongs to $\{0, 1\}^{10}$. The best fit a deterministic model provides for an SCF p can be computed exactly, as it is the maximal γ for which one can find $m \in \mathcal{M}$ such that $-(1-\gamma) \leq p - m \leq 1 - \gamma$. When $\mathcal{M}$ contains all deterministic choice functions, the optimum is reached with $m_i = 1$ if, and only if, $p_i > 0.5$. The optimal fit for p is then $\min_{i=1}^{10} \max\{p_i, 1 - p_i\}$. The goodness of fit when comparing p to a specific deterministic SCF m is $\min_{i=1}^{10} m_i p_i + (1-m_i)(1-p_i)$. Thus, to find the best fit that rationality or the maximization of a crra preferences provides for p, one just has to maximizes this last expression over all m's in the model. Under rationality, there is an isomorphism between the set of rational choice functions over the 10 binary menus and the set of preference orderings over the five options. Thus we get all rational choice patterns by considering the $5! = 120$ possible preference relations. There are 11 choice functions generated by the maximization of a preference relation. To see this, note that the lottery with the larger monetary prize is picked in all binary menus when one is sufficiently risk loving. One can easily check that, as the parameter of risk aversion decreases, the choice in a menu will at some point switch to the other lottery (the one with the larger winning probability is picked) and never switches back (single crossing property). Since there are ten menus (and we don't face a degenerate case where indifference occurs in two different binary menus for a same CRRA parameter), one can indeed generate this way 11 distinct choice functions. Each row of the following

[40] Though we focus on Cash I, similar ideas apply to Cash II and to the union of Cash I and Cash II.

matrix represents a possible vector of choice probabilities p for CASH I as one varies the CRRA coefficient:

$$M_{CRRA-I} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4}$$

The analogous matrix for CASH II is:

$$M_{CRRA-II} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{5}$$

A.2 Stochastic models

By definition, the model $\mathcal{M}$ provides a fit γ to the SCF $p \in [0,1]^{10}$ if, and only if, there exists $m \in \mathcal{M}$ such that

$$-(1-\gamma) \leq p - m \leq (1-\gamma). \tag{6}$$

By Cohen and Falmagne (1990), m is compatible with RUM if, and only if, the following inequalities (whenever $|X| \leq 5$):

$$m(x|xy) + m(y|yz) - m(x|xz) \leq 1,$$

for all triplets x, y, z of distinct elements in X. There are six possible orderings of a triplet x, y, z, and hence six inequalities associated to the unordered set $\{x, y, z\}$. It is easy to check, however, that only two of these six inequalities are distinct while the other four are repeats. In our case, $|X| = 5$. There are thus ten subsets of cardinality three, and hence twenty inequalities that choice probabilities must satisfy for RUM compatibility. Labeling a, b, c, d and e as 1, 2, 3, 4 and 5, they are:

$$0 \leq p(i|ij) + p(j|jk) - p(i|ik) \leq 1, \tag{7}$$

for all $1 \leq i < j < k \leq 5$. Combined with the observation in the previous paragraph, we see that whether or not RUM provides a fit γ to the SCF $p \in [0,1]^{10}$ amounts to checking the feasibility of a system of linear inequalities, which is easily testable numerically.

Consider next SCRUM for CRRA choice functions. An SCF m is a mixture of the 11 Cash-I CRRA choice functions presented in the previous subsection if, and only if, there exists $\alpha \in \Delta_{11}$ such that $\alpha \cdot M_{CRRA_I} = p$. This last condition is equivalent to $\alpha_1 = m_1$, $\alpha_1 + \alpha_2 = m_2$ or $\alpha_2 = m_2 - m_1$, and so on so forth with $\alpha_i = m_i - m_{i-1}$ for all $i \neq 1$. Given that each α_i must be nonnegative, we have that m is a mixture of the rows in M_{CRRA_I} only if $m_{i+1} \geq m_i$ for all $i \neq 10$. But the converse holds as well, simply by picking $\alpha_1 = m_1$, $\alpha_i = m_i - m_{i-1}$ for $i = 2, \ldots, 10$ (automatically ≤ 1; ≥ 0 by assumption), and $\alpha_{11} = 1 - m_1$ (to make sure the α_i's sum up to 1). To summarize, the SCF m is compatible with SCRUM over CRRA preferences if, and only if, $m_{i+1} \geq m_i$ for all $i \neq 10$. (Similar arguments apply in Cash II, and m is compatible with SCRUM over CRRA preferences in Cash II if, and only if, $m_8 \leq m_9 \leq m_6 \leq m_7 \leq m_{10} \leq m_4 \leq m_3 \leq m_1 \leq m_2 \leq m_5$.) Hence, as for RUM, we see that whether or not SCRUM provides a fit γ to the SCF $p \in [0,1]^{10}$ amounts to checking the feasibility of a system of linear inequalities, which can easily testable numerically.

An SCF m is a Luce rule if there exists $u : X \to \mathbb{R}^5_{++}$ such that the probability of picking x from a pair $\{x, y\}$ equals $u(x)/(u(x) + u(y))$ for all $x, y \in X$.[41] Equation 6 can then be rewritten as:

$$-(1-\gamma)(u(x) + u(y)) \leq p(x|xy)(u(x) + u(y)) + u(x) \leq (1+\gamma)(u(x) + u(y)),$$

for all distinct x, y in X. Thus, whether or not the Luce model provides a fit γ to the SCF $p \in [0,1]^{10}$ amounts to checking the feasibility of a system of linear inequalities in the five

[41] From a numerical perspective, we can overlook the constraint that u defines a strict preference. Suppose the γ-fit inequalities are satisfied only by some m that is a Luce rule only if indifferences are allowed. It is easy to check then that the γ'-fit inequalities are satisfied by a Luce rule without indifference for any $\gamma' < \gamma$ (as close as desired to γ).

variables $(u(x))_{x\in X}$, which can easily testable numerically.[42] The same idea can be used to test whether or not the Luce model with respect a specific ranking $\succeq$ of X provides a fit γ to the SCF $p \in [0,1]^{10}$, simply by adding the linear inequalities $u(x) \geq u(y)$ for each x, y such that $x \succeq y$. Then testing whether or not C-Luce model provides a fit γ to an SCF amounts to checking whether or not one the Luce model with respect one of 11 CRRA choice functions provides a fit γ (having to test the feasibility of 11 systems of linear inequalities in the worst-case scenario).

A.3 Monte Carlo simulation

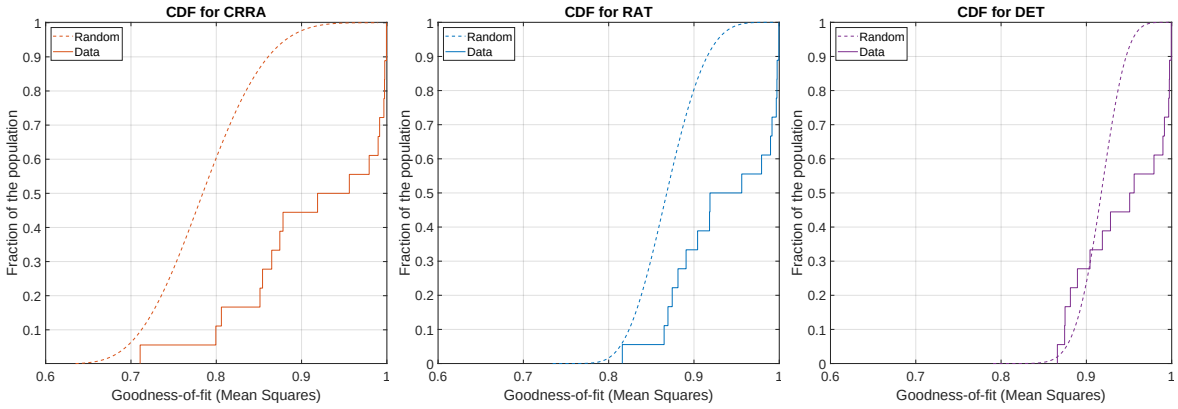
We use Monte Carlo simulation to determine the fraction of randomly drawn SCFs that belong to M^γ, with the set of SCFs drawn according to the uniform prior for the denominators but drawn according to a subject's updated beta distribution for the numerators. For the deterministic models, the speed of computations permits drawing 1 billion SCFs for both numerators and denominators at all levels of fit. For the more computationally-intensive stochastic choice models, we begin with a smaller draw of SCFs. To ensure robustness of our results, we aim to avoid, whenever computationally feasible, that the probability in denominators is determined by fewer than 1000 hits to the model: once this occurs, we scale up the number of draws by an order of magnitude and obtain a more precise computation. For RUM, SCRUM and Luce we begin by drawing 10 million SCFs for the denominator and numerators. For RUM, no further refinements are needed, as the threshold number of 1000 hits is achieved at all fits. For SCRUM, further refinements are needed starting at $\gamma = 94\%$, at which point we increase the draw to 1 billion SCFs and then achieve the threshold for all fits. For Luce, we increase the draw to 1 billion SCFs to achieve at least 1000 hits for $\gamma = 95\%, 96\%, 97\%$, and then increase the draw to 100 billion SCFs for $\gamma = 98\%, 99\%$ (at 99%, we can reach only 545 hits given the rareness of the model). Our most computationally-intensive model is Luce with CRRA preferences, and we begin by drawing 1 million SCFs for the numerators and 10 million SCFs for the denominator (we use fewer for numerators since they must be computed for each subject). For this model, we then increase to 1 billion draws for the denominator for $\gamma = 94\%, 95\%, 96\%, 97\%$, and then increase further to 100

[42] Some bounds must be imposed on these utilities for the numerical test. We require them to be at least one (this restriction alone is without loss, as Luce utilities can be renormalized to satisfy this) and at most 10^{13}. One can view this as limiting how close to deterministic choice functions the model can go. For instance, though rationality can be approximated as a sequence of Luce rules, the bounds limit how closeone can get. To get a sense of magnitudes, notice that the upper bound remains satisfied when normalizing the lowest utility to one, and having the utility function jump by a factor of $1,000$ each time one goes up one step in the ranking. One would get very close to the rational choice function this way, with a margin of error of roughly at most 10^{-3} in all binary menus. So, to the extent that goodness of fit cannot be computed exactly anyway, this provides a satisfactory approximation.

billion draws for $\gamma = 98\%, 99\%$ (at 99%, there are only 71 hits given the rareness of the model).

B Other tables and graphs for Sections 4 and 5

B.1 Robustness to mean-squared errors noted in Section 4

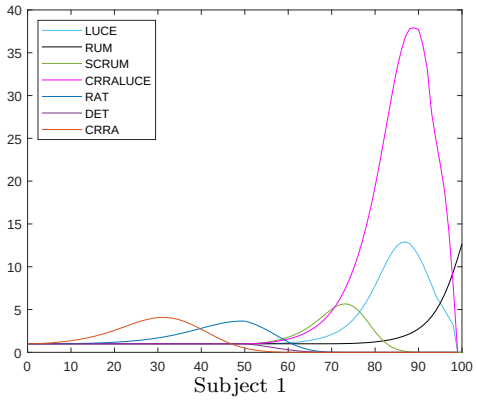


B.2 Table of optimal BFs and fits for Section 5

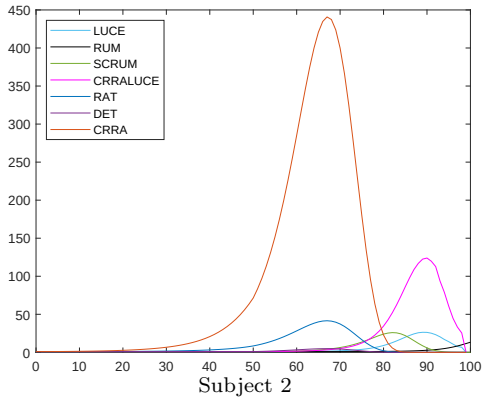
The optimal Bayes factor, with the corresponding fit underset in parenthesis, for each subject and model; with the largest optimal BF per row bolded. Scientific notation, with multiplier rounded to two decimal points, used for Bayes factors larger than 1000.

	CRRA	RAT	DET	CLUCE	LUCE	SCRUM	RUM
Subject 1	4.1 (31%)	3.7 (49%)	1.0 (50%)	**37.9** (89%)	12.9 (87%)	5.7 (73%)	12.7 (100%)
Subject 2	**440.8** (67%)	41.5 (67%)	4.9 (67%)	123.9 (90%)	26.4 (89%)	25.9 (82%)	13.3 (100%)
Subject 3	**9.09×10^{9}** (99%)	8.33×10^{8} (99%)	9.77×10^{7} (99%)	2.52×10^{5} (99%)	3.28×10^{4} (99%)	4.77×10^{3} (95%)	8.3 (98%)
Subject 4	1.0 (4%)	1.0 (18%)	**2.4** (64%)	1.0 (54%)	1.2 (64%)	1.0 (53%)	1.7 (88%)
Subject 5	**2.52×10^{7}** (93%)	2.31×10^{6} (93%)	2.71×10^{5} (93%)	6.48×10^{3} (99%)	844.0 (99%)	1.40×10^{3} (93%)	8.3 (98%)
Subject 6	3.4 (31%)	2.2 (42%)	1.0 (53%)	**21.5** (87%)	5.6 (83%)	1.9 (64%)	5.7 (100%)
Subject 7	**1.49×10^{5}** (84%)	1.37×10^{4} (84%)	1.60×10^{3} (84%)	292.4 (92%)	38.2 (91%)	590.7 (92%)	8.9 (99%)
Subject 8	**1.88×10^{8}** (96%)	1.73×10^{7} (96%)	2.02×10^{6} (96%)	2.71×10^{5} (99%)	3.53×10^{4} (99%)	1.43×10^{3} (94%)	8.6 (98%)
Subject 9	6.2 (39%)	3.3 (47%)	1.0 (50%)	**20.7** (83%)	7.5 (83%)	9.4 (79%)	9.3 (100%)
Subject 10	**3.33×10^{5}** (86%)	3.05×10^{4} (86%)	3.57×10^{3} (86%)	8.17×10^{3} (98%)	1.05×10^{3} (99%)	1.48×10^{3} (94%)	11.0 (100%)
Subject 11	**2.58×10^{7}** (94%)	2.36×10^{6} (94%)	2.77×10^{5} (94%)	4.73×10^{5} (99%)	6.16×10^{4} (99%)	1.05×10^{3} (93%)	9.8 (99%)
Subject 12	1.6 (19%)	1.9 (37%)	1.0 (50%)	3.9 (78%)	3.1 (77%)	1.6 (63%)	**4.6** (98%)
Subject 13	12.4 (44%)	3.2 (46%)	1.0 (50%)	**83.5** (90%)	20.7 (90%)	11.9 (79%)	12.9 (100%)
Subject 14	**9.09×10^{9}** (99%)	8.33×10^{8} (99%)	9.77×10^{7} (99%)	2.13×10^{5} (99%)	2.77×10^{4} (99%)	7.71×10^{3} (96%)	8.4 (98%)
Subject 15	58.4 (54%)	7.2 (53%)	1.0 (51%)	**1.47×10^{3}** (98%)	204.3 (98%)	57.3 (86%)	18.6 (100%)
Subject 16	**5.67×10^{3}** (79%)	519.3 (79%)	60.9 (79%)	4.6 (73%)	1.7 (71%)	47.1 (85%)	1.6 (87%)
Subject 17	4.8 (33%)	1.5 (31%)	1.0 (50%)	4.9 (74%)	2.2 (74%)	**6.4** (76%)	2.3 (92%)
Subject 18	35.3 (50%)	5.2 (50%)	1.0 (50%)	**990.6** (96%)	165.1 (99%)	51.5 (86%)	19.1 (100%)

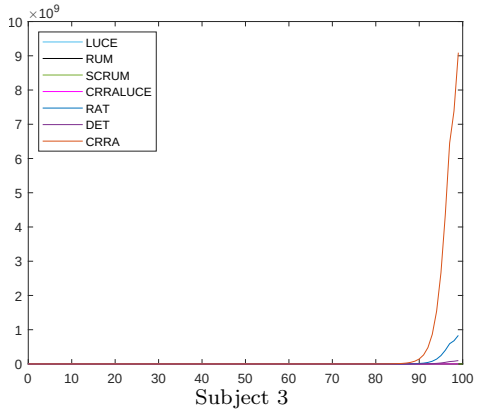
B.3 Plots of BF per model and subject for Section 5



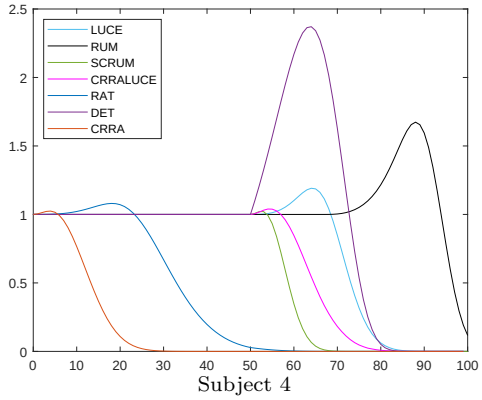
Subject 1



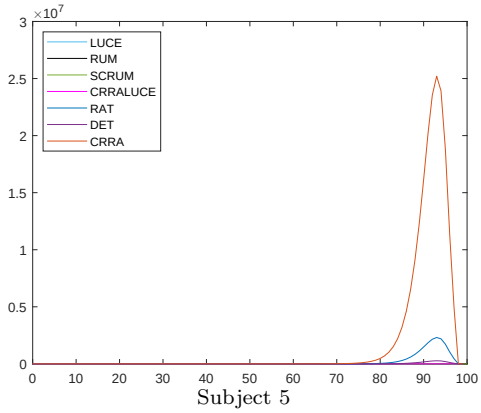
Subject 2



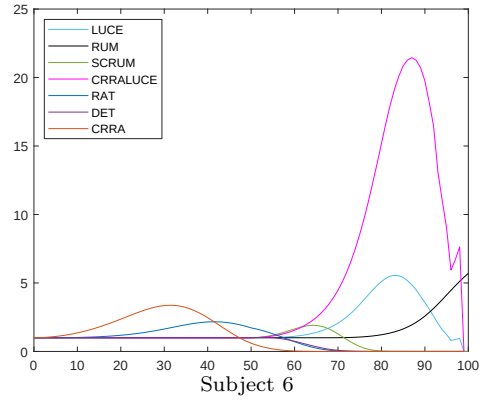
Subject 3



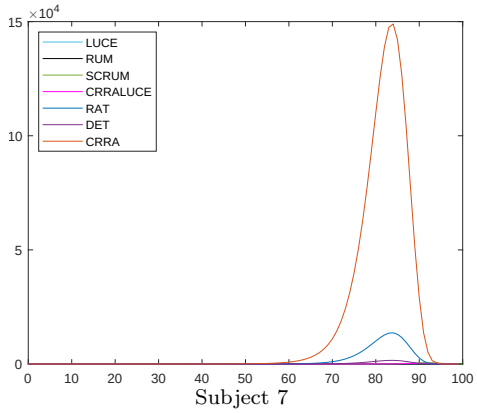
Subject 4



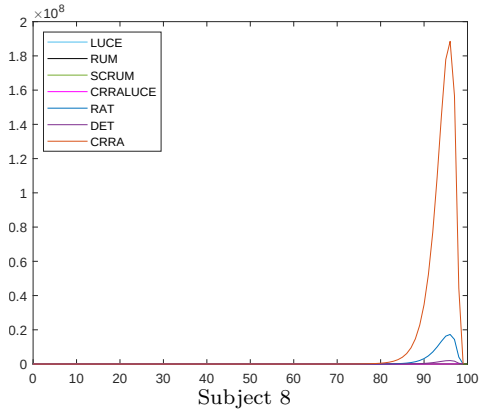
Subject 5



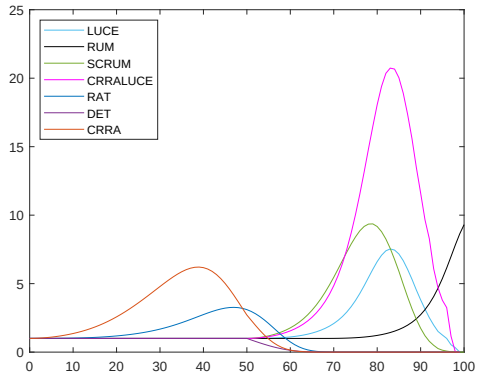
Subject 6



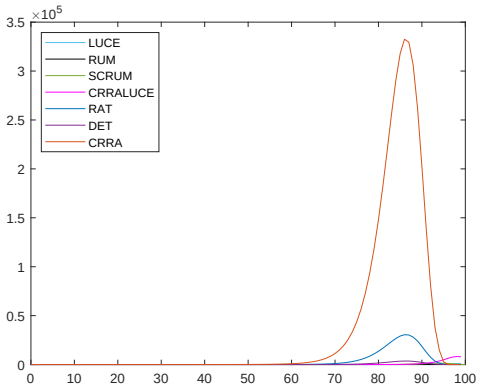
Subject 7



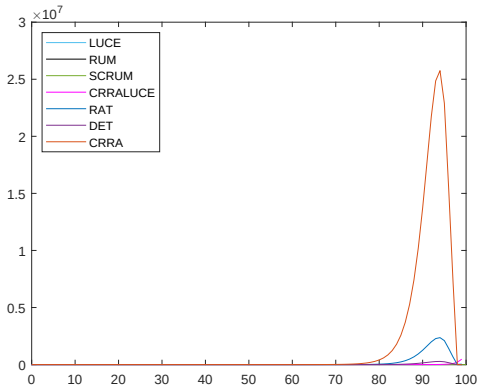
Subject 8



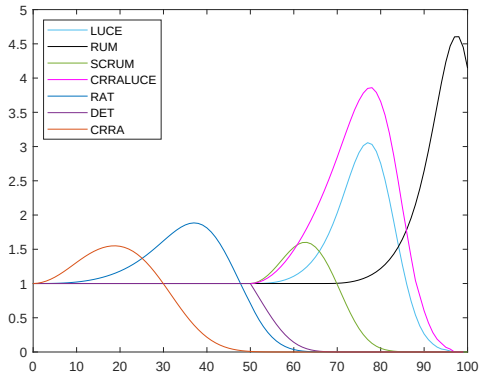
Subject 9



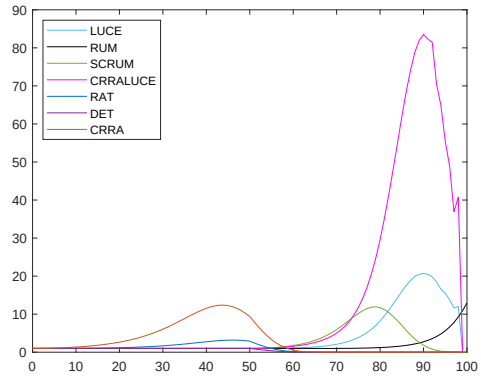
Subject 10



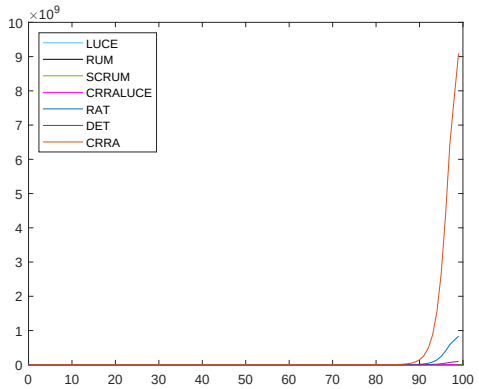
Subject 11



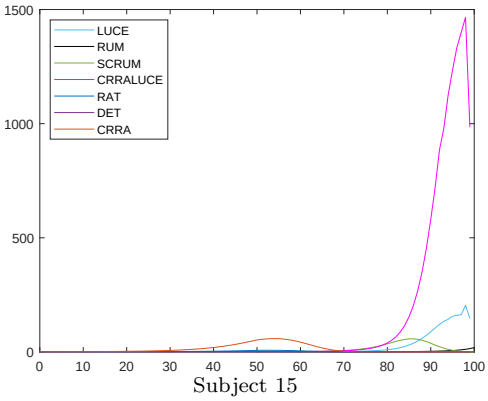
Subject 12



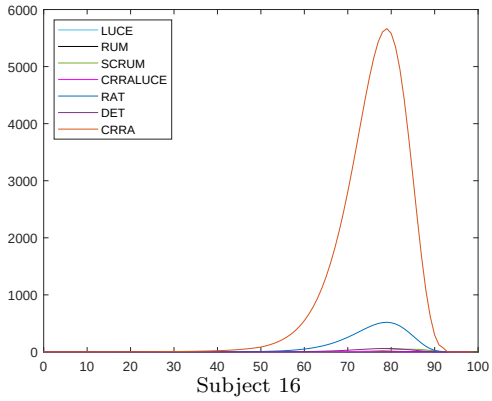
Subject 13



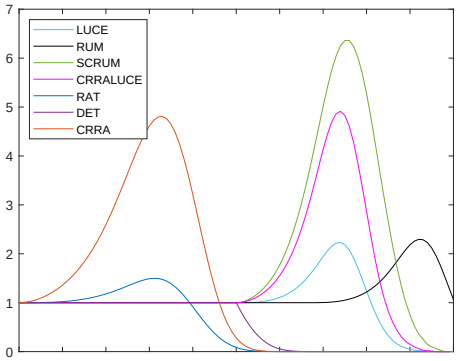
Subject 14



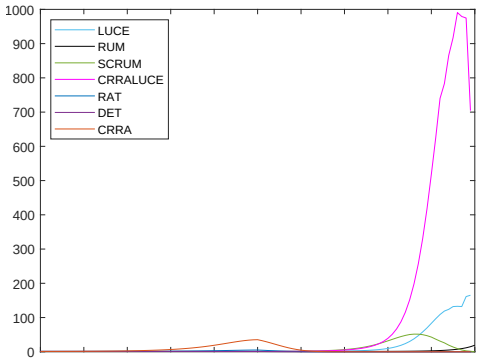
Subject 15



Subject 16



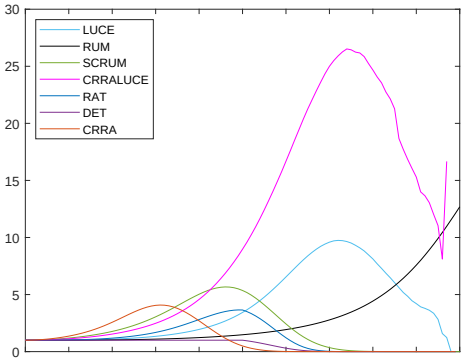
Subject 17



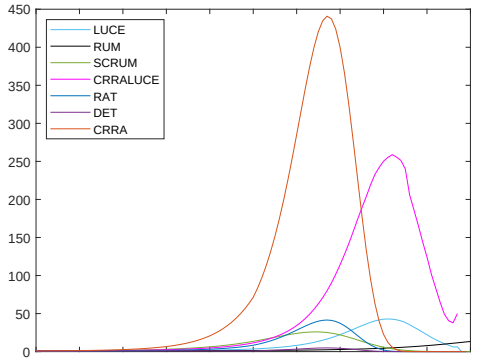
Subject 18

C Plots under model-reliance measure for Section 7.1

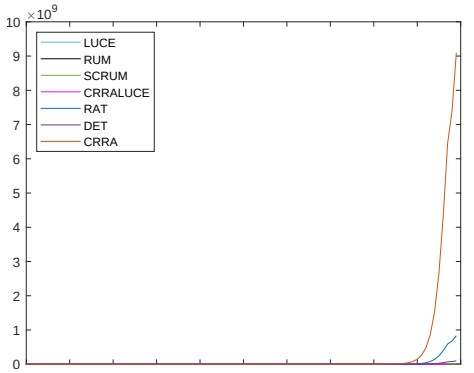
C.1 Subject by subject



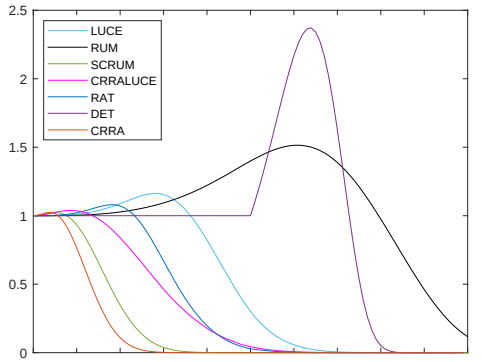
Subject 1



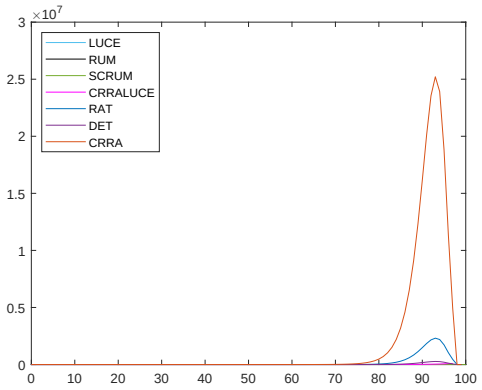
Subject 2



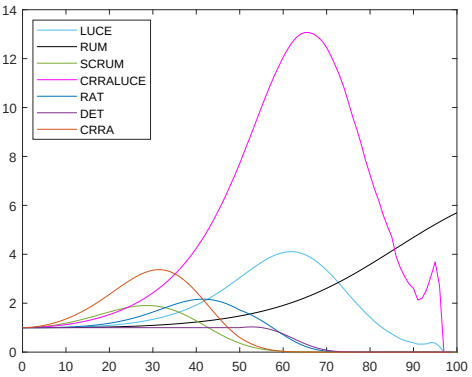
Subject 3



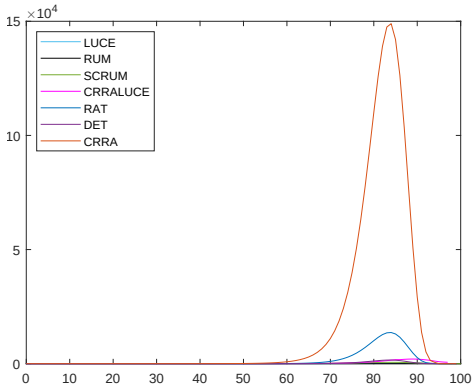
Subject 4



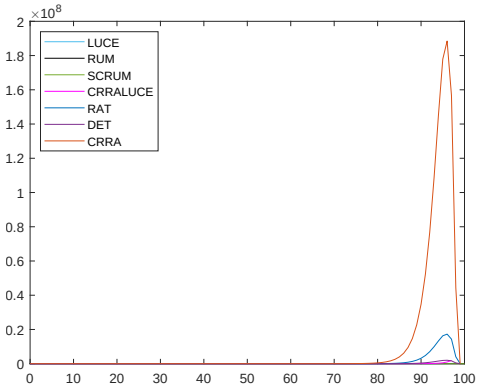
Subject 5



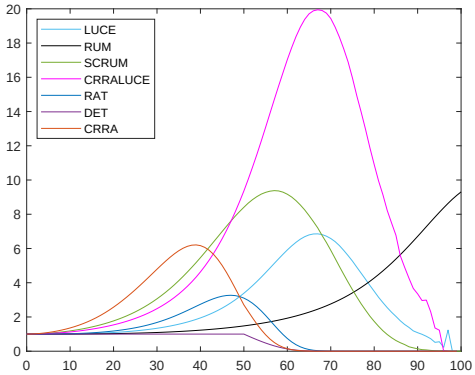
Subject 6



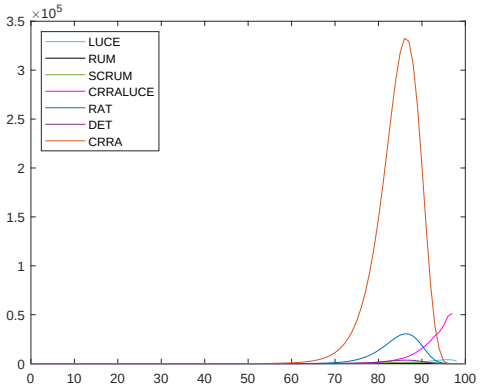
Subject 7



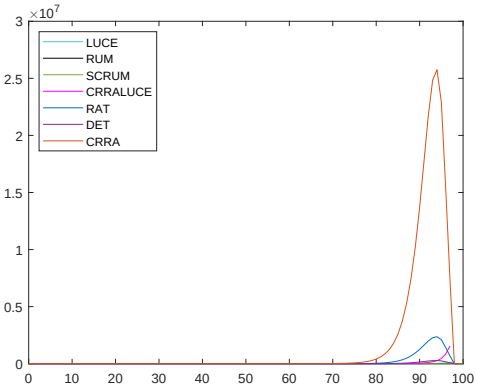
Subject 8



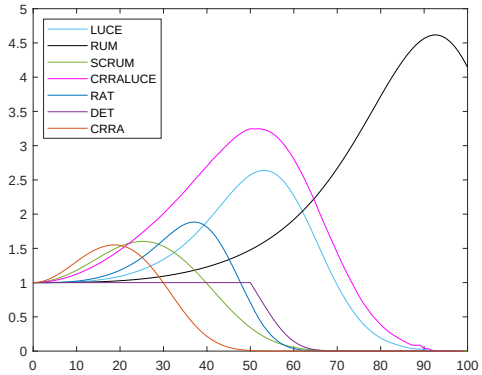
Subject 9



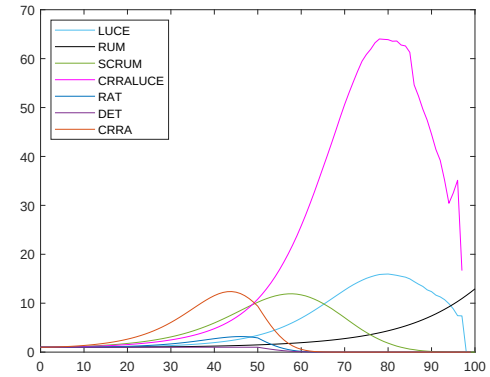
Subject 10



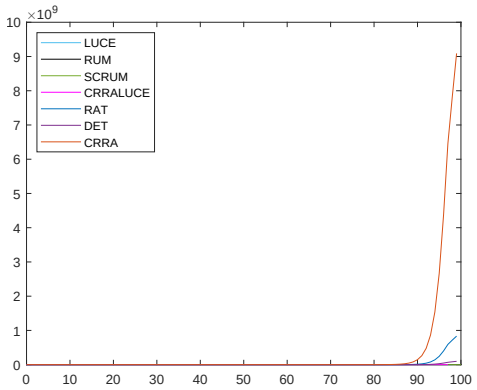
Subject 11



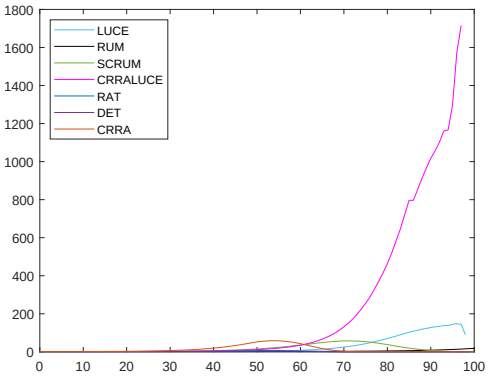
Subject 12



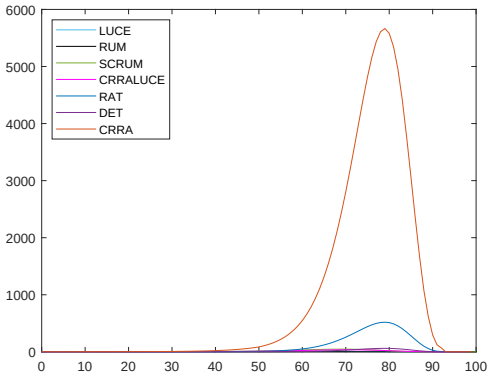
Subject 13



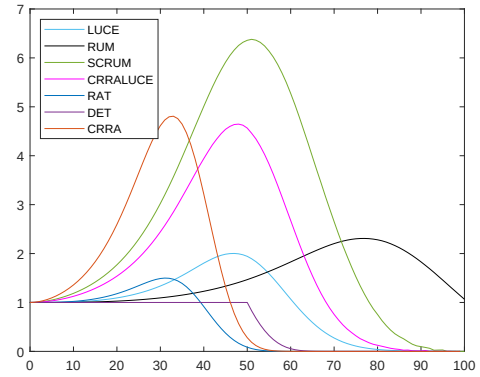
Subject 14



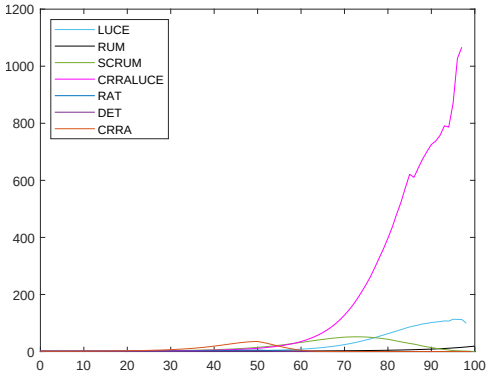
Subject 15



Subject 16

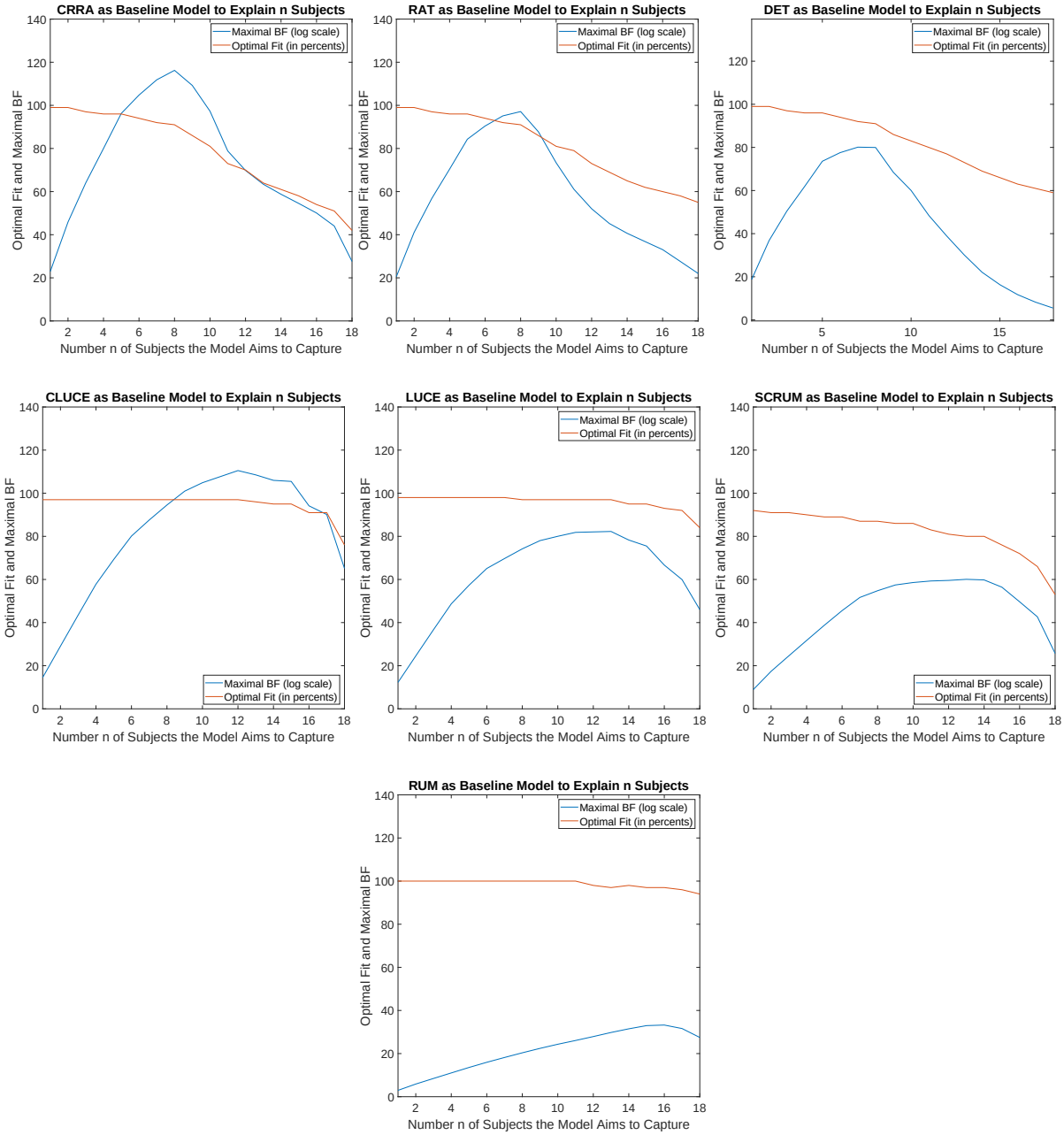


Subject 17



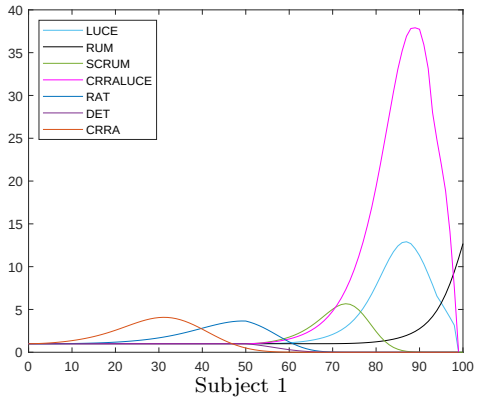
Subject 18

C.2 One size fits most

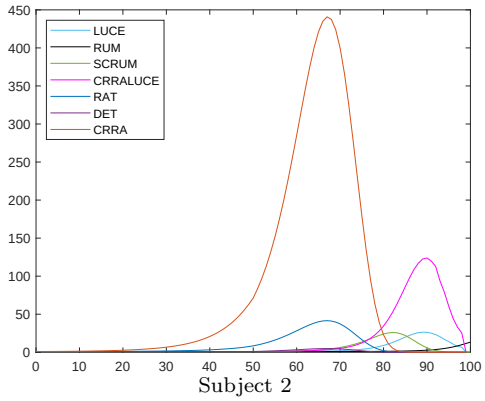


D CASH II plots for Section 7.2

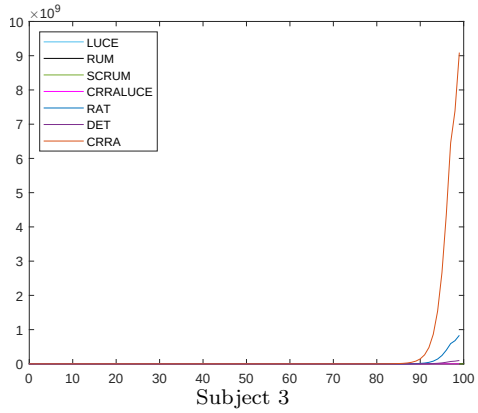
D.1 Subject by subject



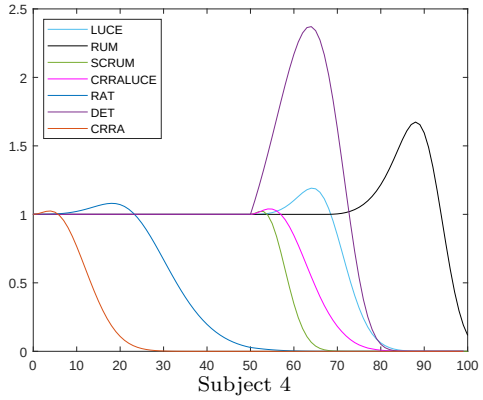
Subject 1



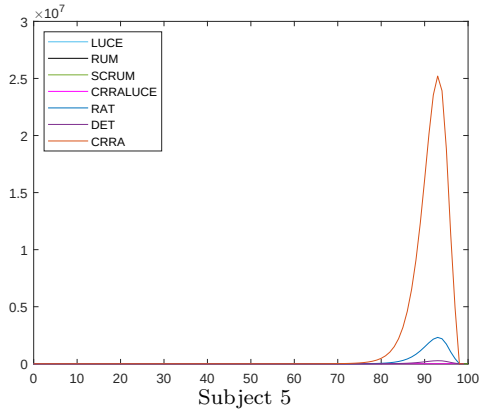
Subject 2



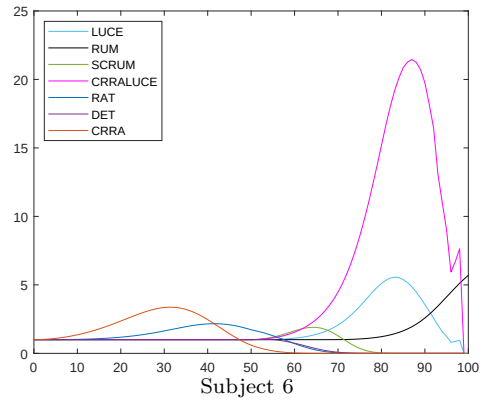
Subject 3



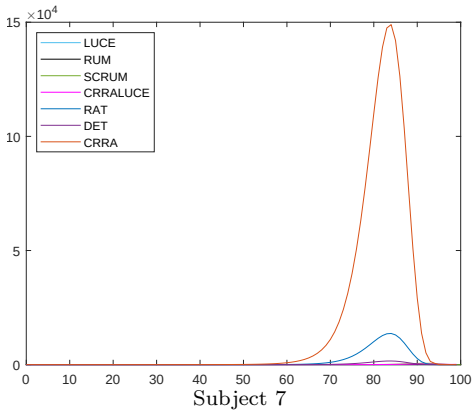
Subject 4



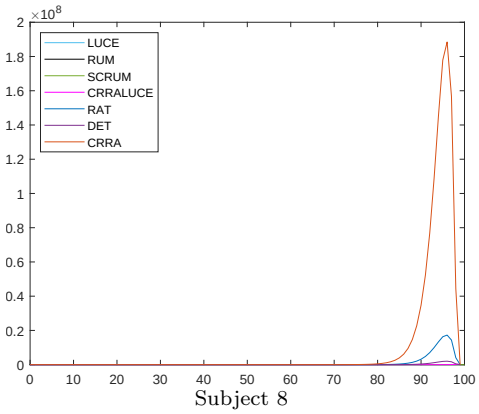
Subject 5



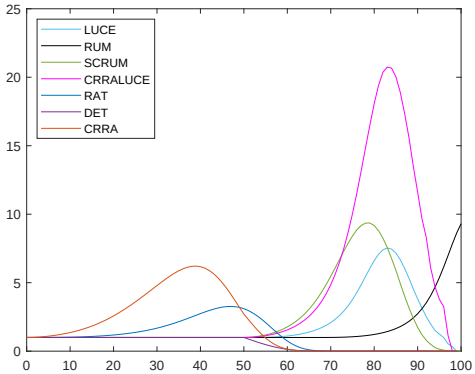
Subject 6



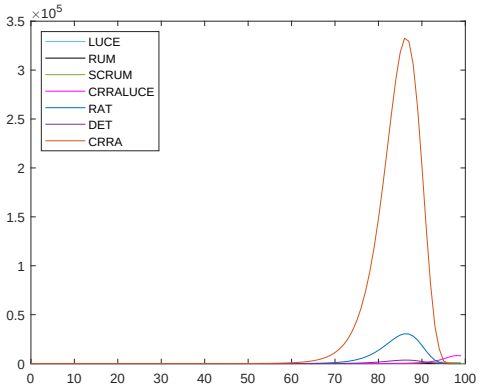
Subject 7



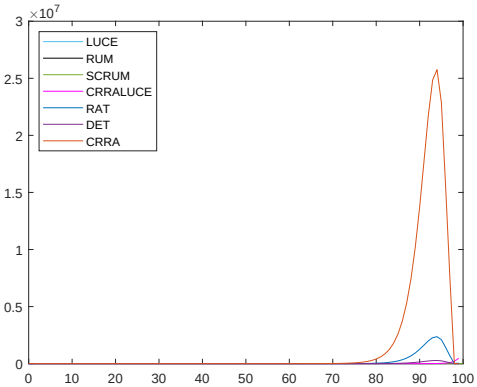
Subject 8



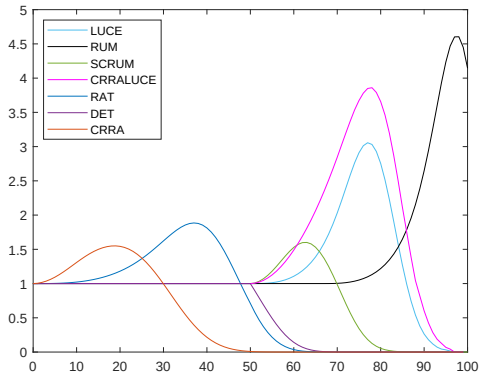
Subject 9



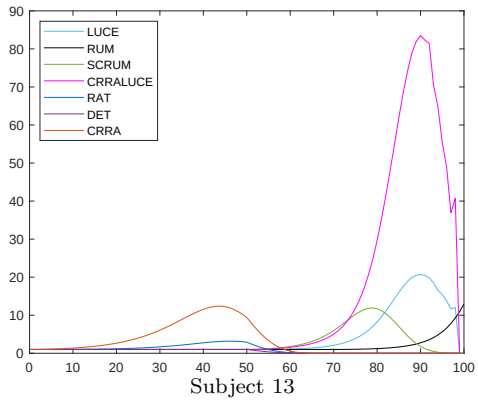
Subject 10



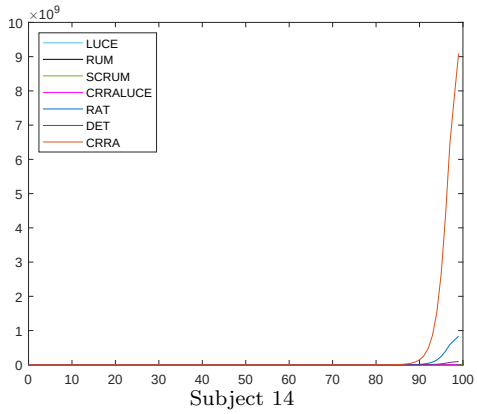
Subject 11



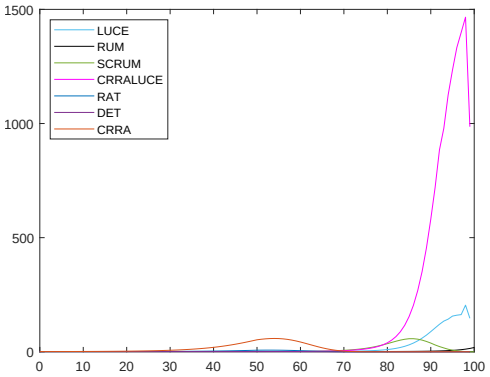
Subject 12



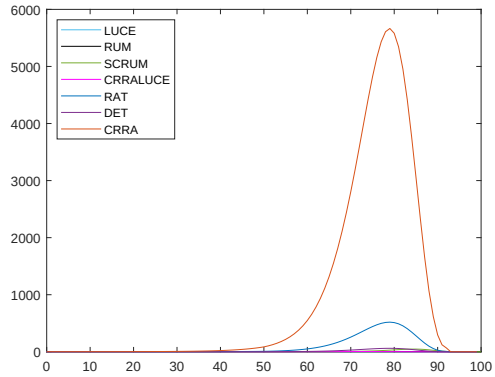
Subject 13



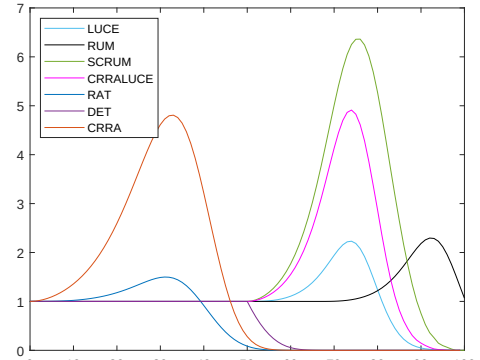
Subject 14



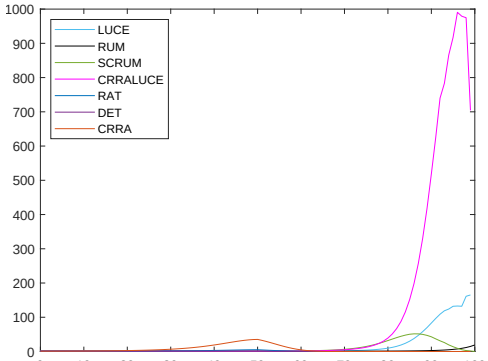
Subject 15



Subject 16



Subject 17



Subject 18

D.2 One size fits most