A Better Test of Choice Overload*

Mark Dean† Dilip Ravindran‡ Jörg Stoye§

July 1, 2024

Abstract

Choice overload – by which larger choice sets are detrimental to a chooser's wellbeing – is potentially of great importance to the design of economic policy. Yet the current evidence on its prevalence is inconclusive. We argue that existing tests are likely to be underpowered and hence that choice overload may occur more often than the literature suggests. We propose more powerful tests based on richer data and characterization theorems for the Random Utility Model. These new approaches come with significant econometric challenges, which we show how to address. We apply our tests to new experimental data and find strong evidence of choice overload that would likely be missed using current approaches.

*We recognize the advice and input from Benjamin Scheibehenne, the members of the Cognition and Decision Laboratory at Columbia University, and the Columbia Experimental Laboratory in the Social Sciences. We thank audiences at Berkeley Haas, Harvard, Princeton, Stanford, UCLA, the University of Queensland, the Pan-Asian Theory Seminar, the MiddExLab Virtual Seminar, the Zurich Workshop on Economics and Psychology, the 2023 Bounded Rationality in Choice Conference, the 2023 Econometric Society European Summer Meeting, and the 2023 Econometric Society North American Summer Meeting for feedback and Brenda Quesada Prallon for a careful reading of the near-final manuscript. We gratefully acknowledge financial support from Columbia's Experimental Laboratory in the Social Sciences and through NSF grant SES-1824375.

†Department of Economics, Columbia University. Email mark.dean@columbia.edu.

‡Humboldt University Berlin. Email: dilip.ravindran@hu-berlin.de.

§Department of Economics, Cornell University. Email stoye@cornell.edu.

1 Introduction

The standard model of utility maximization tells us that increasing the set of available options can make the consumer no worse off, and may well improve their welfare. Since the pioneering study of Iyengar and Lepper (2000),[1] a body of work in psychology has called this assumption into question. The umbrella term "choice overload" covers a number of phenomena by which larger choice sets seem to make people worse off. By now there are a huge number of studies investigating various aspects of choice overload; see Scheibehenne et al. (2010) and Chernev et al. (2015) for recent reviews and meta analyses.

Some of the output measures used to identify choice overload are hard to interpret using the classic tools of economic analysis; examples include a reduction in ex-post reported satisfaction or lower confidence that the right choice was made. Others fall very much in the realm of choice theory. In this paper we focus on the observation that larger choice sets may make people more likely to choose a "default" option.

Iyengar and Lepper (2000) provide an example of this type of behavior in their famous "jam" study. On different days a table was set up in a supermarket displaying a range of jams and offering a money-off coupon if a jam was bought. On "limited" choice days, six jams were available; on "extensive" choice days, customers faced 18 additional options. Strikingly, people were *less* likely to buy a jam on extensive choice days (2% vs 12% on limited choice days).[2] Treating "do not buy" as the default choice option, this violates the Independence of Irrelevant Alternatives (IIA) axiom.

If choice overload could be well established it would have a number of important implications for economics. From a positive perspective, it would clearly violate the standard model of utility maximization. Moreover, it is not well explained by some of the more obvious behavioral theories. For example, most models of reference dependence assume that, for choices that share the same reference point, the decision maker will behave like a standard utility maximizer, ruling out choice overload type effects (Dean, 2008). Models of limited attention can also struggle to replicate the

[1] See also Reibstein et al. (1975).

[2] These are the unconditional probabilities of buying jam. Typically the numbers quoted for this study are the probabilities of buying jam conditional on stopping at the table (3% vs 30%). However these numbers are more difficult to interpret as there may be selection into who stops at the table.

phenomenon, if they cannot explain why an overloaded individual does not simply ignore some of the available options.[3] This points to the need for models including more exotic elements such as regret (Sarver, 2008), rational contextual inference (Kamenica, 2008; Nocke and Rey, 2021) or decision avoidance (Beattie et al., 1994; Dean, 2008; Gerasimou, 2018). From a normative perspective, a key tenet of classic economics is that welfare is weakly improved by increasing the set of options available to the consumer. This assumption forms the basis of many policy recommendations, yet choice overload would call it into question.

Unfortunately, current research into choice overload is inconclusive. Some direct replications of previous experiments have failed (Scheibehenne, 2008; Greifeneder et al., 2010). One recent meta-analysis concluded that the mean measured choice overload effect is zero (Scheibehenne et al., 2010). Another one (Chernev et al., 2015) concludes that whether or not choice overload exists may depend a lot on context.

The aim of this paper is to argue that existing studies are likely to underestimate the degree of choice overload of this type.[4] Almost universally, a data set is said to exhibit choice overload only if the default is *more* likely to be chosen in a larger choice set than it is in some smaller choice set (see section 2). Yet intuitively, under the hypothesis of utility maximization one would expect the default to be chosen *much less* in larger choice sets, because the decision maker (DM) should be more likely to find something that they prefer to the default. This lack of power may potentially explain the inconsistent results from existing tests.

Two examples may further clarify this point.

Example 1.1. *Consider again the jam example. Suppose that the experimental population consists of subjects who have a probability p of liking any particular jam better than the default of no jam, where this preference is realized independently across subjects and jams. For the probability of a jam being purchased from a set of six to be* 12%*, p would have to be approximately* 2%*. This would imply a probability of active choice (i.e. choosing something other than the default) from the 24 jam set of* $1 - .98^{24} \approx .38$*, i.e.* 38%*. Thus, any choice overload effect would have to be exceptionally strong to*

[3] Though see Cattaneo et al. (2020, 2021) for examples of models of consideration sets, and de Lara and Dean (2024) for a model of sequential search that can accommodate choice overload-type behavior.

[4] Despite the above discussion, from now on we will use 'choice overload' to refer specifically to the phenomenon of choosing a default option more often in larger choice sets.

push active choice below 12% *in the larger choice set – and yet this is the benchmark typically used to identify the effect.*

Example 1.2. *The following example is even more stylized but avoids distributional assumptions. Let the choice universe be* $X = \{a, b, c, d\}$*, where* d *stands for default. Suppose* a, b, c *are individually chosen with probability* 20% *from choice sets* $\{a, d\}, \{b, d\}, \{c, d\}$*. Suppose also that the probability of active choice is* 40% *for any of* $\{a, b, d\}, \{a, c, d\}, \{b, c, d\}$*. Assuming that indifference is not allowed, these probabilities* can *be reconciled with a Random Utility Model (RUM, precisely defined later); however, they imply that the preferences* $a \succ d$*,* $b \succ d$*, and* $c \succ d$ *occur in nonoverlapping subsets of the population. Thus, the active choice probability from* X *must be at least* 60%*. If it is lower, we observe choice overload in the sense that observations cannot be explained in terms of a fully rational population, but they can be explained in terms of a population that is fully rational except that some individuals switch to default choice when encountering the grand choice set. This is true even if the active choice probability from* X *strictly exceeds the active choice probability from any smaller set.*

In this paper, we describe new tests for choice overload. We consider a data set based on a finite set of alternatives X that includes a default d. Observations of choice behavior are made from the grand set X, plus some subsets that always contain d. In each case we assume that the probability that d is chosen is observed. We make use of two definitions of choice overload. The first is model free: we say that a data set exhibits choice overload if it violates monotonicity – i.e., the probability of choosing the default option increases as more alternatives are added to the choice set. The second is based on the null hypothesis of the random utility model (RUM) – i.e. there is a set of utility functions, a (choice set independent) probability distribution over that set, and the probability of choosing any option in any choice set is the probability of the set of utility functions that make that option maximal. As example 1.2 illustrates, consistency with RUM implies monotonicity, but not vise versa, so the latter definition provides a more sensitive test for choice overload.[5] The advantage of the former is that it does not rely on the underlying assumption of random utility - monotonicity

[5] In general, it is well known that stochastic monotonicity is necessary but not sufficient for random utility. Example 1.2 clarifies that coarsening the domain of observations to active versus passive choice does not fundamentally change this.

is also implied by a number of other models of stochastic choice, such as Manzini and Mariotti's (2014) Random Consideration Set model.

If the data set consists only of choice probabilities from X and a single subset A (as is the case in Iyengar and Lepper (2000) and most other studies), then both definitions collapse to the conventional test: Assuming that only the probability of default choice is observed, the data set is consistent with monotonicity and RUM if and only if the probability of choosing d from X is weakly lower than from A. We therefore propose as a first modification of the standard approach that data should be collected from multiple subsets of X.

In this richer data, monotonicity implies that the choice of the default from X is lower than the *lowest* probability of choosing d across all observed choice sets A. We call this the *Min bound*. While easy to define, testing the Min bound is subtle because it is a test of multiple hypotheses. Even if monotonicity holds, if a finite number of observations are collected from many subsets there is a high probability that one will exhibit lower default choice than is observed from X. Essentially, the problem is the same as that of the "winners curse" – because the probability of default choice is observed with noise, the minimum of many observations is likely to be below the true minimum. To deal with this, we propose tests that take inspiration from closely related problems in the literature on moment inequalities.

Because monotonicity is necessary but not sufficient for RUM, our model-based definition of choice overload allows us to identify tighter bounds. We adapt the approach of McFadden and Richter (1991), who show that random choice data can be rationalized by the RUM if, and only if, a particular linear expression has a solution. We refer to the highest default choice probability in the large set for which such a solution exists as the *RUM bound*. The fact that it is choice frequencies, not underlying probabilities, that are observed presents a significant econometric challenge. We adapt the nonparametric test of RUM proposed by Kitamura and Stoye (2018), potentially including computational innovations by Smeulders et al. (2020) needed for applications with larger sets X.

We apply our tests to a novel experimental data set based on choices from subsets of size 2 and 3 of 12 different choice objects plus a default. The objects are verbally

described sums of numbers as used in Caplin et al. (2011); see Figures 1 and 2 for a preview. In any given choice set, the subject could choose to stick with a default which provided $3.50 for sure, or could switch to other available choices. Subjects were asked to chose from 10 such sets, one of which was the default plus the entire set of 12 other options –henceforth called the *grand set*– while the other 9 were uniformly randomly selected from all possible sets with 1 or 2 options plus default.

A total of 2000 subjects were recruited through Amazon's Mechanical Turk platform. After removing subjects who failed a comprehension quiz, we were left with 1832 subjects and 18320 choices.

Traditional approaches would be unlikely to find evidence of choice overload in our data set: the average default choice in small sets is higher than in the grand set, and only 15 of the 78 small sets have a default choice which is significantly lower than that of the grand set. Hence if, as in most work in the literature, data was collected from only a single or few small sets, one would be unlikely to detect overload. In contrast, even the most 'brute force' of our proposed tests –a potentially very conservative but finite-sample valid implementation of the Min bound– do detect it.

Our test based on RUM provides an unexpected additional insight: While RUM is indeed rejected on the whole data set, it is also rejected if we remove data from choices in the grand set. This is because subjects are more likely to choose the default option in choice sets of size 3 than their data from size 2 sets and RUM would predict. Thus, our results indicate that choice overload type effects can start in choice sets which are much smaller than have previously been demonstrated.

We hope that our work will have three consequences. First, by providing a higher powered test of choice overload, it should clear up the question of whether this is indeed a real phenomenon. Second, given that (we suspect) it will show choice overload to be more prevalent than previously thought, we hope it will spur further theoretical and policy work designed to understand its causes and mitigate its effects. Finally, by providing a better tool for measuring when choice overload does occur, we hope it will facilitate the above work by providing a better empirical basis on which to theorize.

2 Literature Review

Many studies have examined the phenomenon of choice overload from different perspectives, including different outcome measures and potential moderators. Two recent meta-analyses (Chernev et al., 2015; Scheibehenne et al., 2010) report results from 99 experiments in 53 papers and 63 experiments in 50 papers, respectively. Given these two reviews, our aim here is not to provide comprehensive coverage of the existing research. Instead, we make two points.

First, to the best of our knowledge, no other paper has made use of the methodology we propose. Using the aforecited meta-analyses and Google Scholar, we identified 32 studies from 19 papers that use default choice as a measure of choice overload. Of these, 20 ask subjects to make choices from a single subset of the grand choice set. These studies can do no better than to compare the default choice probability in the small and large choice set. The remaining 12 studies ask subjects to make choices from multiple subsets of the large choice set, and so in principle could have implemented either the min bound or tests based on the RUM. However, all of them instead compare the average default choice across all small choice sets to that in the larger set, a measure which is necessary but sufficient for the aforementioned Min bound, which in turn is necessary but not sufficient for consistency with RUM. Applying our approach to the data from these previous studies (e.g., Chernev (2005), Gingras (2003), or White and Hoffrage (2009)) is an interesting avenue for future research.

While not explicitly designed to test for choice overload, the experiment of Aguiar et al. (2023) does contain the type of data needed to perform our test. Choices were observed from all subsets of a grand set of six lotteries, with a default that is always available. The authors report finding no evidence of choice overload, and our tests confirm this result.[6] This may be due to the fact that the default alternative was chosen to be 'obviously' worse than the other available alternatives.

The second point is that the veracity and scope of choice overload is far from established. Some direct replications have failed (Scheibehenne, 2008). One meta-

[6]While we performed the test on all their treatments, their 'low' complexity treatment is the best fit to apply the test. This is because, in the 'medium' and 'high' treatments, lottery prizes were expressed as sums, and different subjects saw different sums, arguably making them different choice alternatives.

analysis (Scheibehenne et al., 2010) finds a mean effect of set size on measures of choice overload to be zero, but noted a high degree of variance. A more recent analysis (Chernev et al., 2015) identifies four variables which can increase the incidence of choice overload: decision difficulty (for example due to time constraints), choice set complexity (for example due to hard to compare alternatives), preference uncertainty (for example because the decision maker is unsure how to aggregate their preferences across many dimensions), and decision goal (for example because the decision maker is not really committed to making a purchase). Chernev et al. (2015) argue that, taking these mediators into account, there is evidence of a robust choice overload effect.

We believe that the measures we introduce can contribute to this ongoing debate by providing more highly powered evidence on when choice overload is occurring. Low powered tests may in part explain some of the contradictory results in the existing literature – for example, some studies may by chance use a small choice set that leads to choice overload, while others do not. By making it easier to spot cases of overload, we anticipate movement towards a greater acceptance that choice overload does indeed occur.

3 Theory

We now present the theoretical underpinnings of our test for choice overload. We do so in three stages. First, we describe our tests under the assumption that we can perfectly observe the probabilities with which each alternative is chosen in each choice set. Second, we discuss the econometric techniques needed to handle the fact that our experimental data are sample frequencies, not population probabilities. Finally, we describe the steps necessary to implement these econometric tests, with additional considerations that proved unnecessary for our specific experiment relegated to Appendix B.

3.1 The Population-Level Testing Problem

Let X be a finite set of alternatives and d a default alternative contained in X. Let $\mathcal{D} \subset 2^X/\emptyset$ be a collection of choice sets, all of which contain the default alternative, and which includes the grand set X.

Our data set consists of observations of choice behavior in each set in $\mathcal{D}$.[7] Specifically, we assume that we observe a function $p_d : \mathcal{D} \to [0, 1]$, with the interpretation that $p_d(A)$ is the probability of choosing the default d from choice set A.

Two notes on our data are in order. First, we assume that we observe the population probability with which the default is chosen in each choice set, but not that we can track individuals across choice problems; i.e., we observe a repeated cross section rather than a panel. This makes our approach applicable to many between-subject data sets that have this property.[8] Second, we assume that we observe only the probability with which the default was chosen in each choice set, not the probability with which specific non-default options are chosen. This is consistent with our desire to to focus on choice overload effects; the limited data only allows us to detect violations of utility maximization that occur to due too much or too little default choice from a set.

In order to design a test for choice overload, we first need to precisely define the term. Here we take two approaches. The first is to use an explicitly model-free definition, in the same way that risk aversion can be defined without reference to expected utility. An obvious candidate for such a definition is through violations of choice monotonicity, which states that the probability of choosing d cannot increase as more options are added to the choice set.

Definition 3.1. *A data set $\{\mathcal{D}, p\}$ satisfies monotonicity if, for any $A, B \in \mathcal{D}$ such that $A \subset B$,*

$$p_d(A) \geq p_d(B)$$

Specifically, one could declare choice overload to have occurred if there is a higher

[7] We follow established terminology in revealed preference theory by calling this a "data set," but note that statistically speaking, these are population probabilities; inference for finite samples is considered in the next subsection.

[8] We note that the data we collect in our experiment is somewhat richer than this, as it contains multiple observations from the same individual. We maintain this assumption to increase the generalizability of our approach.

probability of default choice in the large choice set than in the smaller one. The canonical choice overload experiment, in which $\mathcal{D} = \{A, X\}$, applies this test.

More generally, one can define

$$p_d^{min}(X) = \min_{A \in \mathcal{D} \setminus X} p_d(A)$$

as the smallest observed probability of choosing d in any set other than X. If $p_d(X) > p_d^{min}(X)$ then monotonicity is violated and the data set exhibits choice overload. We refer to $p_d^{min}(X)$ as the Min bound. Data that violates the Min bound is inconsistent with a number of models – most obviously RUM, but also the stochastic consideration set model of Manzini and Mariotti (2014)[9] and models of reference dependent preferences such as Tversky and Kahneman (1991).

A second approach is to define choice overload as a violation of a specific model. Here, the most obvious candidate would be RUM, and we will work with it, although the basic idea would generalize to any model that we know how to test. Thus, call a data set *stochastically rationalizable* if it could have been generated by a RUM. Then we can think of it as revealing choice overload if it would be stochastically rationalizable except that the probability of default choice in larger choice sets is too large.

To illustrate this idea, consider the following definitions. Note that, while $\mathcal{D}$ in our application contains X, these definitions refer to a generic $\mathcal{D}$ that may or may not do so.

Definition 3.2. *A data set $\{\mathcal{D}, p\}$ is consistent with RUM if there exist a finite collection $\mathcal{U}$ of one-to-one utility functions on X and a probability distribution $\rho \in \Delta(\mathcal{U})$ such that, for every $A \in \mathcal{D}$,*

$$p_d(A) = \sum_{u \in \mathcal{U} | d = \arg\max u(A)} \rho(u)$$

For any set $A \subseteq X$, we can then define a maximal bound on the choice of default using the default choice probabilities from the subsets of A and consistency with RUM.

[9] Other, more general models of stochastic consideration do allow for choice overload type effects - see for example Cattaneo et al. (2020, 2021)

The basic idea here goes back to Varian (1982, 1983): A counterfactual choice behavior is in the predictive bounds if, and only if, that choice behavior and previously observed ones are jointly rationalizable. Formally, we have:

Definition 3.3. *For a dataset* $\{\mathcal{D}, p\}$ *and choice problem* $A \subseteq X$*, define*

$$p_d^{RUM}(A) = \sup x \in [0,1]$$

such that the data set

$$\tilde{p}_d(\tilde{A}) = \begin{cases} x \text{ if } \tilde{A} = A \\ p_d(\tilde{A}) \text{ otherwise} \end{cases}$$

defined on the choice sets $\mathcal{D}_A \equiv \{\tilde{A} \in \mathcal{D} : \tilde{A} \subset A\}$ *is consistent with RUM.*

For most of this paper, we say that a data set exhibits choice overload if $p_d(X) > p_d^{RUM}(X)$. However, the definition of $p_d^{RUM}(\cdot)$ allows for choice overload to "kick in" for smaller choice sets and we will consider that later.

Conceptually, the Min bound and the RUM bound deal with different reasons why observations from a single small choice set might not effectively identify choice overload. The first is heterogeneity in the quality of available items. Consider a grand choice set that consists of a number of not very appealing jam flavors (gooseberry, bark, salmon) and one extremely appealing flavor (strawberry). Absent any choice overload, we would expect to see high levels of default choice in small sets that do not include the strawberry jam, and low levels of default choice in both small sets that included the strawberry jam, and the grand set. Thus, if a researcher randomly selected for analysis a small choice set without the strawberry jam, it would make it very hard to spot choice overload. Collecting data on all small choice sets and applying the Min bound would help to solve this problem, because the researcher would observe choice from a small set in which strawberry jam was included.

A second issue is preference heterogeneity. Consider example 1.2 from the introduction. In this case, all alternatives are ex ante identical, so there is no heterogeneity in item quality. This in turn means that the default choice probability is the same in any choice set with the same cardinality. Thus, collecting (for example) data from multiple sets of size two and applying the Min bound does not help. However, as the example demonstrates, the RUM bound based on observations from choice sets of size

two and three does put further constraints on the data. This is because the assumption that randomness is driven by preference heterogeneity puts extra structure on the allowed choice probabilities. Thus the RUM-based definition of choice overload gives potentially tighter bounds with which to identify choice overload.[10]

A disadvantage of RUM bound is that it requires data on $\mathcal{D}/X$ to be consistent with RUM. There are two possible issues with this. First, it could be that the population choice distributions on on $\mathcal{D}/X$ are inconsistent with RUM; in that case, $p_D^{RUM}(X)$ is not well-defined. Second, it could be that rationalizable population distributions generate non-rationalizable finite sample frequencies; in that case, one could define a feasible version of $p_d^{RUM}(X)$, and we will do so in Section 4.3.

It remains to clarify how we can test stochastic rationalizability of data. In its essence, this question has been resolved in McFadden and Richter (1991). Because we adapt their approach to an environment in which rationalizability is not characterized by standard revealed preference axioms (although they mention the possibility of this generalization) and to keep the treatment self-contained, we will briefly explain the answer. The basic insight is that probabilistic choice data can be rationalized if, and only if, it can be expressed as convex combination of data that would be produced by *deterministic* utility maximizers.[11]

To make this precise, construct a matrix $\boldsymbol{A}$ s.t. each row of $\boldsymbol{A}$ corresponds to a given alternative *within a particular choice set* (i.e., each alternative appears once for every choice set containing it), and each rationalizable deterministic choice pattern corresponds to exactly one column of $\boldsymbol{A}$. For example, consider a data set consisting of choices from $\{a_1, d\}$, $\{a_1, a_2, d\}$ and $\{a_1, a_2, a_3, d\}$. The first row of $\boldsymbol{A}$ indicates the choice of a_1 from $\{a_1, d\}$, the second row the choice of d from $\{a_1, d\}$ and so on. One column of $\boldsymbol{A}$ (namely, the first one in 3.1 below) then represents the choices of someone who picked a_1 from all three choice sets, which can be rationalized as the choices of a decision maker who prefers a_1 over all other alternatives in X. The remaining columns

[10] Note that, without preference heterogeneity, the RUM bound will equal the Min bounds. If all subjects have the same preferences, then in all small choice sets the default choice probability must be either zero or one. Thus, the Min bound must either be zero, if in some choice set the default is never chosen, or one, if the default is always chosen in every choice set. In both cases the RUM bound would be the same as the Min bound.

[11] As is often the case with seminal results, with ample hindsight the intuition informs a much easier than the original proof (Stoye, 2019).

of $\boldsymbol{A}$ represent all other rationalizable choice patterns. Some book-keeping reveals that in this example, one has:[12]

$$\begin{array}{c} a_1|\{a_1,d\} \\ d|\{a_1,d\} \\ a_1|\{a_1,a_2,d\} \\ a_2|\{a_1,a_2,d\} \\ d|\{a_1,a_2,d\} \\ a_1|\{a_1,a_2,a_3,d\} \\ a_2|\{a_1,a_2,a_3,d\} \\ a_3|\{a_1,a_2,a_3,d\} \\ d|\{a_1,a_2,a_3,d\} \end{array} \left\{ \begin{array}{cccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{array} \right\} = \boldsymbol{A}. \tag{3.1}$$

The core insight of McFadden and Richter (1991) is summarized in the following theorem.

Theorem 3.1. *Let π be the observed choice probabilities associated with each row of the matrix $\boldsymbol{A}$ derived from some stochastic choice data set. That data set is rationalizable by RUM if and only if there exists a vector $\nu \in \Delta^{H-1}$ (the $(H-1)$-dimensional unit simplex, where H is the number of columns of $\boldsymbol{A}$) such that*

$$\boldsymbol{A}\nu = \pi$$

In order to adapt this approach to our idealized data set, we need to deal with the fact that we only observe whether or not the default was chosen. To do so, we premultiply the matrix $\boldsymbol{A}$ with a matrix $\boldsymbol{B}$ that merges events we do not separately observe – i.e. the choice of different non-default options. In the above example, we

[12] The size of $\boldsymbol{A}$ increases extremely rapidly with the complexity of data; see Kitamura and Stoye (2018) for numerical examples.

would have

$$\boldsymbol{B} = \begin{Bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{Bmatrix}$$

A corollary to Theorem 3.1 then characterizes $p_d^{RUM}(X)$:

Corollary 3.1. *Let π be the observed choice probabilities associated with each row of the matrix $\boldsymbol{BA}$ derived from a data set $\{\mathcal{D}, p\}$ which reports only the probability of default choice. That data set is rationalizable by RUM if and only if their exists a probability vector ν such that*

$$\boldsymbol{BA}\nu = \pi.$$

Further, let $\boldsymbol{a}$ be the row of $\boldsymbol{A}$ that corresponds to default choice from X, then

$$p_d^{RUM}(X) = \max_{\nu \geqq 0} \{\boldsymbol{a}\nu\} \ s.t. \ \boldsymbol{BA}\nu = \pi. \tag{3.2}$$

For intuition, observe that in (3.1), the vector $\boldsymbol{a}$ is the last row of $\boldsymbol{A}$. Equivalently, it is the indicator of the choice type being the one whose choice is always the default (this is the only type who chooses the default from X). The bound simply maximizes the probability corresponding to this type, subject to overall data being stochastically rationalizable.

In practice, $p_d^{RUM}(X)$ as stated is well-defined only if π is stochastically rationalizable. Since the set of rationalizable probability vectors is "small" (we will elaborate on this in Section 4.3), empirical choice frequencies will typically –and do in our data– fail this. In that case, a feasible version of the bound can be computed by substituting a constrained estimator of π. This will be illustrated later.

We finally note that the above machinery opens the door to using this testing approach on other, for instance non-RUM, models. Nothing in the above development forces $\boldsymbol{A}$ to contain columns that correspond exactly to conventionally rationalizable behaviors. By adding (removing) columns of $\boldsymbol{A}$, one can test less (more) restrictive

models. Indeed, in our empirical application we will use this approach to test:

(i) The RUM as just explained.

(ii) A model which admits all the rationalizable deterministic choice patterns of RUM but also allows for choosing d from choice set X (regardless of behavior on smaller sets).

(iii) The same model as (ii), except that a choice type may also switch to choosing d for *all* choice sets of cardinality 3. (Types that choose d from all sets of cardinality 3 must also choose d from choice set X.)

Models (ii) and (iii) capture 'choice overloaded' behavior in that the only violations of standard rationality they allow are switches to the default when the choice set expands. Moving from (i) to (iii) enlarges $\boldsymbol{A}$, and therefore increases computational cost, but does not add conceptual difficulties. Since the resulting models are nested (listed in increasing order of generality), one can then ask what is the least permissive model that is not rejected in the data. This will allow us to unpack what sort of choice overloaded behavior, if any, could have generated our data; for reasons that will become clear, this line of analysis will be central to our results. As a preview, model (iii) will be the only one not rejected in our data. This analysis is in Section 4.3, where we also argue that model (iii) is itself a restrictive model.

We note that a further generalization which captures choice overload would be to allow switches to the default not as a function of set cardinality (as in (ii) and (iii)), but along any distinct path of choice set expansion. Dean (2008) considers a similar model. We do not consider this generalization due to computational complexity (the corresponding $\boldsymbol{A}$-matrix is very large) and because we know it would not be rejected in our data given it nests (iii).

3.2 Econometric Tests

We next explain some testing strategies that connect the above ideas to recent advances in econometrics. To this purpose, we consider samples that were generated by randomly

drawing individuals and then exposing each individual to an (i.i.d. randomly generated) selection of choice problems, i.e. subsets of $\mathcal{D}$.[13] In particular, data may contain choices from the same individuals in different choice sets, as is the case in our application. Note also that in this section, the term "data" is used in the statistical sense, i.e. referring to a finite sample from an underlying population.

We begin with two tests of the Min bound. We will estimate choice probabilities $p_d(\cdot)$ by the analogous sample frequencies, which can be informally defined as

$$\hat{p}_d(A) \equiv \frac{\sum \mathbf{1}\{\text{choice problem is } A,\ \text{choice made is } d\}}{\sum \mathbf{1}\{\text{choice problem is } A\}}, \tag{3.3}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and the sums are taken over all data points. For all but the finite-sample test that we present first, any estimator of $p_d(\cdot)$ whose asymptotic distribution is normal or approximated by the simple nonparametric bootstrap would suffice. Sample averages have both properties as long as they are not close to degenerate; by direct inspection, this condition is easily met in our data.

3.2.1 A Finite-Sample Test of the Min Bound

Testing the Min bound amounts to testing whether

$$\begin{aligned} & p_d(X) \leq \min_{A \in \mathcal{D} \setminus X} p_d(A) \\ \iff \quad & p_d(X) \leq p_d(A), \forall A \in \mathcal{D} \setminus X. \end{aligned}$$

The second expression clarifies that this is a joint test of potentially many hypotheses. Consider first testing any one of these, i.e. testing whether $p_d(X) \leq p_d(A)$ for a specific $A \subset X$. To this purpose, define $\hat{p}_{d,A}(X)$ in exact analogy to (3.3) but dropping observations from subjects who also saw choice problem A. In words, we estimate $p_d(X)$ by its simple empirical analog among subjects who did not see A. As a result, $\hat{p}_d(A)$ and $\hat{p}_{d,A}(X)$ estimate binomial proportions in two mutually exclusive samples. We will therefore apply Fisher's (1992) exact test for binomial proportions to H_0 : $p_d(X) \leq p_d(A)$ for any given A.

[13] This mirrors our empirical design, which was partly chosen because, unlike stratified sampling or mean-reverting coins, it is easy to bootstrap.

Of course, we need to account for the fact that we will conduct many such tests once (78 in our application) and cannot assume independence. Our first approach is Bonferroni adjustment, that is, all p-values are multiplied by 78. An advantage of this approach is that it ensures finite-sample (as opposed to asymptotic) size control. However, its power is limited through three channels: The estimators $\hat{p}_{d,A}(\cdot)$ discard data; Fisher's exact test is in general conservative due to integer issues; and Bonferroni adjustment is conservative. In practice, with the sample sizes that we generated for our empirical application, we expect only the last channel to have an appreciable effect.

3.2.2 An Asymptotic Test of the Min Bound

The finite sample test adjusts for the fact that, in principle, many tests are conducted simultaneously. A common concern with such adjustments is that, if the results of some of these tests appear obvious, one might needlessly lose power. Indeed, consider the visualization of our data in Figure 3: The default probabilities in the right-hand cluster are obviously much higher than in the grand choice set (corresponding to the dashed line). Can we restrict attention to only those inequality conditions that might reasonably bind?

This question has received considerable attention in the econometric literature on moment inequalities and is by now well understood. We implement a method that can be seen as special case of Andrews and Soares (2010) and also of Chernozhukov et al. (2013), both of whom establish its validity under rather general conditions.

The method can in principle use many test statistics; for concreteness, set

$$t = \hat{p}_d(X) - \min_{A \in \mathcal{D} \setminus X} \hat{p}_d(A).$$

The test will reject if t is too large. The catch is that the distribution of t, and therefore the appropriate critical value, depends on the nuisance parameter $(p_d(X) - p_d(A))_{A \in \mathcal{D} \setminus X}$. This parameter cannot be pre-estimated with sufficient accuracy[14] and so we must conservatively approximate it. This is done in three steps:

[14] Technically, it enters the asymptotic distribution scaled by $\sqrt{n}$. Our exposition is slightly informal because the issue is well understood in the literature; see also Canay and Shaikh (2017, section 4) for a survey.

1. Use the simple nonparametric bootstrap to approximate the distribution of

$$(\hat{p}_d(A) - p_d(A))_{A \in \mathcal{D}}$$

by the (bootstrap) distribution of

$$(\tilde{p}_d(A))_{A \in \mathcal{D}} \equiv (\hat{p}^*_d(A) - \hat{p}_d(A))_{A \in \mathcal{D}},$$

where $\hat{p}^*_d(\cdot)$ denotes the bootstrap analog of $\hat{p}_d(\cdot)$. This bootstrap will be clustered by individual, i.e. we (i.i.d. uniformly with replacement) resample individuals and use all responses from a given resampled individual; this ensures that correlation patterns in $(\hat{\pi} - \pi)$ due to eliciting several responses per individual are captured.

2. Use a pre-test with size converging to 0, e.g. $\alpha_n = 1/\log(n)$, and discard from consideration any sets A s.t. the null hypothesis $H_0 : p_d(X) \geq p_d(A)$ is rejected at significance level α_n. Let $\mathcal{D}^*$ denote the set of choice problems that are retained in this pre-test.

3. The critical value of our test is the appropriate quantile of the recentered bootstrap test statistic

$$t^* \equiv \tilde{p}(X) - \min_{A \in \mathcal{D}^* \setminus X} \tilde{p}(A).$$

This procedure reflects two important ideas from the moment inequalities literature. First, the bootstrap population of data must be on the null hypothesis, which necessitates a recentering. In our case, the least favorable and therefore relevant instance of the null hypothesis is that all relevant probabilities are equal. Since the test statistic is location invariant, for concise notation and implementation we recenter them to 0. This is reflected in the definition of $\tilde{p}_d(\cdot)$. Second, the test may be extremely conservative if we accordingly recenter all 78 estimators. Therefore, we pre-screen choice items whose default probability is likely to much exceed $p_d(X)$. This is a special case of Generalized Moment Selection (Andrews and Soares, 2010). In particular, the size of the pre-test must go to 0 in order for us to claim that we will asymptotically select all binding constraints.[15]

[15] Our depiction of the method is simplified by picking a specific test statistic and also filling in

3.2.3 An Asymptotic Test of RUM and its Generalizations

Statistical testing of random utility models is due to Kitamura and Stoye (2018), with important computational improvement by Smeulders et al. (2020). It has seen application to observational data (Deb et al., 2022) as well as lab experiments (Aguiar et al., 2022). Because we will slightly adapt the approach to our setting, and also for self-contained exposition, we next briefly explain it. We also remind the reader that RUM is here meant in a somewhat loose sense: It is not essential that the population is modelled as mixing a finite number of "admissible" types encoded in the columns of $\boldsymbol{A}$, not that admissibility coincides with standard economic rationality; hence, we can use the machinery to test nonstandard and, in particular, nested models.

The Hypothesis Test The null hypothesis

$$H_0 : \boldsymbol{B}\boldsymbol{A}\nu = \pi, \exists \nu \in \Delta^{H-1}$$

can equivalently be written as

$$\begin{aligned} H_0 : \quad & \boldsymbol{B}\boldsymbol{A}\nu = \pi, \exists \nu \geq 0 \\ \iff \quad & \pi \in \mathcal{C} \equiv \{\boldsymbol{B}\boldsymbol{A}\nu : \nu \geq 0\} \\ \iff \quad & \min_{\nu \geq 0} \{(\pi - \boldsymbol{B}\boldsymbol{A}\nu)' \Omega (\pi - \boldsymbol{B}\boldsymbol{A}\nu)\} = 0, \end{aligned}$$

where Ω is an arbitrary positive definite (and in practice diagonal) weighting matrix. Here, the first step is true because a vector ν can fulfil $\boldsymbol{B}\boldsymbol{A}\nu = \pi$ only if its elements sum to 1, thus we can relax the constraint on ν. The last expression states that the residuals from projecting π onto the cone $\mathcal{C}$ of rationalizable probabilities must equal 0 and suggests the scaled norm of the corresponding sample residuals as natural test statistic. Noting the similarity to specification tests in multiple equation models

specific values for several tuning parameters. For example, the sum $\sum_{A \in \mathcal{D}^* \setminus X} \max\{\hat{p}_d(X) - \hat{p}_d(A), 0\}$ of constraint violations might yield a test statistic that is more powerful against small but uniform violations. Also, rather than letting the size of a pre-test vanish, one could use Bonferroni correction to "spend" some "coverage budget" on the pre-test (Andrews and Barwick, 2012; Romano et al., 2014). Again, we omit these generalizations for conciseness because they are known in the literature. The test that we implement is as stated, and the rejection in our empirical application is so resounding that none of these choices would plausibly have impacted it.

(Sargan, 1958; Hansen, 1982), call this statistic

$$J_n \equiv n \min_{\nu \geq 0} \{(\hat{\pi} - \boldsymbol{B}\boldsymbol{A}\nu)' \Omega (\hat{\pi} - \boldsymbol{B}\boldsymbol{A}\nu)\},$$

where $\hat{\pi}$ stacks estimated choice probabilities $(\hat{p}_d(A), 1 - \hat{p}_d(A))$ in an order corresponding to π and n is sample size.

The asymptotic distribution of J_n is delicate to estimate because it depends discontinuously on where on $\mathcal{C}$ the true π is. Valid inference therefore relies on a bootstrap procedure that not only recenters the empirical distribution to be on H_0 (as in the previous test) but that also shrinks H_0 in a locally unfavorable direction. Heuristically, the idea is once again similar to Generalized Moment Selection in moment inequalities, but the details differ because, while $\mathcal{C}$ is a finite polytope, we can only characterize it in terms of its vertices (the columns of $\boldsymbol{A}$) as opposed to its faces (which would correspond to linear inequality constraints and allow for Generalized Moment Selection if we had them).

Operationally, we recenter the bootstrap distribution of data at the projection onto the tightened cone

$$\mathcal{C}_{\tau_N} \equiv \{\boldsymbol{B}\boldsymbol{A}\nu : \nu \geq \mathbf{1} \cdot \tau_n / H\},$$

where τ_n is a tuning parameter that we set equal to $\sqrt{\log(\underline{n})/\underline{n}}$ with $\underline{n}$ being expected sample cell size for any but the universal choice problem; $\mathbf{1}$ is a vector of 1's.[16] The essential feature here is that τ_n vanishes but slowly compared to estimation uncertainty in $\hat{\pi}$. Then the bootstrap analog of J_n is defined as

$$\begin{aligned} J_n^* &\equiv \min_{\nu \geq \mathbf{1} \cdot \tau_n / H} \{(\hat{\pi}^*_{\tau_n} - \boldsymbol{B}\boldsymbol{A}\nu)' \Omega (\hat{\pi}^*_{\tau_n} - \boldsymbol{B}\boldsymbol{A}\nu)\} && (3.4) \\ \hat{\pi}^*_{\tau_n} &\equiv \hat{\pi}^* + \hat{\eta}_{\tau_n} - \hat{\pi} \\ \hat{\eta}_{\tau_n} &\equiv \arg \min_{\nu \geq \mathbf{1} \cdot \tau_n / H} \{(\hat{\pi} - \boldsymbol{B}\boldsymbol{A}\nu)' \Omega (\hat{\pi} - \boldsymbol{B}\boldsymbol{A}\nu)\} && (3.5) \end{aligned}$$

and $\hat{\pi}^*$ is a simple nonparametric (clustered, as explained earlier) bootstrap analog of $\hat{\pi}$.

[16] Simple algebra reveals that $\underline{n} = \frac{2qn}{k(k+1)}$, where k is the number of nondefault items in X and q is the number of choice problems other than the grand problem faced by each subject. Recall also that H is the length of ν, thus division by H ensures that the above constraint is scaled by τ_n and not by the testing problem's complexity.

Validity of the resulting test under reasonable assumptions is established in Kitamura and Stoye (2018). We go beyond a completely straightforward implementation of their test because we do not weight questions equally. This possibility is anticipated in Kitamura and Stoye (2018) because they require only a diagonal weighting matrix Ω but has not, to our knowledge, been implemented before. It is motivated by the fact that, in our data, choice probabilities pertaining to the universal choice set X will be estimated from a much larger sample cell, and therefore much more precisely, than others. Specifically, both the expected and also realized average relative cell size of the last versus other questions are determined by the experimental design, and it is this number that we will use, i.e. the weighting matrix will be

$$\Omega = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & w & 0 \\ 0 & \cdots & 0 & 0 & w \end{pmatrix},$$

where w is easy to compute and takes a value of $w \approx 9$ in our experimental design.[17]

We close by observing that in general, this test can be expensive to compute. The experimental design that we settled on, partly to ensure reasonable sample cell sizes, is small enough so that this is not the case. However, in preparation, we also implemented an adaptation of the computational improvements in Smeulders et al. (2020). For the benefit of future users, these details are laid out in Appendix B.

4 An Application to Experimental Data

We next describe an application of the above methods to an experimental data set.

[17] Indeed, $w = \frac{k(k+1)}{2q}$, with (k, q) as in the preceding footnote. We do not weight questions by *realized* sample cell sizes, and we also do not estimate cell-specific variances by the binomial variance formula, in order to avoid data dependent weighting. In our application, these modifications would have minimal effect.

4.1 Experimental Design

The aim of the experiment is to collect data of the type needed to implement tests for the Min and RUM bounds. This means we require repeated observations of choices from a grand set of alternatives and a number of subsets, with all sets containing a default option. Moreover, based on the findings of Chernev et al. (2015), we want the choice problems to be non-trivial to increase the probability of finding choice overload. To this end, we ask subjects to make choices between amounts of experimental points expressed as sums, as in Caplin et al. (2011).

Specifically, each non-default option is expressed as a sum of four numbers between zero and ten written in text. The value of choosing an option in experimental points is the value of the sum, and one experimental point is worth 50 cents. There are 12 non-default options in the grand set X. Following Caplin et al. (2011), we generated these as follows: First, we drew the value of an alternative was drawn from an exponential distribution truncated at 10 points with $\lambda = 0.25$. Next, individual terms of sums were chosen stochastically ensuring neither the first nor the maximal summand was correlated with the total value of the option.

Meanwhile, a default option was available in each choice set. This option provided 7 points and was the only option expressed as a single number. The default option also always appeared at the top of the screen, was pre-selected (i.e. the subject had to actively click on another option if they wanted to choose it) and the simplest alternative (as measured by the number of summands in the expression). Thus it fulfilled many criteria that have been proposed to define the "default" in choice overload situations (see for example Iyengar and Kamenica (2010)). Of the non-default options, 9 yielded a number of points strictly lower than the default, 1 yielded the same number of points as the default, and 2 yielded strictly more points. Figure 1 shows an example choice screen with the default and one other option. Figure 2 shows an example choice screen for the grand set.

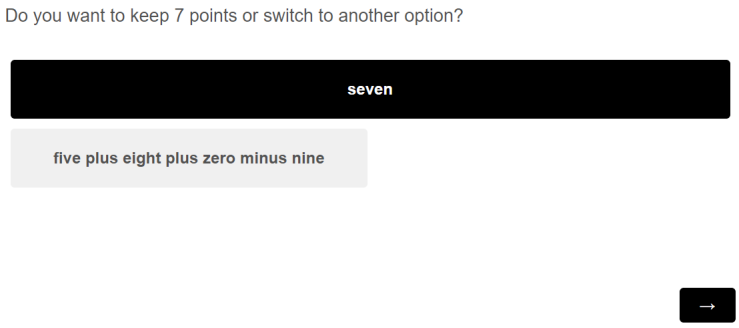


Figure 1: A choice problem comparing one lottery to the default.

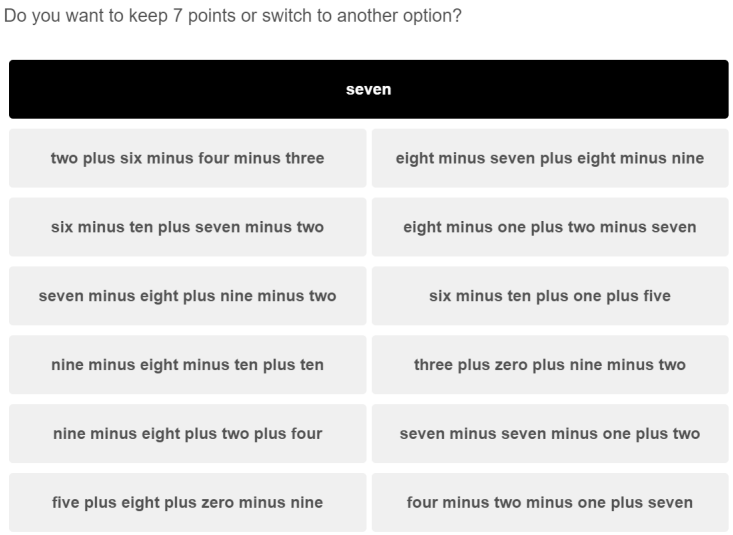


Figure 2: A choice problem comparing the grand set to the default.

The collection of choice sets used in the experiment consisted of the grand set (12 alternatives plus the default), and all 1 and 2 alternative subsets (along with the default), for a total of 78 smaller choice sets. Based on the prior literature, we initially believed that 3 item sets are unlikely to trigger choice overload, while 13 item sets are likely to do so, if such an effect is present. Each subject was presented with 9 randomly selected small sets and the grand set, with the order of choice questions and the order of non-default options in each choice question randomized. One question was randomly selected for payment. Subjects in addition received a $1 participation fee. A complete list of the choice alternatives and choice sets can be seen in Appendix A.2.[18]

[18]The experiment was approved by the Institutional Review Boards of Columbia and Cornell Uni-

Note that we have selected a choice environment that should generate a good deal of heterogeneity in item quality, but in which there should be little preference heterogeneity. Based on the discussion of Section 3.1 we would anticipate the Min bound to be effective in identifying choice overload, whle the RUM bound may not necessarily provide further advantages.

Subject Recruitment. The experiment was run on Amazon's Mechanical Turk (MT) platform, a digital marketplace for work. This platform was chosen to make it easy to get data from a large number of subjects, each of whom needed to answer only a limited number of questions. "Requesters" post Human Intelligence Tasks, or "HITs," which are usually simple, repetitive jobs that typically pay small sums for each completed task. Workers on MT view descriptions of the HITs, decide which to accept, and complete those HITs over the internet. In our case, subjects who accepted the HIT followed a link to an external webpage, where they completed the experiment. Upon completion they were given a randomly generated code, which was used to pay them the appropriate amount given their choices in the experiment.

Instead of posting HITs and recruiting subjects directly on MTurk, we used an online platform called CloudResearch. CloudResearch posts our experiment as a HIT on MTurk and recruits the subjects. All subjects recruited through CloudResearch's platform are first screened through CloudResearch's Sentry data validation system. Sentry is a short (1 minute) pre-experiment survey meant to ensure subjects are attentive, engaged, and ready to participate in the experiment. People who fail Sentry are routed away from the experiment. These pre-study validation surveys have been shown to be effective in increasing data quality (Chandler et al., 2019; Litman and Robinson, 2020).

We recruited 2000 subjects for the experiment. In addition to CloudResearch's screening, we restricted recruitment to MTurk workers who had completed over 1000 HITs and had an approval rating of over 97%. Following the instructions subjects had to answer two quiz questions to ensure the understood the instruction. 1833 subjects passed the quiz and completed the experiment, but one of these subjects had a browser related error which made their data unusable, leaving us with 1832 subjects' data.

versities.

The resulting data provides 1832 choices from the grand set. We have at least 185 observations of choices from each of the small sets, with variation in sample size due to the random selection of choice questions.

4.2 Data Overview

Figure 3 presents a histogram showing the distribution of choice sets by the probability with which the default option was chosen from sets other than X. The distribution is bimodal. The choice sets with low probability of default choice are precisely the ones in which the best option yields more points than the default. The probability of default choice in X is indicated by the dotted vertical line. It falls somewhat above the default choice probability of the lower group, but below that of the higher group. Appendix A.2 lists the default choice probabilities for each choice set.

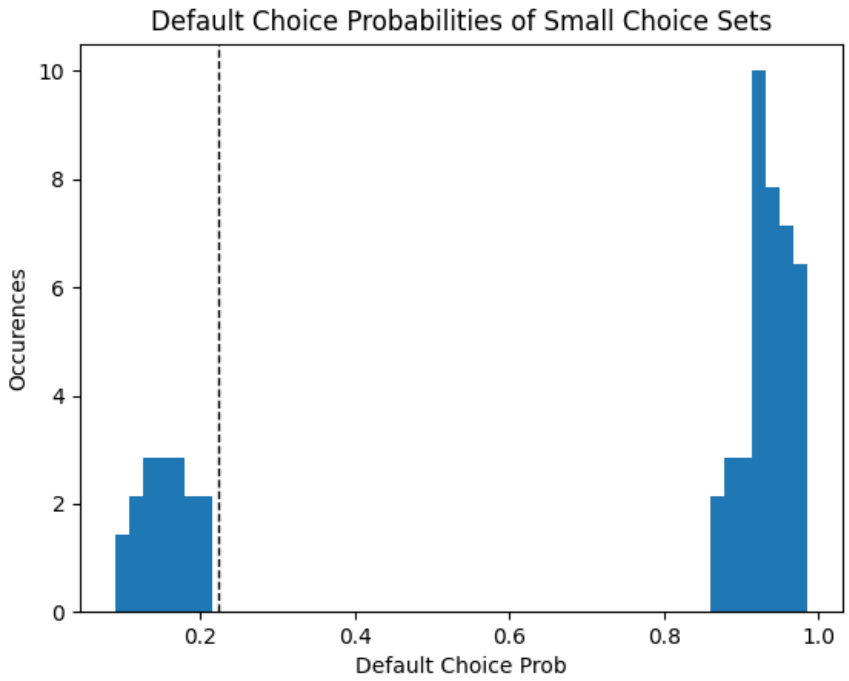


Figure 3: Histogram showing distribution of choice sets by probability that default option was chosen. Dotted line shows the fraction of default choice in the grand set X.

4.3 Analysis

We next ask whether the data we collected are suggestive of choice overload by conducting tests in roughly increasing order of presumed power.

First, we establish that existing approaches would be unlikely to identify choice overload in our data set. The empirical frequency of default choice in the grand set, $\hat{p}_d(X)$, was 22.3%. The analogous empirical frequency across all subsets was 70.7% and therefore considerably higher. A simple comparison of means would not reveal choice overload, but also masks extreme heterogeneity across choice items. A random selection of a single smaller set would also be unlikely to reveal choice overload. 23 of 78 small choice sets had default choice frequency below 22.3%, only 15 of which significantly so at the 5% level (using Fisher's exact test on disjoint samples but without controlling for multiple comparisons). Thus, if a researcher were to randomly select a small choice set to compare to the grand set, they would only find evidence for choice overload 19% of the time.

In a next step, we use the finite sample test to ask whether any of those 23 frequencies is significantly below the analogous frequency in the grand choice set, taking into account sampling uncertainty. Strikingly, the answer is yes: The p-value against the null hypothesis that the grand choice default frequency is lowest is .00003, and 5 choice items are significantly below it at the 5%-level, where the p-values and significance levels just reported reflect Bonferroni adjustment across 78 tests.[19] Similarly, the asymptotic test of the Min bound yields a p-value that we could not distinguish from 0 in $B = 10000$ Monte Carlo simulations. Hence we find strong evidence of choice overload using this approach.

We next apply the test explained in Section 3.2.3 to see if our data is consistent with RUM. At a Monte Carlo replication size of $B = 10000$, the p-value against this null hypothesis is equals 0 as well. This extremely strong rejection comes with a caveat: The p-value against all but the grand choice set equals .005. Therefore, "the data are consistent with RUM except that default choice from X is too frequent" is not an appropriate description of our findings. However, this leaves open the possibility that choice overload had an effect already in some of the smaller choice sets.

[19] The choice sets alluded to are numbers $(9, 11, 20, 32, 74)$ in order of appearance in Table 2.

To test this, we next consider models (ii) and (iii) from the end of Section 3.1 by running the test from Section 3.2.3 with appropriately modified $\boldsymbol{A}$-matrices. The p-value using all data but applying extension (ii), i.e. allowing for choice types that are rationalizable but choose d from X, is also 0.005.[20] In contrast, the more general model (iii), i.e. allowing for subjects to switch to d at X or at all choice sets of size 3 and up, is not rejected ($p = .65$). Given the details of our testing procedure, this also implies that the null hypothesis corresponding to the more restrictive model (ii) would be rejected while imposing the less restrictive model (iii).[21] Subjects' behavior is therefore consistent with exhibiting choice overload already at sets of size 3 (as well as at the grand set) and otherwise being rational. As prior experiments in the literature generally looked for choice overload kicking in at larger set sizes,[22] it is economically interesting that subjects are choice overloaded when facing such small sets.

We close with four additional observations. First, with hindsight our data speaks so loudly that our more sensitive tests were not needed. As an informal illustration of their higher sensitivity, we replicated the entire analysis above on the first 50% of subjects. As expected, all p-values crept up. Of note, model (ii) above, previously rejected with $p = .005$, was not any more rejected ($p = .250$). Furthermore, while the p-value associated with the asymptotic Min bound test (see Section 3.2.2) remains effectively 0, that associated with the exact test (see Section 3.2.1) is above 1%; thus, at the 1% level we would declare choice overload using the former but not the latter test. With the benefit of hindsight, the use case for the asymptotic tests would therefore have been more striking if we had collected only half the data; needless to say, this sort of thing can happen if one designs tests before collecting data (as we did).

Second, to further quantify the sense in which these models fit the data, we compute

[20] This is expected in view of the preceding paragraph's results because economically, model (i) restricted to choice sets excluding X is equivalent to model (ii). That said, p-values need not be numerically the same due to subtleties of how the testing problem gets regularized. However, one would informally expect them to give very similar results, and their unrounded values in our data and using identical bootstrap draws are indeed .0047 vs .0046.

[21] More precisely, this null hypothesis is linear (it can be written as $\boldsymbol{e}'\nu = 0$, where the vector $\boldsymbol{e}$ is an indicator vector of columns of $\boldsymbol{A}$ that would reveal choice overload) and therefore can be tested using results from Deb et al. (2022). However, close inspection reveals that that test will numerically coincide with a direct test of the more restrictive model.

[22] Most papers in the literature employ small sets that are larger than size 3. Tversky and Shafir (1992) report an increase in default choice when switching from size 2 to size 3 choice sets, but ascribe this to the disjunction effect, rather than choice overload.

the largest sample proportion of subjects such that individually rational behavior by these subjects would be compatible with empirical choice frequencies. This proportion is defined by the linear program

$$\max \mathbf{1}'\nu \text{ s.t. } \boldsymbol{A}\nu \leq \hat{\pi}$$

and equals 1 if, and only if, sample proportions are rationalizable in terms of rational matrix $\boldsymbol{A}$. This fraction is .866 for the RUM, i.e. model (i) above, increases to .877 for model (ii), and equals .915 for model (iii).

Third, one may worry that model (iii) is just not very restrictive, while maybe models (i) and (ii) are. This cannot be literally true because our test statistics are positive in all three tests, hence empirical choice frequencies are not rationalizable under any model. But it could certainly be "morally true," notably if all possible data sets are "close" to the model. In order to shed some light on this, we conducted some analyses inspired by Bronars (1987), Selten (1991), and Beatty and Crawford (2011). Specifically, for models (i)-(iii) we calculated the expected mean square error (MSE) when matching data that were generated from the uniform distribution on $[0,1]^{79}$ (i.e. choice probabilities drawn uniformly at random). The resulting values are 0.25 for model (i), 0.23 for (ii) and 0.21 for (iii). These numbers are all close to each other, and far above the MSE when these models are applied to the data (which range from 0.03 to 0.01), showing that even our most permissive model places significant restrictions on the data.

Finally, the data can be used to illustrate the difference between the Min bound, $p_d^{min}(X)$, and the RUM-based bound, $p_d^{RUM}(X)$. The empirical choice frequencies do not imply a well-defined $p_d^{RUM}(X)$ because they are not stochastically rationalizable even if choice set X is ignored. However, one can easily compute the vector $\hat{\eta}$ of choice probabilities that are closest to the empirical ones while being rationalizable; in fact, this computation is a by-product of the statistical tests. This allows us to compute a feasible analog of $p_d^{RUM}(X)$. Its value in our data is 9.8%. To compare apples to apples, we report that the same $\hat{\eta}$ implies a Min bound of $p_d^{min}(X) = 11.4\%$. Using the full implications of RUM is potentially much more informative than just testing monotonicity.

5 Conclusion

In this paper, we have argued that existing tests for choice overload are underpowered, and that this may explain the confusing and contradictory picture that emerges from the current literature. We proposed that, by collecting more data and making use of theory, one can design better tests. We have demonstrated that, in a novel data set, our approach finds choice overload. Indeed, our new data speak so loudly that with hindsight, our methodological innovations would not have been necessary to detect choice overload in our own data set. We believe that the innovations are of interest nonetheless, and we also note that such things are bound to occur if –as we did– one genuinely designs the empirical strategy before going into the field.

One unexpected finding from our experiment is that choice appears to violate the RUM even in smaller choice sets. This is perhaps surprising given the relative simplicity of the choice items. Understanding the cause of these violations strikes us as an interesting avenue for future research. For our purposes, it suggests that RUM may not be an appropriate benchmark against which to look for choice overload, at least in this experimental task.

References

Aguiar, V., M. Boccardi, N. Kashaev, and J. Kim (2022): "Random Utility and Limited Consideration," *Quantitative Economics*, in press.

Aguiar, V. H., M. J. Boccardi, N. Kashaev, and J. Kim (2023): "Random utility and limited consideration," *Quantitative Economics*, 14, 71–116.

Andrews, D. W. K. and P. J. Barwick (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80, 2805–2826.

Andrews, D. W. K. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Beattie, J., J. Baron, J. C. Hershey, and M. D. Spranca (1994): "Psychological determinants of decision attitude," *Journal of Behavioral Decision Making*, 7, 129–144.

Beatty, T. K. M. and I. A. Crawford (2011): "How Demanding Is the Revealed Preference Approach to Demand?" *American Economic Review*, 101, 2782–95.

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

Canay, I. A. and A. M. Shaikh (2017): *Practical and Theoretical Advances in Inference for Partially Identified Models*, Cambridge University Press, vol. 2 of *Econometric Society Monographs*, 271–306.

Caplin, A., M. Dean, and D. Martin (2011): "Search and satisficing," *American Economic Review*, 101, 2899–2922.

Cattaneo, M. D., P. Cheung, X. Ma, and Y. Masatlioglu (2021): "Attention overload," *arXiv preprint arXiv:2110.10650*.

Cattaneo, M. D., X. Ma, Y. Masatlioglu, and E. Suleymanov (2020): "A random attention model," *Journal of Political Economy*, 128, 2796–2836.

Chandler, J., C. Rosenzweig, A. J. Moss, J. Robinson, and L. Litman (2019): "Online panels in social science research: Expanding sampling methods beyond Mechanical Turk," *Behavior research methods*, 51, 2022–2038.

Chernev, A. (2005): "Feature complementarity and assortment in choice," *Journal of Consumer Research*, 31, 748–759.

Chernev, A., U. Böckenholt, and J. Goodman (2015): "Choice overload: A conceptual review and meta-analysis," *Journal of Consumer Psychology*, 25, 333–358.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737.

de Lara, L. and M. Dean (2024): "Rational Choice Overload," .

Dean, M. (2008): "Status quo bias in large and small choice sets," *Unpublished working paper*.

Deb, R., Y. Kitamura, J. K.-H. Quah, and J. Stoye (2022): "Revealed Price Preference: Theory and Empirical Analysis," *Review of Economic Studies*, in press, ceMMAP working paper CWP57/18.

Fisher, R. A. (1992): "Statistical methods for research workers," in *Breakthroughs in statistics*, Springer, 66–70, originally published in 1934.

Gerasimou, G. (2018): "Indecisiveness, undesirability and overload revealed through rational choice deferral," *The Economic Journal*, 128, 2450–2479.

Gingras, I. (2003): "Dealing with too much choice." Ph.D. thesis, ProQuest Information & Learning.

Greifeneder, R., B. Scheibehenne, and N. Kleber (2010): "Less may be more when choosing is difficult: Choice complexity and too much choice," *Acta psychologica*, 133, 45–50.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Iyengar, S. S. and E. Kamenica (2010): "Choice proliferation, simplicity seeking, and asset allocation," *Journal of Public Economics*, 94, 530–539.

Iyengar, S. S. and M. R. Lepper (2000): "When choice is demotivating: Can one desire too much of a good thing?" *Journal of personality and social psychology*, 79, 995.

Kamenica, E. (2008): "Contextual inference in markets: On the informational content of product lines," *American Economic Review*, 98, 2127–49.

Kitamura, Y. and J. Stoye (2018): "Nonparametric Analysis of Random Utility Models," *Econometrica*, 86, 1883–1909.

Litman, L. and J. Robinson (2020): *Conducting online research on Amazon Mechanical Turk and beyond*, Sage Publications.

Manzini, P. and M. Mariotti (2014): "Stochastic choice and consideration sets," *Econometrica*, 82, 1153–1176.

McFadden, D. and K. Richter (1991): "Stochastic rationality and revealed stochastic preference," in *Preferences, Uncertainty and Rationality*, ed. by J. Chipman, D. McFadden, and K. Richter, Boulder: Westview Press, 161–186.

Nocke, V. and P. Rey (2021): "Consumer Search and Choice Overload," .

Reibstein, D. J., S. A. Youngblood, and H. L. Fromkin (1975): "Number of choices and perceived decision freedom as a determinant of satisfaction and consumer behavior." *Journal of Applied Psychology*, 60, 434.

Romano, J. P., A. M. Shaikh, and M. Wolf (2014): "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 82, 1979–2002.

Sargan, J. D. (1958): "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, 26, 393–415.

Sarver, T. (2008): "Anticipating regret: Why fewer options may be better," *Econometrica*, 76, 263–305.

Scheibehenne, B. (2008): "The effect of having too much choice," .

Scheibehenne, B., R. Greifeneder, and P. M. Todd (2010): "Can there ever be too many options? A meta-analytic review of choice overload," *Journal of consumer research*, 37, 409–425.

Selten, R. (1991): "Properties of a measure of predictive success," *Mathematical Social Sciences*, 21, 153–167.

Smeulders, N., L. Cherchye, and B. de Rock (2020): "Nonparametric analysis of random utility models: computational tools for statistical testing," *Econometrica*, forthcoming.

Stoye, J. (2019): "Revealed Stochastic Preference: A one-paragraph proof and generalization," *Economics Letters*, 177, 66–68.

Tversky, A. and D. Kahneman (1991): "Loss aversion in riskless choice: A reference-dependent model," *The quarterly journal of economics*, 106, 1039–1061.

Tversky, A. and E. Shafir (1992): "The Disjunction Effect in Choice under Uncertainty," *Psychological Science*, 3, 305–309.

Varian, H. (1982): "The Nonparametric Approach to Demand Analysis," *Econometrica*, 50, 945–972.

——— (1983): "Non-parametric Tests of Consumer Behaviour," *Review of Economic Studies*, 50, 99–110.

White, C. M. and U. Hoffrage (2009): "Testing the tyranny of too much choice against the allure of more choice," *Psychology & Marketing*, 26, 280–298.

A Supplementary Appendix: Experimental Details

Here we present the experiment instructions, quiz, and details about the choice objects.

A.1 Instructions and Quiz

Subjects were first shown this instructions screen before proceeding to the quiz (below). Subjects were given two attempts at the quiz.

Instructions

You will be paid $1 for completing this HIT. After you finish, you will be given an MTurk completion code. You must enter this code into MTurk to receive payment.

In addition, you will receive a bonus payment which will be between $0 and $5 dollars. You will be asked to answer 10 questions. At the end of the HIT, one of these questions will be chosen at random. Each question is equally likely to be chosen. Your bonus will only depend on how you answered this question. Because all questions have a chance of determining your bonus payment, it is important that you choose your preferred answer to each question.

The questions you are asked will determine how many experimental points you receive. At the end of the experiment, any points you earn will be converted into your bonus payment: each point is worth $0.50.

At the start of each question, you will be given the option of getting 7 points (this option will be highlighted in black). If you want to keep this option, then simply click the arrow on the bottom of the screen. Alternatively, you can switch to another option by clicking that option which will then turn black. All the other options will be written as 4 numbers added and/or subtracted together; each of these options is worth the number of points between 0 and 10 that is expressed by the total sum.

Below is a sample question. If your bonus payment was based on this question, you would receive 7 points if you stayed with the first (top) option. You would receive 8 points if you chose the bottom left option (the total of the sum) and 6 if you chose the bottom right option.

In this example question there are 3 choices. During the task, your questions may ask you to choose between 2, 3, or 13 different options.

Sample Question:

Do you want to keep 7 points or switch to another option?

seven	
two plus ten minus one minus three	six minus two plus nine minus seven

Quiz

To ensure you understand the instructions, please answer the two following quiz questions. You must successfully complete the quiz to continue. If you do not answer correctly, you will not earn any bonus payment and the survey will end immediately.

1. True or false: 1 of the 10 questions you answer during the HIT will be randomly chosen and your bonus payment will depend on your answer to this question. Type "True" into the box below if you think this is true and "False" if you think it is false.

→

For the next quiz question, recall that:

- Your bonus payment will be based on the sum of points given by your answer in the randomly chosen question.
- You will receive $0.50 bonus per point you earn from that answer.

→

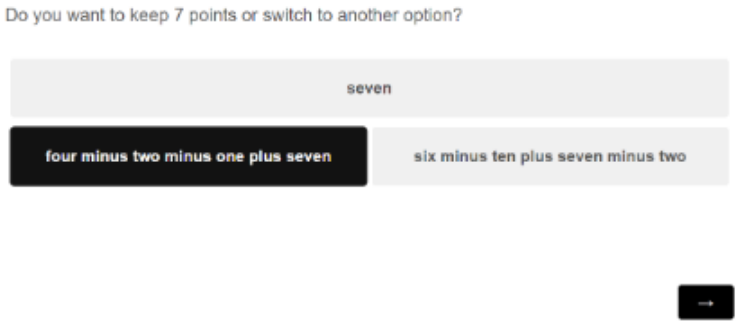
Quiz

2. Suppose you answered the question in the screenshot below as shown (i.e. you selected the option highlighted black).

If your bonus payment was based on this question: how many points would you receive and what would your bonus payment be?

Remember: the value of each option is expressed in points. Each point is worth $0.50 in bonus.

Screenshot



Answer Below:

5 points, $2.50 bonus

6 points, $3 bonus

2 points, $1 bonus

8 points, $4 bonus

16 points, $8 bonus

→

Table 1: List of Options with Values

ID	Option	Value
0	seven	7
1	eight minus seven plus eight minus nine	0
2	eight minus one plus two minus seven	2
3	seven minus seven minus one plus two	1
4	six minus ten plus one plus five	2
5	seven minus eight plus nine minus two	6
6	five plus eight plus zero minus nine	4
7	nine minus eight plus two plus four	7
8	nine minus eight minus ten plus ten	1
9	four minus two minus one plus seven	8
10	two plus six minus four minus three	1
11	three plus zero plus nine minus two	10
12	six minus ten plus seven minus two	1

A.2 Choice Alternatives with Summary Data

Table 1 gives full list of the choice objects with their values in experimental points.

Table 2 shows how often each choice set was shown to subjects and how often the default was chosen from it. Choice sets are given by the alternatives they contain: $[0, 1, 2]$ represents the choice set with the default (0) and options with IDs 1 and 2 from Table 1.

B Implementing recent computational innovations

We implemented the computational procedure in Smeulders et al. (2020). To our knowledge, this is the first such implementation beyond their own illustrative example. The implementation and some not entirely obvious modifications are described next.

The procedure is motivated by the fact that computation of the matrix $\boldsymbol{A}$ and also computation (3.4) is hard, and yet the latter needs to be repeated many times. It exploits that, because $\boldsymbol{BA}$ has many more columns than rows, there always exists a sparse (in the loose sense of having relatively few nonzero entries) arg max to problem

Choice Set	# Choices	# Default	Choice Set	# Choices	# Default
$[0,1]$	204	199	$[0,3,10]$	199	190
$[0,2]$	193	188	$[0,3,11]$	200	38
$[0,3]$	218	210	$[0,3,12]$	231	219
$[0,4]$	232	225	$[0,4,5]$	219	190
$[0,5]$	227	214	$[0,4,6]$	215	200
$[0,6]$	230	226	$[0,4,7]$	213	191
$[0,7]$	221	212	$[0,4,8]$	216	187
$[0,8]$	229	212	$[0,4,9]$	193	35
$[0,9]$	201	18	$[0,4,10]$	219	204
$[0,10]$	199	190	$[0,4,11]$	210	28
$[0,11]$	194	20	$[0,4,12]$	197	186
$[0,12]$	195	184	$[0,5,6]$	224	200
$[0,1,2]$	209	203	$[0,5,7]$	209	183
$[0,1,3]$	225	217	$[0,5,8]$	210	186
$[0,1,4]$	185	175	$[0,5,9]$	224	45
$[0,1,5]$	204	190	$[0,5,10]$	199	189
$[0,1,6]$	208	199	$[0,5,11]$	214	36
$[0,1,7]$	203	188	$[0,5,12]$	213	199
$[0,1,8]$	225	203	$[0,6,7]$	205	189
$[0,1,9]$	211	24	$[0,6,8]$	200	178
$[0,1,10]$	218	215	$[0,6,9]$	218	44
$[0,1,11]$	223	30	$[0,6,10]$	209	204
$[0,1,12]$	235	219	$[0,6,11]$	223	31
$[0,2,3]$	229	219	$[0,6,12]$	221	210
$[0,2,4]$	213	202	$[0,7,8]$	223	202
$[0,2,5]$	218	202	$[0,7,9]$	206	36
$[0,2,6]$	215	208	$[0,7,10]$	199	182
$[0,2,7]$	250	231	$[0,7,11]$	205	30
$[0,2,8]$	193	178	$[0,7,12]$	221	205
$[0,2,9]$	207	36	$[0,8,9]$	226	32
$[0,2,10]$	192	185	$[0,8,10]$	205	182
$[0,2,11]$	194	24	$[0,8,11]$	222	33
$[0,2,12]$	192	182	$[0,8,12]$	221	204
$[0,3,4]$	191	182	$[0,9,10]$	192	31
$[0,3,5]$	218	203	$[0,9,11]$	202	23
$[0,3,6]$	198	194	$[0,9,12]$	223	42
$[0,3,7]$	205	192	$[0,10,11]$	219	33
$[0,3,8]$	215	194	$[0,10,12]$	193	187
$[0,3,9]$	207	44	$[0,11,12]$	224	36
			$[0,\ldots,12]$	1832	409

Table 2: Choice Set and Default Choice Frequencies

(3.4). We will avoid solving (3.4) as stated, or ever computing $\boldsymbol{BA}$ (though the latter is feasible here), by guessing the nonzero entries. Formally, this goes as follows. (We drop N for ease of notation.):

1. Initialize the matrix $\tilde{\boldsymbol{BA}}$ by constructing relatively few columns of $\boldsymbol{BA}$.

2. Compute
$$\tilde{J} \equiv \min_{\nu \geq 0} \{(\hat{\pi} - \tilde{\boldsymbol{BA}}\nu)'\Omega(\hat{\pi} - \tilde{\boldsymbol{BA}}\nu)\}.$$
Let $\tilde{\eta} \equiv \tilde{\boldsymbol{BA}}\tilde{\nu}$, where $\tilde{\nu}$ solves this problem. (While $\tilde{\nu}$ may not be unique, $\tilde{\eta}$ is.)

3. Maximize $(\hat{\pi}-\tilde{\eta})'\Omega(a-\tilde{\eta})$ subject to the constraint that a is a column of $\boldsymbol{BA}$. This is called the "pricing problem." Its constraint must be expressed in an application specific, computable way, and we do so below.

4. If the value of the problem just solved is positive, append column a to $\tilde{\boldsymbol{BA}}$. Repeat until the value of the problem is nonpositive or another convergence criterion is met.

The basic idea is that, as long as the deficient matrix $\tilde{\boldsymbol{BA}}$ does not contain all columns that receive positive weight in one solution to the original problem, the value of the simplified problem can be improved by appending such a column. But a column improves this value iff the supporting hyperplane separating the current feasible set from $\hat{\pi}$ does not separate the new column from $\hat{\pi}$. The program in step 3 simply checks this. (We solve it but in principle, it suffices to sign its value.) If the solution is sparse, it will be found while only generating a fraction of all possible columns of $\boldsymbol{BA}$.

Our implementation is again not completely off the shelf. Modifications are as follows:

(i) We take account of the weighting matrix Ω not being the identity matrix. This is already reflected in expressions above.

(ii) The requirement that the vector a be a possible column of $\boldsymbol{BA}$ can be expressed by writing the pricing problem as follows. To enforce that a is binary and any two entries corresponding to the same choice problem sum to 1, parameterize it in

terms of a vector ρ that only collects indicators of active choice. Then $a = d + D\rho$, where

$$d = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \end{pmatrix}.$$

The objective function of the pricing problem becomes

$$(d + D\rho - \hat{\eta})^\top \Omega(\hat{\pi} - \hat{\eta}) = (D\rho)^\top \Omega(\hat{\pi} - \hat{\eta}) + const.$$

Constraints on ρ must reflect that (i) $\mathbf{0} \leq \rho \leq \mathbf{1}$; (ii) if choice from one set is active, choice from all supersets thereof is active, (iii) if the default option is chosen from all subsets of a set, then it is chosen from the set as well.

In sum, the pricing problem can be expressed as the following integer linear program:

$$\begin{aligned} \max_{\rho \in \{0,1\}^I} & (D\rho)^\top \Omega(\hat{\pi} - \hat{\eta}) \\ \text{s.t. } \rho_i - \rho_j &\geq 0 \text{ whenever choice problem } i \text{ contains problem } j \\ \rho_i &\leq \sum_{j=1,\dots,k: x_j \in X_i} \rho_j. \end{aligned}$$

(iii) At first glance, the tightened optimization problem (3.4) has no sparse solution, but Smeulders et al. (2020) remedy this. Heuristically, the vector $\mathbf{1} \cdot \tau_N / H$ can be concentrated out of the problem and a problem with sparse solution remains. A catch is that this requires the initial guess $\tilde{\boldsymbol{B}}\tilde{\boldsymbol{A}}$ to have the same dimension as the true A (its column cone cannot be contained in a face of the $\mathcal{C}$). Smeulders et al. (2020) generate columns at random and verify that this constraint is met. This will not work here because only one of possibly millions of choice types makes a default choice on the universal set. Random column generation would be unlikely to discover that type, and so we seed $\tilde{\boldsymbol{B}}\tilde{\boldsymbol{A}}$ with the corresponding column, 300 additional random columns, and verify the rank condition. This is a problem and a fix that is likely to apply to other applications of the method as well.