A Model of Justification

Sarah Ridout*

October 14, 2020

Most recent version here.

Abstract

I model decision-making constrained by morality, rationality, or other virtues. In addition to a primary preference over outcomes, the decision maker (DM) is characterized by a set of preferences that he considers justifiable. In each choice setting, he maximizes his primary preference over the subset of alternatives that maximize at least one of the justifiable preferences. The justification model unites a broad class of empirical work on distributional preferences, charitable donations, prejudice/discrimination, and corruption/bribery. I provide full behavioral characterizations of several variants of the justification model as well as practical tools for identifying primary preferences and justifications from choice behavior. I show that identification is partial in general, but full identification can be achieved by including lotteries in the domain and allowing for heterogeneity in both primary preferences and justifications. Since the heterogeneous model uses between-subject data, it is robust to consistency motives that may arise in within-subject experiments. I extend the heterogeneous model to information choice and show that it accounts for observed patterns of information demand and avoidance on ethical domains.

1 Introduction

1.1 Background

Consequential decisions are often subject to the demands of principle. At the same time, complex situations provide some freedom to interpret a general principle or to decide which principle to prioritize. For instance:

*ridout@g.harvard.edu. For helpful comments and suggestions, I am indebted to Christine Exley, Ed Glaeser, Ben Golub, Jerry Green, Yoram Halevy, Peter Klibanoff, David Laibson, Shengwu Li, Yusufcan Masatlioglu, Efe Ok, Pietro Ortoleva, Fernando Payró Chew, and participants at an online decision theory conference in July 2020. For years of guidance, I am especially indebted to Matthew Rabin and Tomasz Strzalecki.

- A politician is charged with pursuing justice, but has some freedom in deciding which notion of justice to apply.
- A judge is charged with applying the law, but has some freedom in interpreting legal language and resolving conflicting precedents.
- A hiring manager is charged with choosing the most qualified candidate, but has some freedom in determining which qualifications matter most.

To capture these situations and others, this paper presents a two-tier model of preference maximization in which a "primary" preference resolves conflicts among "justifiable" preferences. The set of justifiable preferences captures the DM's notion of acceptable behavior in the domain at hand, while the primary preference captures the DM's own inclinations. On any choice set, the DM maximizes his primary preference over those alternatives that maximize at least one of the justifiable preferences. He thereby pursues his own objectives without doing anything obviously objectionable. For instance:

- The politician selects among the policies that could be implemented by a disinterested seeker of justice.
- The judge selects among the rulings that could be made by an impartial instrument of the law.
- The hiring manager selects among the candidates who could be ranked highest by an unbiased appraiser of talent.

The reader may wonder how an analyst limited to choice data could possibly disentangle the DM's underlying inclinations from his notion of acceptable behavior. Disentangling the two is indeed impossible if the DM never appeals to more than one justification. In that case, behavior is consistent with standard preference maximization, so a more general model is not required. The justification model becomes useful when the DM finds it optimal to appeal to different justifications in different situations, creating telltale inconsistencies with preference maximization.

The expansive "moral wiggle-room" literature provides evidence of these inconsistencies. We review this literature in Section 2, but provide one instructive example here. Exley (2016) offered subjects binary decisions involving (possibly random) payments to themselves and/or donations to a charity. She found that subjects' treatment of risk shifted in response to their selfish interests, resulting in cyclic choice patterns. For instance, subjects confronted with pairs of outcomes in

$$a = \text{experimenter pays \$2.50 to DM}$$
$$b = 50\% \text{ chance experimenter donates \$10 to charity}$$
$$d = \text{experimenter donates \$4 to charity}$$

might choose

$$a = c(\{a, b\}) \quad b = c(\{b, d\}) \quad d = c(\{a, d\}).$$

This pattern is inconsistent with preference maximization, but it can be generated by a selfish primary preference and a set of generous justifications with different risk attitudes. Intuitively, the DM chooses a over b, but not a over d, because choosing a over b can be attributed to risk aversion rather than selfishness. To see this in more detail, suppose the primary preference $\succsim$ is represented by

$$\mathbb{E}\left[3(\$ \text{ to DM}) + (\$ \text{ to charity})\right],$$

while the set of justifiable preferences $\mathcal{M}$ is represented by

$$\left\{\mathbb{E}\left[1.5(\$ \text{ to DM})^x + (\$ \text{ to charity})^x\right] : x \in [0.5, 2]\right\}.$$

The primary preference has $a \succ b \succ d$. The DM can justify choosing a over b because $a \succ_m b$ for sufficiently risk-averse $\succ_m \in \mathcal{M}$. He can justify choosing b over d because $b \succ_m d$ for the risk-loving $\succ_m \in \mathcal{M}$. However, he cannot choose a over d because $d \succ_m a$ for all $\succ_m \in \mathcal{M}$.

The justification model is intended as a general framework that can unite a wide variety of choice settings. To focus on the conflict between primary preferences and justifications, the model abstracts away from other interesting features of justifying behavior. First, it does not offer a theory of justice, rationality, or any other virtue or principle. While the analyst can learn about the DM's notion of acceptable behavior ex post, the model does not impose ex ante restrictions on that notion. An extension of the model allows for restrictions on the justifiable preferences, but leaves the analyst to determine what those restrictions should be.

Second, the model is agnostic about the DM's motivations for limiting himself to justifiable alternatives. One interpretation is that the DM is pretending to be a better person than he actually is. By limiting himself to alternatives that good people could select, he conceals his true inclinations from anyone observing his choice. The other interpretation is that the DM wants to remain within the bounds imposed by principle. He is not pretending to be a particularly virtuous person, but simply refraining from circumscribed actions. The reader may object that the latter interpretation is more plausible than the former, as a DM who appeals to different justifications in different situations reveals that his primary preference is not fully aligned with any of them. The only way to conceal one's primary preference, the objection continues, is to maximize a single justification across all decisions. This objection is most forceful when the DM must make several closely connected decisions in rapid succession before the same audience. If the connection between decisions is obscured, the time interval is long, or the set of observers varies, the DM may not attempt to maintain consistency across decisions. In any case, this objection applies only to versions of the justification model that rely on within-subject data. The random justification model of Section 5

uses between-subject data, so it accommodates both interpretations equally well.

1.2 Overview of results

The main results of this paper fall into two categories. First, the paper provides full behavioral characterizations of several variants of the justification model. With the exception of continuity conditions, which appear only in Section 3.3, all the axioms used in these characterizations are simple enough for practical application. For instance, they can be used to test whether apparently altruistic choices are consistent with a selfish primary preference, whether apparently irrational choices are consistent with *any* primary preference, and whether choices that fluctuate with the decision environment are consistent with a stable primary preference. Second, the paper provides tools for identifying both primary and justifiable preferences from choice data. For instance, the identification results can be used to disentangle the DM's true level of generosity from the minimal level he thinks acceptable, or to determine the range of allocation rules the decision-maker considers fair. In most cases, the identification results are byproducts of the behavioral characterizations, so the two groups of results are complementary. While the identification procedures are not always guaranteed to deliver full identification, they deliver all of the information contained in the data. Like the axioms, they rely on simple patterns of choice and remain useful on limited datasets.

The Justification model with Observable primary preference (JO) is introduced in Section 3. As the name suggests, this version of the model takes the primary preference as well as the choice correspondence as primitive. There are both expository and practical reasons for this assumption. On the expository side, the results for the Justification model with Unobservable primary preference (JU) build neatly on the JO results. On the practical side, the analyst may wish to test a particular candidate for the primary preference. Some choice settings may suggest a natural candidate (e.g. monetary self-interest). Alternatively, the analyst may be able to elicit a candidate by conducting a separate treatment in which subjects face less pressure to justify their decisions (e.g. because decisions are anonymous or implemented by someone else). Some experiments lend support to this idea. For instance, Hamman et al. (2010) found that subjects made blatantly selfish decisions when those decisions were carried out by an intermediary.

Theorem 1 is the representation result for JO. Given the primary preference, it provides necessary and sufficient conditions for an individual DM's behavior to be consistent with the justification model. The key axiom is Irrelevance of Unjustifiable Alternatives (IUA). Say that it is "unjustifiable" to choose alternative a from set A if a is not selected from A, but the primary preference likes a at least as much as everything that is selected. IUA says that a is irrelevant for choice on any superset of A if it is unjustifiable to choose a from A. The proof of Theorem 1 shows that a preference belongs to the maximal set of justifiable preferences if and only if it does not sanction any unjustifiable choices: the preference does not rank a above the rest of A if it is unjustifiable to

choose a from A. Therefore, the unjustifiable choices are fully informative about the DM's notion of acceptable behavior.

The remainder of Section 3 characterizes three extensions of JO. All three maintain the primary preference as a primitive; this is not dropped until Section 4. Section 3.2 allows the analyst to restrict the universe of possible justifications, ruling out preferences she considers obviously unjustifiable. The analyst's preconceptions are captured by an (observable) asymmetric and transitive relation on the domain that all justifiable preferences are required to respect. In addition to stochastic or Pareto dominance, an appropriately chosen relation can capture natural ethical requirements such as impartiality or nondiscrimination. Proposition 1 shows that a simple strengthening of IUA is necessary and sufficient for a representation in which all justifications respect the desired restrictions.

Section 3.3 extends JO to require continuity of the justifications. Continuity is desirable because it ensures the existence of utility representations, but it raises significant technical difficulties. The proof of Theorem 2, the representation theorem for the continuous case, addresses these difficulties with a continuous version of the Szpilrajn Extension Theorem from Herden and Pallack (2002).

The final extension of JO, in Section 3.4, takes a simpler route to utility representations. It restricts the domain to the set of lotteries on a finite prize space, and requires both primary and justifiable preferences to take an expected-utility form. Theorem 3 provides a behavioral characterization for the EU extension of JO. A nice feature of the EU extension is that behavior on an arbitrary choice set A is entirely pinned down by behavior on binary subsets of the convex hull of A. This makes the EU extension particularly tractable, so it can be used as a building block for other models. In fact, the random justification model in Section 5 is built up from the EU extension of JO.

Section 4 dispenses with the primary preference as a primitive and recovers it as a component of the representation. It provides two complementary characterizations for the Justification model when the primary preference is Unobservable (JU). These characterizations demonstrate that the justification model imposes substantial restrictions on behavior even if the primary preference is a "free parameter." The first characterization, Proposition 2, is a simple corollary to Theorem 1. As a byproduct, Proposition 2 provides a simple procedure for constructing the full set of primary preferences consistent with choice data. The second characterization, Theorem 4, is structurally similar to Theorem 1. Both results say that the justification model is characterized by irrelevance of unjustifiable alternatives, and both proofs show that the maximal set of justifications is precisely the set of preferences that do not sanction any unjustifiable choices. The only difference is the procedure for classifying some choices as "unjustifiable:" Theorem 1 relies on the primary preference, while Theorem 4 relies only on choice patterns. The required choice patterns are simple and easy to spot. By picking out these patterns, the analyst extracts all of the information contained in the data about the DM's notion of acceptable behavior.

Section 4.3 extends JU to account for evidence that justifying behavior is sensitive to features of the decision environment, such as anonymity (Charness and Gneezy, 2008; Franzen and Pointner, 2012) or whether anyone else stands to be disappointed by one's decision (Dana et al., 2006). It is plausible that changes in behavior across environments are driven by variations in the standard for acceptable behavior rather than shifting inclinations. Proposition 4, which builds on Theorem 4, allows the analyst to test this conjecture. It provides necessary and sufficient conditions for a pair of choice functions to be consistent with the same primary preference, but nested sets of justifications. A special case of Proposition 4, spelled out in Corollary 4, ties neatly back to Theorem 1. If (c_1, c_2) satisfies the conditions in Proposition 4, and c_1 is consistent with preference maximization, then c_2 satisfies IUA conditional on the preference maximized by c_1. Thus, it does not matter whether the analyst uses private choice (for example) to elicit the primary preference and subsequently applies Theorem 1 to public choice, or whether she applies Proposition 4 to public and private choice together.

Section 5 is motivated by the concern that some DMs may limit themselves to justifications consistent with their previous choices. As mentioned above, DMs who wish to conceal their primary preferences have a strong incentive to maintain consistency in justifications, since appeals to incompatible justifications produce the choice patterns that drive identification. Section 5 obviates consistency concerns by studying a large population of DMs, each of whom makes only one choice (as in a between-subject experiment). The primitive is a stochastic choice function, and the representation is a pair of distributions: one over primary preferences, and the other over sets of justifications. As in Section 3.4, the domain is a set of lotteries, and both primary and justifiable preferences have an EU form. The model permits arbitrary heterogeneity in primary preferences within the EU framework, but requires the sets of justifications to exist on a spectrum from strict to permissive. Intuitively, DMs within the same population are required to have similar principles, but may differ in their commitment to those principles.

Just as choice in the deterministic justification model violates the Weak Axiom of Revealed Preference (WARP), choice in the Random Justification model (RJ) violates Regularity, the stochastic analogue of WARP. Regularity says that the probability of selecting any given item must weakly decline when more items are added to the choice set. Proposition 5 shows that Regularity violations are ubiquitous in RJ. If it is sometimes unjustifiable to choose q over p, then it is always possible to find nested menus with p and q in their intersection such that the probability of choosing p is larger on the larger menu.

The main result for RJ is Theorem 5, which establishes uniqueness of the random justification representation. This result is important because it shows that an analyst with enough data can separately identify the distribution of primary preferences and the distribution of sets of justifications, even though she cannot tell whether any *particular* decision in her data set was driven by inclination or principle. The proof of Theorem 5 provides an explicit identification procedure. It shows that

the analyst can fully identify the justification distribution using choice sets of manageable size (no more than four elements).

In both the deterministic and random justification models, the DM's only degree of freedom comes from disagreement between justifications. A range of experiments, reviewed in Section 2, suggest that richer choice settings can provide additional degrees of freedom. Information avoidance is a major area of study in this literature, so Section 5.2 extends RJ to a simple information choice setting. Propositions 6 and 7 together show that RJ can predict information avoidance, but only if the standards that govern information acquisition are weaker than the standards that govern choice between final outcomes. If the same standards govern both types of choices, the DM will either feel unable to avoid information, or find it unprofitable to do so. The simplest way to generate information avoidance is to assume that information choice does not need to be justified at all. This assumption is adopted in the rest of Section 5.2, but relaxed in Appendix B.4.

Although Proposition 7 predicts information avoidance, it does not imply that information provision is entirely futile. Some (but not all) DMs with less-than-virtuous primary preferences will voluntarily acquire information that induces them to behave more virtuously. This is not because they feel obligated to become informed, but because information reduces the risk of sacrificing for nothing. For instance, a selfish DM might research a donation opportunity to make sure that his money is well spent when he feels compelled to donate. More generally, Proposition 7 shows that information can be used to bring behavior into alignment with principle, even if information choice does not have to be justified and remaining ignorant is always an option. The qualifier "judiciously chosen" is important, though. The final result of the paper, Proposition 8, shows that injudiciously chosen information can worsen behavior by allowing DMs with less-than-virtuous primary preferences to exploit disagreement between justifications. For instance, a biased hiring manager might treat disfavored candidates' resumes as a source of justifications for rejecting them.

Section 6 relates the justification model to existing theoretical work. Formally, the justification model is a generalization of the models of willpower in Masatlioglu et al. (2020) and dynamic choice in Strotz (1955), as formalized by Gul and Pesendorfer (2005). It is a special case of the models of limited attention in Masatlioglu et al. (2012) and of rationalization in Cherepanov et al. (2013). The latter deserves special attention because it is motivated by some of the same evidence as the justification model, and the two models are interpreted along similar lines. The key difference is that justifications are preferences, while rationales are completely unstructured binary relations. Thus, justifiers can be viewed as pooling with more virtuous types, while rationalizers cannot. Moreover, the additional structure of the justification model leads to strictly stronger identification than the rationalization model on any dataset that violates WARP.

2 Empirical motivation

Although the formal models introduced in this paper can be applied to non-moral domains, the vast majority of relevant empirical research covers moral decision-making, particularly tradeoffs between oneself and others. This work can broadly be divided into two strands. The first strand shows that subjects appeal to different principles, or pretend to have different tastes, when their personal interests change. The second strand provides evidence that subjects attempt to evade the constraints imposed by principle, e.g. by avoiding information about the effects of their actions.

Loewenstein et al. (1993) is a classic reference in the first strand. Each subject was assigned the role of plaintiff or defendant in a tort case involving a motorcycle accident. After reviewing the case material, each subject reported two values: the amount she expected the judge to award to the plaintiff, and the amount she considered fair. On average, subjects assigned to the role of plaintiff expected the judge to award $\$39,000$, and considered $\$37,000$ to be fair. Subjects assigned to the role of defendant expected $\$24,000$, and considered $\$19,000$ to be fair. Loewenstein et al. (1993) interpreted the gap between plaintiff and defendant as evidence of "self-serving assessments of fairness."[1]

Rodriguez-Lara and Moreno-Garrido (2012) studied a similar tension between disinterested allocation rules and personal gain. In an initial phase of the experiment, subjects earned money by completing a multiple-choice test. After the test was complete, each subject's correct answers were converted into money at a random "wage," which was fixed within-subject but could differ between-subject. Finally, subjects were paired up, and one member of each pair (the "dictator") allocated the money earned by the pair. Rodriguez-Lara and Moreno-Garrido (2012) found that dictators' choices were well explained by self-serving choice between three impartial allocation principles: egalitarian, accuracy-based, and earnings-based. Dictators who had low wages and accuracy tended to favor an egalitarian division, while dictators who had high wages and accuracy tended to favor division on the basis of earnings. Dictators who had low wages but high accuracy favored division on the basis of accuracy alone (correcting for the unequal "wages" assigned by the experimenter). We return to Rodriguez-Lara and Moreno-Garrido (2012) in Example 1 of Section 3.1, and to the notion of impartiality at the end of Section 3.2.

Norton et al. (2004) studied hiring decisions in the presence of gender bias. Subjects were asked to choose between a male and a female candidate for a traditionally male role. In one treatment, the male had more education and the female had more experience; in the other, these attributes were flipped. Subjects chose the male candidate a majority of the time in both cases. When asked to explain their decisions, very few subjects mentioned gender, instead citing the attribute (education or experience) in which their preferred candidate was superior. The proportion of subjects who said that they considered education more important than experience dropped from 50% when the male

[1] Self-serving assessments had real effects on behavior. When plaintiffs and defendants were paired up to negotiate a settlement, pairs with more divergent assessments were more likely to reach an impasse.

was more educated to less than 25% when the female was more educated. We return to Norton et al. (2004) at the end of Section 5.2.

Gneezy et al. (2019) took a different approach to showing that principles shift with self-interest. They conjectured that subjects who had already thought about or expressed their principles would find it harder to reshape them when their interests changed. To test this, they placed subjects in the role of investment advisors. Each advisor chose an asset to recommend to another subject, who was not informed about any of the assets. The key feature of the experiment was a small commission for recommending one of the assets. In line with the self-deception hypothesis, Gneezy et al. (2019) found that the bribe mattered if and only if two conditions were met. First, advisors had to learn about the bribe before they had a chance to decide which asset was best. Second, there had to be some ambiguity in the task: advisors did not accept a bribe to recommend a strictly dominated asset. Gneezy et al. (2020) found similar results for a different task, in which subjects were told to select the funniest joke from a set of jokes written by others.

Haisley and Weber (2010) tested the same idea in a different setup. Subjects faced a choice between an equitable allocation that gave moderate payments to the subject and a stranger, and an inequitable allocation that gave a large prize to the subject and a chance of a small prize to the stranger. As in Gneezy et al. (2019), ambiguity increased the scope for selfishness. Subjects were much more likely to choose selfishly when the chance of the small prize was uncertain rather than fixed at 50%. Also as in Gneezy et al. (2019), subjects were constrained by evaluations they had already made. Some subjects faced a choice between a risky lottery and ambiguous prospect earlier in the experiment. These subjects did not choose more selfishly in the ambiguous case. Haisley and Weber (2010) concluded that these subjects were unable to inflate the value of the ambiguous prospect because the vast majority of them had already expressed ambiguity aversion. We return to the consistency motives demonstrated in Haisley and Weber (2010) and Gneezy et al. (2019) at the beginning of Section 5, which motivates the random justification model.

We now proceed to the second strand of the literature, which shows that subjects take advantage of opportunities to evade the constraints imposed by principle. Most papers in this literature study information avoidance. The classic example is Dana et al. (2007), in which about 40% of subjects blatantly avoided learning how their choice would affect another subject's payoffs. More recent papers extend these findings to a wide variety of choice settings. Kajackaite (2015) found that one-third of subjects avoided learning how their actions would affect contributions to a negatively perceived lobbying organization. Serra-Garcia and Szech (2019) found that a majority of subjects chose not to learn about an opportunity to donate to a charity, and that almost half paid to avoid information. Woolley and Risen (2018) found that a majority of subjects preferred not to learn the calorie content of a tempting dessert, or the monetary bonus for a boring task. Ehrich and Irwin (2005), d'Adda et al. (2018), and Freddi (2017) found evidence of information avoidance about the ethical attributes of consumer goods, the environmental impacts of air conditioning, and the

refugee crisis, respectively. On the other hand, Fong and Oberholzer-Gee (2011) found that about a third of subjects facing a self-other allocation decision paid a substantial fee to learn whether the other subject was "deserving." We return to Fong and Oberholzer-Gee (2011), and to information choice more broadly, in Section 5.2.

A few papers study other ways in which subjects avoid feeling compelled to behave morally. Dana et al. (2006) found that about a third of subjects paid a fee to exit a dictator game. From a purely financial point of view, exiting is a dominated choice. It makes sense only for dictators who would feel compelled to give more than they would like if they remained in the game. Exiting is attractive to these dictators because the recipient does not observe the exit decision, so the sense of obligation governing this decision is presumably weaker than the sense of obligation governing the game itself. In a related field experiment, Andreoni et al. (2017) found that subjects went out of their way to avoid being asked to donate.

Hamman et al. (2010) studied delegation as a way to avoid responsibility. Each subject had the opportunity to hire an agent to make a self-other allocation on the subject's behalf. Subjects typically favored agents who had a record of sharing very little, or who announced the intention to share very little. This drove down sharing relative to a control condition in which subjects did not have the opportunity to delegate. We return to the effects of the choice setting on standards for acceptable behavior in Section 4.3.

3 Primary preference observable

This section introduces the Justification model with Observable primary preference (JO). Formally, JO does not place any restrictions on the domain of choice, $\mathcal{A}$. That said, JO is primarily intended for choice settings with ties to ethics, virtue or law. JO delivers interesting predictions on domains in which the "right" choice is not always obvious, but some choices are outright unacceptable.

JO has two primitives. The first is the primary preference $\succsim$, which is a complete and transitive relation on $\mathcal{A}$. Intuitively, it is what the DM would choose in the absence of any need to justify his decision. Formally (as the representation theorem will show), it breaks ties between justifications. The paper does not rest on the assumption of observable $\succsim$, which is dropped in Section 4. The second primitive is a choice correspondence c that maps each non-empty finite set of alternatives to a non-empty subset. To formalize this, let $\mathcal{F}(\mathcal{A})$ be the set of non-empty finite subsets of $\mathcal{A}$. Then we have $c : \mathcal{F}(\mathcal{A}) \rightrightarrows \mathcal{F}(\mathcal{A})$ such that $c(A) \subseteq A$ for all $A \in \mathcal{F}(\mathcal{A})$.

Definition 1 presents the JO representation, which is a set $\mathcal{M}$ of complete, transitive and antisymmetric orders on $\mathcal{A}$.[2] Intuitively, the elements of $\mathcal{M}$ are the set of preferences that the DM considers justifiable. These preferences are not required to be continuous, so they are not guaranteed to have utility representations. However, readers who prefer to think in terms of utility

[2]There is no loss of generality in taking the justifications to be antisymmetric.

functions will not lose anything by doing so. A continuous extension is presented in Section 3.3.

Definition 1 (JO Representation)**.** *A JO representation for* $(\succsim, c)$ *is a nonempty set* $\mathcal{M}$ *of strict preferences such that, for all* $A \in \mathcal{F}(\mathcal{A})$*,*

$$c(A) = \arg\max\left(\mathcal{M}(A), \succsim\right)$$
$$\text{where } \mathcal{M}(A) = \bigcup_{\succ_m \in \mathcal{M}} \arg\max\left(A, \succ_m\right).$$

3.1 Characterization

JO is fully characterized by a pair of axioms. The first one, Optimization, says that the DM is truly indifferent between all the items he actually selects. Although this is a standard assumption, it is possible to imagine morally-motivated DMs who violate it. For instance, a DM might feel that it is acceptable to select a selfish alternative as long as he also selects an unselfish one. This behavior is not captured by the justification model. In any case, Optimization has no bite when c is a choice function rather than a choice correspondence.

Axiom 1 (Optimization)**.** *For any* $A \in \mathcal{F}(\mathcal{A})$*, for any* $a, b \in c(A)$*:* $a \sim b$*.*

The second axiom, Irrelevance of Unjustifiable Alternatives (IUA), is the heart of the model. Consider a menu A and an item $a \in A$. If the primary preference likes a at least as much as everything that is selected from A, but a is not itself selected, then it must be unjustifiable to choose a from A. (Otherwise, the DM would have chosen it.) IUA says that a is irrelevant for choice on any superset of A.

Axiom 2 (Irrelevance of Unjustifiable Alternatives (IUA))**.** *For any* $a \in \mathcal{A}$ *and* $A \in \mathcal{F}(\mathcal{A})$ *such that* $a \in A$*: if* $a \succsim c(A)$ *and* $a \notin c(A)$*, then for all* $B \supseteq A$*,* $c(B \setminus \{a\}) = c(B)$*.*

Example 1, loosely based on Rodriguez-Lara and Moreno-Garrido (2012), clarifies the restrictions that IUA imposes on choice data.

Example 1. *Suppose the DM must allocate* \$12 *that he and two other subjects (A and B) earned in an earlier phase of the experiment. The DM earned* \$4*, while A and B earned* \$6 *and* \$2 *respectively. Let*

$$a = (5 \text{ to self}, 5 \text{ to } A, 2 \text{ to } B)$$
$$b = (4 \text{ to self}, 6 \text{ to } A, 2 \text{ to } B)$$
$$d = (4 \text{ to self}, 4 \text{ to } A, 4 \text{ to } B).$$

Suppose that the DM's primary preference is given by $a \succ b \succ d$*: he believes in rewarding other people's good performance, but above all wants to keep more for himself. IUA says that the DM*

cannot flip from choosing the performance-rewarding option b when all three options are present to choosing the egalitarian option d when the selfish option a is removed. Intuitively, the DM's notion of fairness cannot change when the attractive but unjustifiable option a is made unavailable.

To see why IUA is necessary for JO, suppose that it is unjustifiable to choose a from A. Toward a contradiction, fix some $B \supseteq A$, and suppose that $c(B) \neq c(B \setminus \{a\})$. Since the primary preference is fixed, the set of justifiable alternatives must be changing. Specifically, some alternative $b \in B$ must be unjustifiable when a is present, but justifiable when a is absent. That is, there must be some justifiable preference that prefers a to b, but b to everything else. Since there is no justifiable preference that prefers a to everything else in B, we have the desired contradiction.

Theorem 1 is the representation theorem for JO.

Theorem 1. $(\succsim, c)$ *has a JO representation if and only if it satisfies IUA and Optimization.*

The proof of Theorem 1 proceeds in two parts. The first part defines $\mathcal{M}$ and establishes that it is not too big: it does not justify any choice that the DM would like to make, but does not. The second part establishes that $\mathcal{M}$ is big enough: it justifies every choice that the DM actually makes.

The notion of "exclusion from below" is central to the proof. Intuitively, a menu A excludes an item b from below if the DM likes b better than everything in A, but does not choose b over A. A preference "respects exclusion from below" if it does not rank any item above a set that excludes that item from below.

Definition 2 (Exclusion from below)**.** $A \in \mathcal{F}(\mathcal{A})$ *excludes* $b \notin A$ *from below if* $b \succsim A$ *and* $b \notin c(A \cup \{b\})$*. A preference* $\succ_m$ *respects exclusion from below if for all* A, b *such that* A *excludes* b *from below, there exists* $a \in A$ *such that* $a \succ_m b$*.*

We define $\mathcal{M}$ to be the set of preferences on $\mathcal{A}$ that respect exclusion from below. To see why $\mathcal{M}$ is not too big, suppose that the DM would like to choose b from $A \cup \{b\}$, but does not: $b \succsim c(A \cup \{b\})$ and $b \notin c(A \cup \{b\})$. We need to show that $b \notin \mathcal{M}(A \cup \{b\})$. This is not entirely obvious because A might not exclude b from below: there might be items in A that are strictly better than b according to the primary preference. Fortunately, IUA says that any such item can be removed without changing choice:

$$c(A \cup \{b\}) = c(\{a \in A : b \succsim a\} \cup \{b\}).$$

Since b is not chosen over A, it is not chosen over $\{a \in A : b \succsim a\}$. Conclude that $\{a \in A : b \succsim a\}$ excludes b from below, so no preference in $\mathcal{M}$ ranks b above everything in $\{a \in A : b \succsim a\}$. We have $b \notin \mathcal{M}(A \cup \{b\})$ as desired.

It remains to show that $\mathcal{M}$ is big enough: it contains a justification for every choice the DM makes. This part is more involved. Fix an item b and a menu A such that $b \in c(\{b\} \cup A)$. For

simplicity, assume that $A = \{a\}$; this affects the argument very little. We need to find a preference $\succ_m$ that respects exclusion from below (so belongs to $\mathcal{M}$) and has $b \succ_m a$. We will construct an appropriate $\succ_m$ by carefully extending the "exclusion from below" relation.

Exclusion from below satisfies three convenient properties. It is irreflexive, meaning no menu containing x excludes x from below. It is proper, meaning $\emptyset$ does not exclude any item. Finally, it is transitive: if X excludes x from below, Y excludes y from below, and $x \in Y$, then $X \cup Y \setminus \{x, y\}$ excludes y from below. It turns out that a relation with these properties can be extended in a neat way, captured in the following Lemma. If $R \subset \mathcal{F}(\mathcal{A}) \times \mathcal{A}$ is irreflexive, proper and transitive, and if $\neg(y\ R\ x)$, then the transitive closure of $R \cup (\{x\}, y)$ is irreflexive and proper too.

The Lemma is used to show that exclusion from below can be extended to an irreflexive, proper and transitive relation R that has $\{b\}\ R\ a$ and has $\{x\}\ R\ y$ or $\{y\}\ R\ x$ for every distinct $x, y \in \mathcal{A}$. This is done in two steps. First, we take R_0 to be the transitive closure of the union of $(\{b\}, a)$ and exclusion from below. Since it cannot be that $\{a\}$ excludes b from below, the Lemma implies that R_0 is irreflexive and proper. Second, we follow the proof of the Szpilrajn Extension Theorem (SET) to extend R_0 to R. SET says that every irreflexive and transitive relation on a set X can be extended to an irreflexive and transitive relation that contains (x, y) or (y, x) for every distinct $x, y \in X$. Since R_0 relates menus to items, not items to items, and since such relations require a non-standard notion of transitivity, SET is not directly applicable. However, the arguments used to prove SET are easily adapted to deliver the desired R.

We use R to define $\succ_m$ in the natural way: for any distinct x, y, $x \succ_m y$ if and only if $\{x\}\ R\ y$. It is not difficult to see that $\succ_m$ is complete, antisymmetric and transitive, and satisfies $b \succ_m a$. To see why it respects exclusion from below, take any X, y such that X excludes y from below. Since R extends exclusion from below, we have $X\ R\ y$. Since R is transitive and proper, it cannot be that $\{y\}\ R\ x$ for all $x \in X$. Conclude that $\{x\}\ R\ y$ for some $x \in X$, so $x \succ_m y$ for some $x \in X$. We have found $\succ_m \in \mathcal{M}$ that justifies choosing b over a. With minor modifications for $|A| > 1$, we can apply the same argument to find an appropriate justification for each decision the DM makes. Thus, $\mathcal{M}$ is indeed a justification representation for $(\succsim, c)$.

3.2 Extension: restricting allowable constraints

JO abstracts away from the details of the choice setting to focus on the conflict between primary preferences and justifiable preferences. This allows the model to unite a range of disparate choice settings at a high level. However, particular applications may demand more structure. If the domain is a set of lotteries, a preference that violates first-order stochastic dominance cannot reasonably be considered justifiable. If the domain is a set of payments to a group of experimental subjects, a preference that violates Pareto dominance is presumably not justifiable. Some more interesting, but more involved, examples are deferred to the end of this section.

This section modifies Theorem 1 to exclude obviously unjustifiable preferences from the representation $\mathcal{M}$. The setting at hand determines which preferences count as "obviously unjustifiable." Specifically, the domain of choice is endowed with an (observable) asymmetric and transitive relation that embodies the basic requirements of rationality and/or morality on that domain. To remind the reader of the FOSD and Pareto examples, we refer to this relation as "dominance" and denote it $\succ_D$.

We will construct a JO representation in which all justifiable preferences are strictly monotone in $\succ_D$.

Definition 3 (Strict D-monotonicity)**.** *A relation $\succsim_R$ on $\mathcal{A}$ is strictly D-monotone if, for any $a, b \in \mathcal{A}$: $a \succ_D b$ implies $a \succ_R b$.*

Definition 4 (Monotone JO representation)**.** *A JO representation $\mathcal{M}$ is monotone if each $\succ_m \in \mathcal{M}$ is strictly D-monotone.*

Unsurprisingly, the key axiom for the monotone representation is a strengthening of IUA. It says that dominated items, as well as unjustifiable items, are irrelevant. To formalize this, let $S(B)$ be the members of B that are dominated or unjustifiable in B:

$$S(B) \;:=\; \bigcup_{b \in B} \{b' \in B : b \succ_D b'\} \cup \bigcup_{B' \in \mathcal{F}(B)} \{b' \in B' : \text{it is unjustifiable to choose } b \text{ from } B'\}.$$

The key axiom for the monotone representation, Irrelevance of Submaximal Alternatives (ISA), says that choice is unchanged when any subset of $S(B)$ is removed.

Axiom 3 (Irrelevance of Submaximal Alternatives (ISA))**.** *For any $B \in \mathcal{F}(\mathcal{A})$, for any $A \subseteq S(B)$: $c(B) = c(B \setminus A)$.*

The reader may wonder why ISA allows removal of several items, while IUA only allows removal of one item. Intuitively, this is because exclusion from below (defined in Section 3.1) satisfies a nice transitivity property, which allows us to remove unjustifiable items sequentially rather than all at once. This transitivity property doesn't hold once we introduce dominance, so we aren't always able to remove submaximal items sequentially. We have to explicitly allow removing multiple items at once.

Proposition 1 says that replacing IUA with ISA delivers a JO representation in which all the justifiable preferences respect dominance. The proof is along the same lines as that of Theorem 1, but the construction must now keep track of dominance as well as exclusion from below.

Proposition 1. *$(\succsim, c)$ has a monotone JO representation if and only if c satisfies ISA and Optimization.*

As promised, Example 2 shows how notions of disinterestedness or impartiality can be captured by a dominance relation.

Example 2.

1. *As in Gneezy et al. (2019), let the DM be an investment advisor, and $\mathcal{A}$ be a set of investments. Each investment is characterized by a distribution over payoffs p and a real number b. The real number is the bribe the DM will receive if he recommends that investment to his client. The following dominance relation is a natural way to formalize disinterestedness: $(p, b) \succ_D (p', b')$ if $p >_{FOSD} p'$.*

2. *As in Rodriguez-Lara and Moreno-Garrido (2012), let $\mathcal{A}$ be a set of allocations to n people. Subjects are indexed from least deserving (1) to most deserving (n). An allocation is $p \in \mathbb{R}^n_+$, where the ith entry is the payment to the ith subject. The following dominance relation is a natural way to formalize impartiality: $p \succ_D q$ if $p > \pi(q)$, where π is a permutation of $\{1, \ldots, n\}$ such that*

$$i \leq j \quad \Longrightarrow \quad q(\pi(i)) \leq q(\pi(j)).$$

Intuitively, p dominates q if it Pareto-dominates a reshuffling of q that gives larger payments to more deserving subjects.

3.3 Extension: continuity

If the domain $\mathcal{A}$ is uncountably infinite, utility representations are not guaranteed to exist for the justifiable preferences in the JO representation. This section extends Theorem 1 to require the existence of utility representations. This material is more technical than the preceding. An application-focused reader can safely skip to Section 3.4.

For this section alone, we take the domain $\mathcal{A}$ to be a separable metric space. Let $\mathcal{Z} = \{z_1, z_2, \ldots\}$ denote a countable dense subset of $\mathcal{A}$. A continuous JO representation is built from continuous utility functions on $\mathcal{A}$ rather than preferences on $\mathcal{A}$. Let $C(\mathcal{A}, \mathbb{R})$ denote the set of continuous functions from $\mathcal{A}$ to $\mathbb{R}$.

Definition 5 (Continuous JO representation)**.** *$(u, \mathcal{M}) \in C(\mathcal{A}, \mathbb{R}) \times 2^{C(\mathcal{A}, \mathbb{R})}$ is a continuous JO representation for $(\succsim, c)$ if u represents $\succsim$ and, for all $A \in \mathcal{F}(\mathcal{A})$,*

$$c(A) = \arg\max_{a \in \mathcal{M}(A)} u(a)$$
$$\text{where } \mathcal{M}(A) := \bigcup_{m \in \mathcal{M}} \arg\max_{a \in A} m(a).$$

In addition to continuity, the representation theorem imposes three technical conditions on $(u, \mathcal{M})$. All three conditions use the following bits of terminology. For any $u \in C(\mathcal{A}, \mathbb{R})$ and any

$A, B \in \mathcal{F}(\mathcal{A})$, say that A is strictly (weakly) preferred to B by u if

$$\max_{a \in A} u(a) > (\geq) \max_{b \in B} u(b).$$

For any $\mathcal{M} \subseteq C(\mathcal{A}, \mathbb{R})$ and any $A, B \in \mathcal{F}(\mathcal{A})$, say that A is strictly (weakly) preferred to B by $\mathcal{M}$ if, for all $m \in \mathcal{M}$, A is strictly (weakly) preferred to B by m.

The three technical conditions are closedness, local non-satiation and recoverability. Closedness is a continuity-like condition for sets of utilities. A *finite* set of utilities is guaranteed to be closed if all its members are continuous. This implication does not hold for infinite sets of utilities, so closedness has to be imposed separately.

Definition 6 (Closed)**.** *$\mathcal{M} \subseteq C(\mathcal{A}, \mathbb{R})$ is closed if, for all $B \in \mathcal{F}(\mathcal{A})$, the set*

$$\{a \in \mathcal{A} : B \text{ is strictly preferred to } a \text{ by } \mathcal{M}\}$$

is open.

Like closedness, local non-satiation extends a familiar condition to a set of utilities. A set of utilities is locally non-satiated if, for any item a, we can find a menu Z arbitrarily close to a such that all utilities in the set strictly prefer Z to a. Notice that the utilities in the set do not have to agree on *which* item in Z is better than a. For technical reasons, we require all elements of Z to be in the countable dense subset $\mathcal{Z}$.

Definition 7 (Locally non-satiated)**.** *$\mathcal{M} \subseteq C(\mathcal{A}, \mathbb{R})$ is locally non-satiated if, for any $a \in \mathcal{A}$, there exists $Z \in \mathcal{F}(B_\epsilon(a) \cap \mathcal{Z})$ such that Z is strictly preferred to a by $\mathcal{M}$.*

The final condition, recoverability, is the least familiar. A continuous JO representation is recoverable if the justifiable utilities agree on ranking B (strictly) above a only if the primary utility ranks a (weakly) above B. Intuitively, the justifiable utilities do not prevent the DM from choosing a over B unless the DM would actually like to do so. This restriction is not as strong as it may seem. Recall that a DM who wishes to choose a over B only needs one justifiable utility to justify his choice. He doesn't care whether all justifiable utilities rank a over B, or only some do. Thus, agreement in the justifiable utilities has no effect *unless* it is in opposition to the primary utility.

This argument suggests that one can impose recoverability without loss of generality. This is true when continuity is not required; the representation constructed in the proof of Theorem 1 is recoverable.[3] When continuity is required, recoverability is restrictive—but only slightly. Any $(\succsim, c)$ with a "nice" continuous representation has a recoverable quasi-representation that makes

[3]Of course, there may not be utility representations in that case—but recoverability can easily be expressed in terms of preferences rather than utilities.

the right predictions on all menus without ties.[4] For the interested reader, the formal result is Proposition 9 in Appendix B.1.

Definition 8 (Recoverable). $(u, \mathcal{M}) \in \left(C(\mathcal{A}, \mathbb{R}), 2^{C(\mathcal{A},\mathbb{R})}\right)$ *is recoverable if, for every* $(B, a) \in \mathcal{F}(\mathcal{A}) \times \mathcal{A}$,

$$B \text{ is strictly preferred to } a \text{ by } \mathcal{M} \implies \{b \in B : u(b) \leq u(a)\} \text{ is strictly preferred to } a \text{ by } \mathcal{M}.$$

The continuous representation is characterized by four axioms. Optimization is already familiar. C-IUA and IUA are very similar, but C-IUA accounts for additional information that can be gleaned about the justifiable preferences when they are required to be continuous. To see how this works, fix any menu A, and let $W(A)$ be the set of items in $\mathcal{A}$ that are excluded from below by a subset of A. As established in Section 3.1, every justifiable preference must strictly prefer A to every item in $W(A)$. Now consider a sequence of items $a_i \to a$ and a sequence of menus $A_i \to A$ such that $a_i \in W(A_i)$ for all i. Since every justifiable preference strictly prefers A_i to a_i for all i, every justifiable preference must weakly prefer A to a. (This is the step that uses continuity.) If every justifiable preference strictly prefers a to b, then every justifiable preference strictly prefers A to b. Thus, b must be irrelevant for choice when A is present. C-IUA generalizes this logic.

Formally, let

$$W(A) := \bigcup_{A' \subseteq A} \{a \in \mathcal{A} : a \succsim A' \text{ and } a \notin c(\{a\} \cup A')\}$$
$$\bar{W}(A) := \{a \in \mathcal{A} : \exists A_i \to A, a_i \to a \text{ s.t. } \forall i \; a_i \in W(A_i)\}.$$

Axiom 4 (C-IUA). *If* $D \in \mathcal{F}(\bar{W}(B))$ *for some* $B \in \mathcal{F}(W(A))$, *or if* $D \in \mathcal{F}(W(B))$ *for some* $B \in \mathcal{F}(\bar{W}(A))$, *then* $c(A) = c(A \setminus D)$.

The final two axioms, Continuity and Improvability, are the behavioral analogues of Closedness and Local Non-Satiation respectively. Both are integral to the construction of $\mathcal{M}$ in the sufficiency proof.

Axiom 5 (Continuity).

1. $\succsim$ *is continuous.*
2. *For all* $A \in \mathcal{F}(\mathcal{A})$, $W(A)$ *is open.*

Axiom 6 (Improvability). *For any* $a \in \mathcal{A}$ *and any* $\epsilon > 0$, *there is some* $Z \in \mathcal{F}(B_\epsilon(a) \cap \mathcal{Z})$ *such that* $a \in W(Z)$.

[4] Specifically, the quasi-representation predicts that $\{a \in A : a \sim c(A)\}$ is chosen from A. If no unchosen item is tied with $c(A)$, this is the right prediction. Otherwise, it is a superset of the right prediction.

Theorem 2. *The following are equivalent:*

1. $(\succsim, c)$ *satisfies Continuity, C-IUA, Improvability and Optimization.*
2. $(\succsim, c)$ *has a recoverable continuous JO representation* $(u, \mathcal{M})$ *such that* $\mathcal{M}$ *is closed and locally non-satiated.*

Just as the proof of Theorem 1 defines $\mathcal{M}$ to be the set of preferences that respect exclusion from below, the proof of Theorem 2 defines $\mathcal{M}$ to be the set of continuous utility functions that respect exclusion from below. Just as Theorem 1 constructs a preference in $\mathcal{M}$ to justify each decision the DM makes, Theorem 2 constructs a continuous utility function in $\mathcal{M}$ to justify each decision the DM makes. The proofs diverge after that. The proof of Theorem 2 appeals to a result in Herden and Pallack (2002) (HP), which provides sufficient conditions for an incomplete binary relation to have a continuous extension. The main part of the proof shows that a variant of the "exclusion from below" relation satisfies the HP conditions. (This step is non-trivial because the HP conditions are not easily checked, and bear no obvious relation to the axioms.) The HP theorem delivers a continuous utility representation for this relation, which implicitly defines a justifiable utility function.

3.4 Extension: expected utility

This section covers the expected-utility version of JO, in which the primary preference and all the justifiable preferences have expected-utility representations. A larger set of axioms is needed to achieve this additional structure, but the result is a tractable model suited to application. Additionally, the model in this section is the precursor to the random justification model in Section 5.

The set of prizes Z is assumed to be finite, although the domain $\mathcal{A} := \Delta(Z)$ is not. It is convenient to assume that there is a dominance relation $\succ_D$ on Z. As in Section 3.2, $\succ_D$ is asymmetric and transitive, but need not be complete. It is enough to have two payoffs ranked by dominance. For instance, Z could contain two monetary payments to the DM as well as payments to others, or donations to various charities. A version of the representation theorem is available without this assumption, but the axioms are slightly more cumbersome.[5]

We require the primary utility to be strictly D-monotone and the justifiable utilities to be weakly D-monotone. This assumption is motivated by convenience: weak monotonicity of the justifiable utilities falls out of the neatest set of axioms. Note that the DM will never actually *choose* a strictly dominated alternative. If the dominated alternative is justifiable, the dominating alternative will be too, and the DM will choose the latter but not the former.

[5] In the absence of dominance, the justifiable set may contain a preference exactly opposite the primary preference. This case is inconvenient and needs to be handled separately.

Definition 9 (D-monotonicity)**.** *A Bernoulli utility $u : Z \to \mathbb{R}$ is weakly (strictly) D-monotone if $a \succ_D b$ implies $u(a) \geq (>)\ u(b)$.*

We require the set of justifiable utilities to be compact and convex.

Definition 10 (Expected-utility JO representation)**.** *An expected-utility JO representation consists of a strictly D-monotone Bernoulli utility u and a compact, convex set $\mathcal{M}$ of weakly D-monotone Bernoulli utilities such that $p \mapsto \mathbb{E}_p u$ represents $\succsim$, and*

$$c(A) = \arg\max_{p \in \mathcal{M}(A)} \mathbb{E}_p u$$
$$\text{where } \mathcal{M}(A) := \bigcup_{m \in \mathcal{M}} \arg\max_{p \in A} \mathbb{E}_p m.$$

In the EU model, it is often easier not to work directly with $\mathcal{M}$, but with the set

$$B(p) := \{q \in \Delta(Z) : p \succsim q \text{ and } \{q\} = c(\{p, q\})\} \tag{1}$$

for some $p \in \text{int}(\Delta(Z))$. Intuitively, $B(p)$ is the set of lotteries that the DM feels compelled to choose over p, contrary to his primary preference. Formally, $B(p)$ is like a sufficient statistic: it encodes everything we need to know about the representation. It appears several times in the axioms.

Each of the four axioms in this section has two parts. For all axioms but Convexity, the first part is a condition on the primary preference, and the second part is a condition on choice behavior that ultimately translates into a condition on justifiable preferences. Since the conditions for a given preference to have a FOSD-monotone EU representation are well known, the first part is completely standard.

Axiom 7 (Independence)**.**

1. *For all $p, q, r \in \Delta(Z)$ and $\alpha \in (0, 1)$, $p \succsim q$ implies $\alpha p + (1-\alpha) r \succsim \alpha q + (1-\alpha) r$.*
2. *For any $A \in \mathcal{F}(\Delta(Z))$ and $p \in \Delta(Z)$,*

$$c(\alpha A + (1-\alpha)\{p\}) = \alpha c(A) + (1-\alpha)\{p\}.$$

The second part of Independence says that the DM's preference does not flip when every option he faces is mixed with a fixed lottery in a fixed proportion. The necessity of this axiom for an expected-utility JO representation is obvious. If p is the best item in A that is top-ranked by some justifiable preference, then mixing everything in A (including p) with q will not change this. Formally, this axiom ensures that $B(p)$ is a convex cone. The justifiable preferences are among the supporting hyperplanes of this cone.

Axiom 8 (Continuity)**.**

1. *For any* $p \in \Delta(Z)$*,* $\{q \in \Delta(Z) : q \succsim p\}$ *and* $\{q \in \Delta(Z) : q \precsim p\}$ *are closed.*

2. *For any* $p \in \Delta(Z)$*,* $B(p)$ *is open in* $\{q \in \Delta(Z) : q \succsim p\}$*.*

There is nothing unusual about the continuity axiom. If the second part fails, the set of justifiable preferences constructed in the proof will be slightly too permissive. Some preferences will be indifferent between pairs of items that they should strictly rank.

For the monotonicity condition, we need to extend first-order stochastic dominance (FOSD) to allow for an incomplete ranking over prizes. Notice that the definition provided here reduces to the usual one when $\succ_D$ is complete.

Definition 11 (FOSD)**.** $p >_{FOSD} q$ *if* $p = \sum_i \alpha_i \delta_{p_i}$*,* $q = \sum_i \alpha_i \delta_{q_i}$*,* $p_i \succ_D q_i$ *or* $p_i = q_i$ *for all* i*, and* $p_i \succ_D q_i$ *for some* i*.*

Axiom 9 (Monotonicity)**.**

1. $\succsim$ *is strictly FOSD-monotone.*

2. *For any* $p, q \in \Delta(Z)$ *such that* $p >_{FOSD} q$ *and any* $A \in \mathcal{F}(\Delta(Z))$ *containing* p *and* q*,* $c(A) = c(A \setminus \{q\})$*.*

The second part of Monotonicity is really Irrelevance of Dominated Alternatives (IDA). It says that any lottery is irrelevant in the presence of a lottery that strictly dominates it. This type of condition will be familiar from Section 3.2.

The final axiom, Convexity, is the least familiar. The first part says that it is unjustifiable to choose lottery p over menu A only if it is unjustifiable to choose p over some mixture of lotteries in A. The second part is a partial converse of the first. If it is unjustifiable to choose p over some mixture of lotteries in A, then p cannot be chosen over A.

Axiom 10 (Convexity)**.** *For any* $A \in \mathcal{F}(\Delta(Z))$ *and* $p \notin A$*:*

1. *If* $p \succsim c(\{p\} \cup A)$ *and* $p \notin c(\{p\} \cup A)$*, then* $co(A) \cap B(p) \neq \emptyset$*.*

2. *If* $co(A) \cap B(p) \neq \emptyset$*, then* $p \notin c(A \cup \{p\})$*.*

The word "mixture" is important. Convexity does not say that it is unjustifiable to choose p over A only if it is unjustifiable to choose p over some member of A. This is typically not the case. Unless the set of justifiable preferences is a singleton, we can find a lottery p and menu A such that it is justifiable to choose p over each member of A, but unjustifiable to choose p over A. Indeed, this inconsistency between choices in binary menus and choices in larger menus is part of what makes the justification model interesting.

The name Convexity arises because necessity of the first part follows from convexity of $\mathcal{M}$. To see what role Convexity plays in the proof of sufficiency, suppose that the DM prefers p to everything in A. Convexity implies that the DM will choose p over A if and only if $B(p)$ can be separated from A by an appropriately chosen hyperplane. This hyperplane will turn out to be the indifference curve of a justifiable preference.

Theorem 3 is the EU version of Theorem 1. The proof is quite different from that of Theorem 1; it uses mostly geometric rather than order-theoretic arguments. First, Independence and Continuity are used to show that $B(p)$ is always a convex cone, open in $\{q \in \Delta(Z) : p \succsim q\}$. Monotonicity is used to establish the relationship between $B(p)$ and the primary preference. $B(p)$ is used to identify a candidate set of justifiable preferences. Each candidate preference has an indifference curve that is a supporting hyperplane of $B(p)$. Finally, Convexity and the axioms from Theorem 1 are used to show that the candidate set is neither too large nor too small. No candidate preference would allow the DM to justify a choice he would have liked to make, but did not; and for each choice the DM actually makes, some candidate preference justifies that choice.

Theorem 3. *Suppose there exist $y, z \in Z$ such that $y \succ_D z$. The following are equivalent:*

1. *$(\succsim, c)$ satisfies IUA, Optimization, Independence, Continuity, Monotonicity and Convexity.*
2. *$(\succsim, c)$ has an expected-utility JO representation.*

The expected-utility JO model is subject to the usual uniqueness issues for EU representations: both the primary utility and the justifiable utilities can be arbitrarily (and independently) shifted and scaled. This issue can be dismissed by shifting attention from the set of justifiable utilities to the set of preferences they represent.

There is also a more subtle uniqueness issue: some EU preferences that are *not* positive affine transformations of other preferences can be added to or removed from the set of justifiable preferences without changing anything. Intuitively, this is because these preferences are too far from the primary preference to be of much use. Any choice they justify is also justified by some preference that is closer to $\succsim$. Corollary 1 says that we can obtain a unique representation (up to shifting and scaling) by dropping all these surplus preferences, retaining only those preferences needed to explain the DM's behavior. At the other extreme, we can include all the preferences not ruled out by the DM's behavior or by dominance.

Corollary 1. *If $(\succsim, c)$ has an expected-utility JO representation, it has a unique maximal and a unique minimal set of justifiable preferences.*

A nice feature of the EU model is the ease of comparing different DMs' standards for acceptable behavior. Since the proof of Theorem 3 uses $B(p)$ to obtain a set of justifiable preferences, we may expect a connection between the size of $B(p)$ and the strictness of the DM's standards. There is

indeed a connection. If $B_1(p)$ and $B_2(p)$ are nested, then we can find representations in which the sets of justifiable utilities are also nested. If $B_2(p)$ is the smaller set, then DM 2 can appeal to the larger set of justifiable utilities. This is convenient because we can figure out each DM's standards for acceptable behavior just by looking at his choices on a small set of menus: binary menus that contain a fixed item p. If DM 1 chooses a weakly inferior item q over p whenever DM 2 does, we can assume that DM 1's standards are at least as strict as DM 2's.

Corollary 2. *Suppose that $(\succsim, c_1)$ and $(\succsim, c_2)$ have expected-utility JO representations. The following are equivalent:*

1. *$B_1(p) \supset B_2(p)$ for some $p \in int(\Delta(Z))$.*
2. *The maximal set of justifiable preferences for $(\succsim, c_1)$ is strictly smaller than the maximal set of justifiable preferences for $(\succsim, c_2)$.*

The result is not true if "maximal" is replaced with "minimal." This is because there may be justifications that the more liberal DM never feels the need to use. He may agree that these are acceptable justifications, but he doesn't appeal to them because something better is always available. Thus, his minimal set of justifiable preferences doesn't include them.

Corollary 2 is connected to the random justification model of Section 5. There, we consider a population of DMs that can be totally ordered by the strength of their standards for acceptable behavior. The maximal/minimal distinction does not arise in the stochastic setup because heterogeneity in primary preferences eliminates non-uniqueness in sets of justifiable preferences. Intuitively, non-uniqueness persists when the primary preference is fixed because the sets of justifiable preferences are identified through conflict with the primary preference. If the DM prefers p to q and chooses p over q, we simply cannot tell whether he felt able to choose q. This problem goes away when the dataset contains multiple DMs with the same justifications but different preferences.

4 Primary preference unobservable

This section characterizes the Justification model when the primary preference is Unobservable (JU). For simplicity, we now take c to be a choice function rather than a choice correspondence.[6] A JU representation consists of a strict preference $\succ$ as well as a set of preferences $\mathcal{M}$.

This section provides two complementary behavioral characterizations of JU. The first characterization is a straightforward modification of Theorem 1. It shows that a single easy-to-check axiom is necessary and sufficient for a JU representation. It also provides insight into the DM's primary preference: it tells the researcher exactly which primary preferences are consistent with the DM's behavior, and which are not. However, it does not provide direct insight into the preferences

[6] This will be relaxed in the next draft.

the DM considers justifiable. The second characterization fills this gap. It shows how to identify the preferences that the DM may consider justifiable, and how to rule out the rest. Conveniently, the constraints on the justifiable preferences come from simple, easily recognized patterns of choice.

4.1 First characterization

This section builds directly on Theorem 1. Recall that IUA is a necessary and sufficient condition for a JO representation when c is a choice function. It says that a is irrelevant when A is present if $a \succ c(B \cup \{a\})$ for some subset B of A. We can flip IUA backward to derive conditions on the primary preference. To see how this works, suppose that choice from set A changes when item a is added. Clearly, it is not the case that a is irrelevant when A is present. Thus, it cannot be that $a \succ c(B \cup \{a\})$ for any subset B of A. Intuitively, if the addition of a affects choice on set A, then it is justifiable to choose a over A. If the DM fails to choose a over some subset B of A, he must be acting out of inclination rather than obligation. We can repeat this argument to obtain a full set of restrictions on $\succ$. If the resulting restrictions form a cycle, no strict preference can satisfy them, so there is no representation. But if the restrictions do not form a cycle, c satisfies IUA conditional on any strict preference that obeys the restrictions, so Theorem 1 delivers a representation. This is exactly what Proposition 2 says.

Definition 12 (Revealed Preference)**.** *If $a \neq c(B \cup \{a\})$ and, for some $A \supseteq B$, $c(A) \neq c(A \cup \{a\})$, then $c(B \cup \{a\})$ is revealed preferred to a.*

Axiom 11 (Acyclicity)**.** *The revealed preference relation for c is acyclic.*

Proposition 2. *c satisfies Acyclicity if and only if it has a JU representation. Moreover, a preference $\succ$ extends the revealed preference relation for c if and only if there is some $\mathcal{M}$ such that $(\succ, \mathcal{M})$ represents c.*

Corollary 2 is helpful not just because it delivers a representation, but because it reveals the set of primary preferences consistent with choice behavior. Sometimes, it pins down a unique preference. Example 3 illustrates.

Example 3. *Recall the example from Exley (2016) in Section 1:*

$$
\begin{aligned}
&a = \textit{experimenter pays } \$2.50 \textit{ to DM}\\
&b = 50\% \textit{ chance experimenter donates } \$10 \textit{ to charity}\\
&d = \textit{experimenter donates } \$4 \textit{ to charity}\\
&c(\{a,b\}) = a \quad c(\{b,d\}) = b \quad c(\{a,d\}) = d.
\end{aligned}
$$

Exley (2016) considered only binary menus, but choice from menus with more than two alternatives are important for identification in JU. We assume $b = c(\{a, b, d\})$ *since we have already seen a plausible* $(\succsim, \mathcal{M})$ *that generates this choice.*

Notice that choice from the grand set changes when d *or* b *is removed. Since* b *affects choice on a menu containing* a*, but* b *is not chosen over* a*, we must have* $a \succ b$*. Since* d *affects choice on a menu containing* d*, but* d *is not chosen over* b*, we must have* $b \succ d$*. Putting these two restrictions together, we get* $a \succ b \succ d$*. The DM prefers the risky donation over the safe donation, but above all prefers to keep more for himself.*

Conditional on any primary preference $\succ$, we can work out the set of justifiable preferences. We saw how to do this following Theorem 1. First, find each menu A and item $a \in A$ such that $a \succ A$ but $a \notin c(A \cup \{a\})$. Translate each of these pairs into a restriction on the justifiable preferences: every justifiable preference strictly prefers something in A to a. Finally, take the set of justifiable preferences to be the preferences that satisfy all these restrictions.

This process leaves something to be desired. It would be better to learn about the set of justifiable preferences just by looking at the data, not by constructing a set of primary preferences and then working out the constraints associated with each one. The next section explains how to do that, via an alternative characterization of the same model. The second characterization is more involved than the first one, but should still be of interest to the empirically inclined reader.

4.2 Second characterization

Two patterns of choice behavior are key to understanding the set of justifiable preferences. We define these patterns, explain their implications for the primary preference and the justifiable preferences, and leverage them to obtain another representation theorem for JU.

The first key pattern is a three-element cycle. As suggested in Example 3, the primary preference on any three-element cycle is uniquely pinned down. The item chosen from the full three-element set is middle-ranked, and the item that beats it is top-ranked. Since the bottom-ranked item beats the top-ranked item, it must be unjustifiable to choose the latter over the former. (In Example 3, it is unjustifiable to choose the safe payment a over the safe donation d.) Further restrictions on the justifications may be obtained by chaining together multiple cycles. For instance: if one cycle reveals a to be better than b, and another reveals b to be better than d, but d is chosen over a, it must be unjustifiable to choose a over d.

Definition 13 (Cycle/Chain)**.** (a_1, a_2, a_3) *is a cycle if*

$$c(\{a_1, a_2\}) = a_1 \quad c(\{a_1, a_2, a_3\}) = a_2 \quad c(\{a_1, a_3\}) = a_3.$$

For $k \geq 3$*,* $(a_1, \ldots, a_k)$ *is a chain if for each* $i \in \{2, \ldots, k-1\}$*,* (a_{i-1}, a_i, a_{i+1}) *is a cycle and/or*

both (a_{i-2}, a_{i-1}, a_i) *and* (a_i, a_{i+1}, a_{i+2}) *are cycles.*

The second key pattern of choice is an almost-WARP set. (The reasons for the name will soon become clear.) Fix some set A, and suppose that choice satisfies WARP on all its proper subsets. If $|A| = 3$, suppose further that pairwise choice is not cyclic. (Pairwise choice cannot be cyclic if $|A| > 3$.) Then, there is a unique preference on A that is maximized by choice from each proper subset. This preference is pinned down by pairwise choice. If choice on A violates WARP, so $c(A)$ is pairwise-defeated by some other item in A, the primary preference on A is uniquely pinned down. In fact, it is the preference given by pairwise choice. (This is not immediately obvious, but it is straightforward to prove.) Like cycles, almost-WARP sets are informative about the justifications as well as the primary preference. Since the item that pairwise-beats everything else in A is better than everything else in A, but is not chosen from A, it must be unjustifiable to choose that item from A.

Example 4. *This example is a continuation of Example 1, and uses the same notation. Suppose that the DM chooses the selfish allocation a over the performance-based allocation b, and separately chooses a over the equitable allocation d. Suppose further that the DM chooses b over d. If the DM selects b from the grand set $\{a, b, d\}$, then the grand set is almost-WARP. The DM's primary preference must be $a \succ b \succ d$, which is the preference given by pairwise choice. Moreover, a must be unjustifiable in the presence of $\{b, d\}$. Intuitively, the DM can choose a over b by pretending that he is averse to inequity, and he can choose a over d by pretending that he believes in rewarding performance, but he cannot do both at once.*

Definition 14 (Almost-WARP set)**.** *Suppose that A is not a cycle. A is an almost-WARP set if choice violates WARP on $\mathcal{F}(A)$, but satisfies WARP on $\mathcal{F}(A) \setminus A$.*

The implications of cycles and almost-WARP sets for justifiable preferences are summed up in Definition 15.

Definition 15 (Revealed exclusion)**.** *An item a is revealed excluded by a menu B if:*

1. *For $|B| > 1$: $B \cup \{a\}$ is an almost-WARP set, and a pairwise-defeats $c(B \cup \{a\})$.*
2. *For $B = \{b\}$: $b = c(\{a, b\})$, and a comes before b in a chain.*

Since three-element cycles and almost-WARP sets are easy to spot, so is revealed exclusion. One may wonder whether more restrictions on the justifiable preferences could be obtained from more complicated patterns of choice. The answer is no: cycles and almost-WARP sets tell the analyst all she could hope to know about the preferences the DM considers justifiable. Every preference that respects revealed exclusion appears in some JU representation. (In fact, there is a JU representation in which the set of justifiable preferences is *precisely* the set of preferences that respect revealed exclusion. We return to this point after the next representation result.) Proposition 3 summarizes.

Proposition 3. *Suppose c has a JU representation. For any menu A and item a such that $a \notin A$, the following are equivalent:*

1. *a is revealed excluded by a subset of A.*

2. *No justifiable preference in any representation ranks a above A.*

The next axiom is the analogue of IUA for the unknown-primary-preference case. It says that an item a is irrelevant for choice from B if a is revealed excluded by a subset of B.

Axiom 12 (Irrelevance of Excluded Alternatives (IEA))**.** *If each $a \in A \subset B$ is revealed excluded by a subset of B, then $c(B) = c(B \setminus A)$.*

To understand the restrictions imposed by IEA, consider an almost-WARP set A. By definition, there is a unique preference maximized by choice on the proper subsets of A. We can index the items in A from best to worst according to this preference: $a_1 \succ \cdots \succ a_n$. (As noted above, the primary preference agrees with the WARP-implied preference, hence the notation.) Since choice on A violates WARP, we can't have $c(A) = a_1$. It turns out we can only have $c(A) = a_2$—the item chosen from A is the DM's second-favorite item.[7] This means the DM's choice cannot deteriorate too quickly as we expand the choice set. He can move from always choosing his favorite item to choosing his second-favorite, but not to his third-favorite or worse. For instance, the DM cannot choose d from $\{a, b, d\}$ in Example 4.

Of course, IUA has no bite if nothing is revealed excluded. The reader may wonder how often cycles and almost-WARP sets actually arise. The answer is reassuring: unless choice satisfies WARP (in which case a standard preference-maximization model is perfectly adequate), there will be at least one cycle or almost-WARP set, so at least one item will be revealed excluded. This will provide an opportunity to falsify the model.

Theorem 4. *c satisfies IEA if and only if c has a JU representation.*

The proof of Theorem 4 constructs a particular JU representation for c, which we call the "canonical representation." This representation is notable because the set of justifiable preferences is precisely the set of preferences that respect revealed exclusion. Thus, the set of justifiable preferences is maximal: it includes the set of justifiable preferences from every other representation. Maximality of justifications corresponds to minimality of constraints: the more justifications are available to the DM, the fewer constraints he faces in making his decision. Therefore, the canonical representation for c is the most parsimonious model of constrained decision-making that explains c.

[7] Suppose $c(A) = a_3$. Then, a_2 is revealed excluded by $A \setminus \{a_2\}$. By IEA, we can remove a_2 from A without changing choice. But we know that $c(A \setminus \{a_2\}) = a_1 \neq c(A)$, so we have a contradiction. This argument generalizes to $i > 3$.

The primary preference in the canonical representation is easily constructed. First, impose $a \succ b$ whenever a comes before b in a chain. Then, if a and b have not yet been ranked, impose $a \succ b$ if $a = c(\{a, b\})$, and $b \succ a$ otherwise. Although this may not be the only preference consistent with behavior, it is the only preference consistent with the *maximal* set of justifiable preferences. Corollary 3 summarizes the properties of the canonical representation.

Definition 16 (Canonical representation)**.** *JU representation $(\succ^*, \mathcal{M}^*)$ is canonical if (1) $\mathcal{M}^*$ is the set of preferences that respects revealed exclusion, and (2) $\succ^*$ is the preference that has $a \succ b$ whenever a comes before b in a chain, and that agrees with pairwise choice on pairs not connected by any chain.*

Corollary 3. *Suppose c has a JU representation $(\succ, \mathcal{M})$. Then, it has a unique canonical representation $(\succ^*, \mathcal{M}^*)$, and $\mathcal{M}^* \supseteq \mathcal{M}$.*

4.3 Extension: comparing decision environments

Sections 4.1 and 4.2 cover identification in a single, static setting. Intuitively, we should be able to learn more about the primary preference and/or set of justifiable preferences by varying the pressure to find a good justification. For instance, we would expect the DM's choices to be closer to his primary preference when he chooses anonymously than when he must announce his choice to some ethically conscious peers. Several experiments find evidence that the choice setting matters in this way. As mentioned in Section 2, Hamman et al. (2010) find more selfish behavior when decisions are implemented by an intermediary, and Dana et al. (2006) find more selfish behavior when decisions are unobserved by those affected. Falk (2017) finds less selfish behavior when subjects are forced to watch themselves in a mirror, and Haley and Fessler (2005) find less selfish behavior among subjects who are "watched" by a pair of stylized eyespots.

To formalize the two-setting case, let c_L be the choice function corresponding to the low-pressure setting, and let c_H be the choice function corresponding to the high-pressure setting. We are now looking for a pair of JU representations with the same primary preference and nested sets of justifiable preferences. We accomplish this by building on Theorem 4. First, we ensure that c_L has a JU representation by requiring it to satisfy IEA. Second, we impose consistency between c_H and c_L. This consistency condition, IREA, is essentially a stronger version of IEA. It says that any item revealed excluded in the low-pressure setting is irrelevant in the high-pressure setting. This is clearly necessary for the set of justifications to be smaller in the high-pressure case. IREA also says that anything the DM chose in the low-pressure setting, but not in the high-pressure setting, is irrelevant in the high-pressure setting. Intuitively, this is because an item is replaced in the high-pressure setting only if it no longer counts as justifiable.

Definition 17 (Replacement)**.** *a is replaced in A if $a = c_L(A) \neq c_H(A)$.*

Axiom 13 (Irrelevance of Replaced or Excluded Alternatives (IREA))**.** *If each $a \in A \subset B$ is revealed excluded in L by, or replaced in, a subset of B, then $c_H(B) = c_H(B \setminus A)$.*

Example 5. *This example is very loosely based on Norton et al. (2004). Let*

$$m_1 = \textit{educated male applicant}$$
$$m_2 = \textit{experienced male applicant}$$
$$f_1 = \textit{educated female applicant}$$

Suppose there are two treatments, a low-pressure setting (L) in which the DM's choice is observed only by the experimenter, and a high-pressure setting (H) in which the DM's choice is observed by a female peer. Suppose that choice in the low-pressure setting is given by

$$c_L(\{m_1, m_2\}) = m_1 \quad c_L(\{m_2, f_1\}) = c_L(\{m_1, m_2, f_1\}) = m_2 \quad c_L(\{f_1, m_1\}) = f_1.$$

This is a cycle, so the DM must feel unable to choose the educated male applicant over the educated female applicant. Suppose further that $c_H(\{m_1, m_2, f_1\}) = f_1$. Since m_2 is replaced in $\{m_1, m_2, f_1\}$, IREA implies that

$$c(\{m_1, f_1\}) = c(\{m_2, f_1\}) = c(\{m_1, m_2, f_1\}) = f_1.$$

The DM now feels unable to choose either of the male applicants over the educated female applicant.

Proposition 4 is the representation result for the two-setting case. At the expense of additional notation, it could easily be extended to more than two settings.

Proposition 4. *c_L and c_H have JU representations $(\succ, \mathcal{M}^L)$ and $(\succ, \mathcal{M}^H)$ such that $\mathcal{M}^H \subseteq \mathcal{M}^L$ if and only if (c_L, c_H) satisfies IREA and c_L satisfies IEA.*

Proposition 4 also helps to interpret JO, the justification model with observable primary preference. It may seem mysterious for the analyst to observe a component of the representation. Corollary 4 explains what is really going on. Rather than directly observing the primary preference, the analyst observes choice behavior in a low-pressure situation. Provided this behavior satisfies WARP, she identifies the WARP-implied preference with the primary preference. Corollary 4 says there is little harm in this: if there is any JU representation, there is one in which the primary preference is the WARP-implied preference.

Corollary 4. *Suppose that c_L satisfies WARP, so the restriction of c_L to binary menus pins down a unique preference $\succ$. c_H has a JU representation if and only if it satisfies IUA conditional on $\succ$.*

5 Random justification model

In Section 4, WARP violations are used to reveal inconsistencies between the DM's preferences and his notion of acceptable behavior. Thus, the DM is ultimately unable to maintain the illusion that his preferences are beyond reproach. A sophisticated DM may recognize this danger and adjust his behavior accordingly, reducing or eliminating the WARP violations needed to identify the representation. This section addresses the problem by moving to a stochastic setup, in which each data point can be collected from a different DM.

The Random Justification model (RJ) builds on the deterministic EU model in Section 3.4. An RJ representation has two components: a distribution μ over primary preferences, and a distribution ν over sets of justifications. (Primary preferences and justifications are assumed to be drawn independently.) Just as in Section 3.4, both primary preferences and justifications have an EU form, and sets of justifications are assumed to be closed and convex. Formally, let $\mathcal{U}$ be the set of EU preferences on $\Delta(Z)$. Say that $U \subset \mathcal{U}$ is convex (closed) if

$$U_R := \bigcup_{\succsim \in U} \{u \in \mathbb{R}^Z : u \text{ represents } \succsim\}$$

is convex (closed). Let $\mathfrak{U}$ be the set of nonempty closed, convex subsets of $\mathcal{U}$.

Following Gul and Pesendorfer (2006), we assume that ties in the primary preferences happen with zero probability. Some of the results in this section also require μ to have full support.

Definition 18 (Preference distribution)**.** *$\mu \in \Delta(\mathcal{U})$ is a preference distribution if, for any distinct $x, y \in \Delta(Z)$, $\mu(\{\succsim : x \sim y\}) = 0$.*

For tractability, we restrict attention to a particular form of heterogeneity in justifications. The key assumption is that the support of ν is ordered by set inclusion, so the sets of justifications can be ranked from "most permissive" to "most strict." This assumption is appropriate for populations of decision-makers who agree on the values to be promoted, but disagree on the level of commitment needed to effectively promote those values.

Example 6. *This example is based on Fong and Oberholzer-Gee (2011). Subjects can make a small transfer to a poor person who suffers from a physical disability or a drug addiction. Subjects with any degree of generosity may choose a transfer to a disabled recipient over keeping the money. Especially generous subjects may choose the transfer regardless of recipient type. The remaining subjects may choose the transfer only if the recipient is sufficiently likely to be disabled.*

For technical reasons, we also assume that (1) a positive mass of DMs can justify anything, (2) the constraint distribution has no other mass points, and (3) the support of the constraint distribution has no gaps. These restrictions are formalized in Definition 19.

Definition 19 (Constraint distribution)**.** *$\nu \in \Delta(\mathfrak{U})$ is a constraint distribution if it satisfies the following conditions:*

1. *For any distinct $\mathcal{M}_1, \mathcal{M}_2 \in supp(\nu)$, there exists $\mathcal{M}_3 \in supp(\nu)$ such that*

$$\mathcal{M}_1 \subset int(\mathcal{M}_3) \subset \mathcal{M}_3 \subset int(\mathcal{M}_2) \quad or \quad \mathcal{M}_2 \subset int(\mathcal{M}_3) \subset \mathcal{M}_3 \subset int(\mathcal{M}_1).$$

2. *$\nu(\mathcal{U}) > 0$, and for each $t \in (\nu(\mathcal{U}), 1]$, there exists a unique $\mathcal{N} \in supp(\nu)$ such that*

$$\nu(\mathcal{M} : \mathcal{M} \supset \mathcal{N}) = t.$$

Once the primary preference $\succsim$ and set of justifications $\mathcal{M}$ are realized, a choice is made according to the deterministic EU model in Section 3.4.

Definition 20 (Random justification (RJ) representation)**.** *(μ, ν) is an RJ representation for ρ if μ is a preference distribution, ν is a constraint distribution, and*

$$\rho(p|A) = \int_{\mathcal{U}} \int_{\mathfrak{U}} \mathbb{1}\{p \in \arg\max(\succsim_m, \mathcal{M}(A))\} \, d\nu(\mathcal{M}) d\mu(u)$$
$$\text{where } \mathcal{M}(A) := \bigcup_{\succsim_m \in \mathcal{M}} \arg\max(\succsim_m, A).$$

5.1 Properties

Recall from Sections 3 and 4 that the deterministic justification models violate WARP: unchosen alternatives can affect choice by restricting the set of justifiable alternatives. It will not be surprising that RJ violates Regularity, the stochastic analogue of WARP.

Definition 21 (Regularity)**.** *For all $p \in \Delta(Z)$ and all $A, B \in \mathcal{F}(\Delta(Z))$,*

$$\rho(p|A) \geq \rho(p|A \cup B).$$

A simple type of Regularity violation recurs systematically within RJ. Example 7 illustrates.

Example 7. *Suppose that the DM can keep \$10 (p) or donate it to Charity A (q) or Charity B ($\tilde{q}$). Both charities represent good causes, neither of which is obviously more pressing than the other. In RJ, two groups of DMs choose q from $\{p, q, \tilde{q}\}$: DMs with $q \succ p, \tilde{q}$, and DMs with $p \succ q \succ \tilde{q}$ who feel unable to choose p. Now suppose that the donation to Charity B is implemented with error, so the DM is sometimes able to keep the money. Formally, $\tilde{q}$ is replaced with $\alpha\tilde{q} + (1-\alpha)p$, for α close to 1. DMs with $q \succ p, \tilde{q}$ continue to choose q, but some DMs with $p \succ q \succ \tilde{q}$ now choose $\alpha\tilde{q} + (1-\alpha)p$ when they feel unable to choose p. Intuitively, mixing virtuous option $\tilde{q}$ with p*

makes it more attractive to some DMs with less-than-virtuous preferences, reducing the probability of choosing the other virtuous option q.

This is not quite a Regularity violation because $\alpha\tilde{q} + (1-\alpha)p$ does not belong to the original menu $\{p, q, \tilde{q}\}$. However, $\alpha\tilde{q} + (1-\alpha)p$ can be added to that menu without changing choice. (In RJ, adding items in the convex hull of a menu never affects choice on that menu.) This delivers the required Regularity violation.

Definition 22 and the first part of Proposition 5 show that Example 7 generalizes. Say that q is more virtuous than p if it is sometimes unjustifiable to choose p over q, i.e. $\nu(\{\mathcal{M} : p \notin \mathcal{M}(\{p,q\})\}) > 0$. (Going forward, we abbreviate this as $\nu(p \notin \mathcal{M}(\{p,q\})) > 0$.) For any two lotteries p and q such that q is more virtuous than p, there is a pair of nested menus along the lines of Example 7 such that q is chosen more often from the larger menu. The intuition is always the same. When one virtuous option is implemented with error (so a DM who selects it sometimes ends up with a less virtuous option), constrained DMs are more likely to select it. They substitute away from the competing virtuous option, causing a Regularity violation.

The second part of Proposition 5 says that Regularity violations along the lines of Example 7 can be used to tell which of two lotteries is the more virtuous. Lottery q is more virtuous than lottery p if and only if the smallest $\mathcal{M}$ in the support of ν does not justify choosing p over q. Thus, Regularity violations can be used to identify this $\mathcal{M}$, which represents the strictest notion of virtue present in the population being studied.

Definition 22 (Anomalous)**.** *$(p, q) \in \Delta(Z)^2$ is anomalous if, for every $\epsilon > 0$, there exist $\tilde{q} \in B_\epsilon(q)$ and $\alpha, \lambda \in (0,1)$ such that*

$$\rho\left(q \middle| \{p, q, \alpha\tilde{q} + (1-\alpha)p, \lambda\tilde{q} + (1-\lambda)q\}\right) < \rho(q|\{p, q, \tilde{q}\}). \tag{2}$$

Proposition 5. *Suppose that ρ has an RJ representation (μ, ν), where μ has full support. For any $(p, q) \in int(\Delta(Z))^2$:*

1. *If $\nu\left(p \notin \mathcal{M}(\{p,q\})\right) > 0$, then (p, q) is anomalous.*
2. *If there exists $\epsilon > 0$ s.t. $(p, \tilde{q})$ is anomalous for all $\tilde{q} \in B_\epsilon(q)$, then $\nu\left(p \notin \mathcal{M}(\{p,q\})\right) > 0$.*

The second part of Proposition 5 is used as a Lemma in the proof of Theorem 5, which establishes the uniqueness of the RJ representation.

Theorem 5. *Any RJ representation (μ, ν) with full-support μ is unique.*

In the absence of Theorem 5, an analyst who suspects that her data is consistent with RJ can do only two things. She can show that the data includes Regularity violations, so it is inconsistent

with random EU. Second, she can try to assess the extent of the violations by computing

$$\rho(p|B) - \rho(p|A)$$

whenever this quantity is positive and $p \in A \subset B$. Theorem 5 shows that she can go beyond these simple bounds. For each menu, she can recover precisely the proportion of DMs who are maximizing their primary preferences, and she can tell what choice would be in a hypothetical world in which all DMs maximize their primary preferences.

The proof of Theorem 5 provides an explicit, relatively simple procedure for recovering the components of the RJ representation. It proceeds in three steps. First, for each $p \in \Delta(Z)$, we identify

$$D(p) := \arg\max_{x \in \Delta(Z)} \nu\left(x \notin \mathcal{M}(\{p, x\})\right).$$

The set $D(p)$ reveals the most permissive notion of acceptable behavior present within the society being studied, just as $\{x \in \Delta(Z) : \nu\left(x \notin \mathcal{M}(\{p, x\})\right) > 0\}$ reveals the strictest notion of acceptable behavior. It is easier to identify the rest of ν once these bounds are known.

We exploit Regularity violations to identify $D(p)$. If $q \in D(p)$, then q is the first item to become unjustifiable in any menu containing both p and q. Thus, a constrained DM can never choose q over p. Since Regularity violations come from DMs who face binding constraints, it is not possible to find menus A, B such that $\{p, q\} \subset A \subset B$ and $\rho(q|A) < \rho(q|B)$. This is always possible if $q \notin D(p)$, as we can construct menus containing p and q in which q is selected by some constrained DMs.

The second step identifies ν given the "bounds" from the first step. Like Proposition 5, it uses simple three- or four-element menus. The trick is to identify two groups of DMs who have exactly the same (distribution of) preferences, but different sets of justifications. Since preferences are held constant, any difference in behavior across the two groups must be attributed to justifications. This allows us to pin down $\nu(q \notin \mathcal{M}(\{p, q\}))$ for each pair of lotteries p, q. Since the support of ν is ordered by set inclusion, these probabilities pin down ν itself.

The final step identifies μ given ν. This is done by "correcting" $\rho(p|A)$ for the DMs who choose p because they are constrained, leaving only the DMs who genuinely prefer p. To see how this works, fix a three-element menu A, and index the elements from hardest-to-justify to easiest-to-justify. (Again, this is possible because the support of ν is ordered by set inclusion.) Identifying $\mu(p_1 \succsim A)$ is easy because the only DMs who choose p_1 from A are those who like p_1 best and are able to justify it. We have already identified the probability that p_1 is justifiable, so we can identify the probability that it is best. Identifying $\mu(p_2 \succsim A)$ is slightly more complicated because two groups of DMs choose p_2 from A. The first group consists of DMs who like p_2 best and are able to justify it. The second group consists of DMs who like p_1 better than p_2 better than p_3, and are able to justify

p_2 but not p_1. We already know the probability that p_2 is justifiable but p_1 is not. To identify $\mu(p_1 \succ p_2 \succ p_3)$, notice that DMs with this preference pattern substitute away from p_2 when p_1 is added to the choice set:

$$\rho(p_2|\{p_2, p_3\}) - \rho(p_2|A) = \mu(p_1 \succ p_2 \succ p_3)\nu(\{\mathcal{M} : p_1 \in \mathcal{M}(A))\}).$$

Deducting the DMs in the second group from $\rho(p_2|A)$ (and adjusting for the probability that p_2 is justifiable), we recover $\mu(p_2 \succsim A)$. We now know the full distribution of preferences on A. The procedure is essentially the same for larger menus, although it must account for additional groups of DMs.

5.2 Extension: information choice

5.2.1 Model of information

Some of the best-known evidence that people exploit "moral wiggle room" comes from experiments that allow subjects to acquire (or avoid) information before making a decision. Several experiments find that subjects fail to acquire, or even pay to avoid, information that would have affected their ultimate decision. The classic example is Dana et al. (2007), in which each DM was told that his interests could be aligned with or opposed to the interests of another subject. DMs who could not avoid learning the state, and found that the conflict state had been realized, nearly always chose unselfishly. DMs who could avoid learning the state did so about 40% of the time, and nearly always chose selfishly.

The results of Dana et al. (2007) raise two questions about the information demand of justifiers. First, can information be used to bring people's behavior into alignment with their own notions of virtue? Dana et al. (2007) found that it could, and the effects were diminished but not eliminated when people could choose not to look at the information. Proposition 7 says that these results generalize: in RJ, there are almost always opportunities to improve average behavior through information, although some individuals will avoid virtue-promoting information when they have the opportunity. Second, is virtue best promoted by encouraging people to take as much information as possible? Proposition 8 says that the answer is no: an injudicious choice of information may make behavior unambiguously worse by providing new justifications for bad behavior. This possibility is typically ignored in the literature on information and moral wiggle room, although there is empirical evidence of it. The generality of RJ is helpful here; a model tailored to a particular situation would not be able to capture both the "moral suasion" and "excuse" functions of information.

We now formally extend RJ to information choice. We restrict attention to a simple kind of information, designed to match a typical experimental setup. The DM faces a binary choice set. In the first stage, he is offered a single (binary) signal about one of the items in the set. If he accepts

the signal, he observes it before making his selection in the second stage. Definitions 23 and 24 formalize this setup.

Definition 23 (Signal)**.** $(q_1, q_2, \alpha) \in \Delta(Z)^2 \times [0,1]$ *is a signal about* $q \in \Delta(Z)$ *if* $q = \alpha q_1 + (1-\alpha) q_2$.

Definition 24 (Information choice problem)**.** *An information problem is*

$$\left\{\delta_{\{p,q\}}, \alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right\}, \tag{3}$$

where (q_1, q_2, α) *is a signal about* q*, and* $p, q \in \Delta(Z)$.

Perhaps the most natural way to extend RJ to information choice is as follows. Recall that a DM endowed with menu $\{p, q\}$ limits himself to

$$\mathcal{M}(\{p,q\}) = \bigcup_{m \in \mathcal{M}} \arg\max_{\{p,q\}} m.$$

This allows him to pool with DMs who have preferences in $\mathcal{M}$. Now consider a DM who faces information problem (3). If the DM chooses $\delta_{\{p,q\}}$, then he can only pool with DMs who have preferences in

$$\mathcal{M}^{\text{avoid}} := \left\{ m \in \mathcal{M} : \arg\max_{\{p,q\}} m = \alpha \arg\max_{\{p,q_1\}} m + (1-\alpha) \arg\max_{\{p,q_2\}} m \right\}. \tag{4}$$

He must therefore limit himself to $\mathcal{M}^{\text{avoid}}(\{p,q\})$ at the second stage. (If $\mathcal{M}^{\text{avoid}} = \emptyset$, the DM cannot choose $\delta_{\{p,q\}}$ in the first place, as doing so would automatically separate him from all DMs with preferences in $\mathcal{M}$.)

Unfortunately, Proposition 6 shows that this version of RJ cannot generate a strict preference for information avoidance. In some cases, the DM feels compelled to become informed because he believes every virtuous person would do so. In all other cases, he is free to forego information, but does not gain by it. Intuitively, the DM cannot use information to get closer to his second-stage ideal if information choice and second-stage choice are subject to the same constraints.

Proposition 6. *For any justification representation* $(u, \mathcal{M})$ *and any information problem* (p, q_1, q_2, α)*: either the set* $\mathcal{M}^{avoid}$ *given by (4) is empty, or*

$$\alpha \max_{\mathcal{M}(\{p,q_1\})} u + (1-\alpha) \max_{\mathcal{M}(\{p,q_2\})} u \geq \max_{\mathcal{M}^{avoid}(\{p,q\})} u.$$

To account for the empirical prevalence of information avoidance, the constraints on information choice must be relaxed. In the body of the paper, we adopt the simplifying assumption that information choice is fully unconstrained: the DM chooses information to maximize his expected

utility, given his beliefs about his own second-stage behavior. Appendix B.4 presents a more general model that allows information choice to be *less* constrained than second-stage choice without being fully unconstrained. As shown in the Appendix, the results in this section carry over to the more general model.

To complete the model, we still need to specify the DM's first-stage beliefs about his own second-stage behavior. We assume that he knows his primary utility in both stages, but may have imperfect information about his justifications until he reaches the second stage. This allows for the possibility that the set of justifications is partly influenced by passing sentiments, such as feelings of sympathy. We refer to DMs' first-stage information about their second-stage justifications as "self-knowledge."

Definition 25 (Self-knowledge)**.** *Self-knowledge about constraint distribution ν is $N \in \Delta^2(\mathfrak{U})$, where*

$$\int_{\tilde{\nu}} \tilde{\nu}(\cdot) dN_\nu(\tilde{\nu}) = \nu(\cdot).$$

Ties are inevitable in the first stage because the DM must be indifferent to information that does not affect his behavior. Since we are interested in a strict preference for information avoidance, we assume that each DM breaks ties in favor of becoming informed. Information "avoidance" that results from indifference is thereby ruled out. This makes no difference to the results.

We are now ready to extend RJ to information choice. Fix a preference distribution μ and a self-knowledge N. Let $U_{(u,\tilde{\nu})}$ denote the expected utility of a DM with primary utility $u \in \text{supp}(\mu)$ and beliefs $\tilde{\nu} \in \text{supp}(N)$. We use these expected utilities to define a stochastic choice function $\rho_{(\mu,N)}$ on the set of information choice problems:

$$\begin{aligned}\rho_{(\mu,N)}&\left(\delta_{\{p,q\}} \middle| \left\{\delta_{\{p,q\}}, \alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right\}\right) \\ &:= \int_u \int_{\tilde{\nu}} \mathbb{1}\left\{U_{(u,\tilde{\nu})}\left(\delta_{\{p,q\}}\right) > U_{(u,\tilde{\nu})}\left(\alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right)\right\} dN(\tilde{\nu}) d\nu(u).\end{aligned}$$

To assess the effects of information, it is helpful to have a standard for desirable behavior. To this end, we introduce a social planner who must decide how much information to provide to her society. The planner has her own Bernoulli utility $s \in \mathbb{R}^Z$, which need not match that of any DM. For instance, the planner could seek to maximize the expected value of charitable donations from a society of selfish DMs. For each dilemma $\{p, q\}$ confronting members of her society, the planner has three options: provide no information, design a signal about q and provide it for free, or design a signal about q and require everyone to observe it. The planner knows the preference distribution μ and self-knowledge N that characterize her society, and maximizes her expected utility $S_{(\mu,N)}$ given her (rational) expectations of behavior.

5.2.2 Information demand and avoidance

Despite the simplicity of the setup, RJ allows for complex attitudes to information. We use Example 8, which will recur throughout this section, to illustrate.

Example 8. *This is a simplified version of the experiment in Fong and Oberholzer-Gee (2011). Let*

$$p = \textit{experimenter pays } \$10 \textit{ to DM}$$
$$q_1 = \textit{experimenter pays } \$10 \textit{ to physically disabled poor person}$$
$$q_2 = \textit{experimenter pays } \$10 \textit{ to poor person with drug addiction}$$
$$q = \frac{1}{2}q_1 + \frac{1}{2}q_2.$$

Consider a DM who prefers p to q, but may feel that it is unjustifiable to choose p over q.

Information may affect the DM in Example 8 in two ways. First, she may feel able to keep the money more or less often on average. Second, the relative weights on the two recipient groups may change, as the DM may feel compelled to donate to one type of recipient more frequently than the other. The directions of these two effects may be hard to predict. Even if the directions are known, it may be hard to predict which effect will dominate and, by extension, whether the DM will choose to become informed. Finally, it may not be clear whether a virtuous social planner would want the DM to become informed in the first place. For instance, information might reduce total donations but raise donations to a favored recipient group.

In fact, experimental subjects do exhibit complex attitudes to information, and informational interventions often have mixed effects. The goal of this section is not to provide sharp predictions for all or even most information choice problems, but to identify two broad patterns of information choice within RJ. The first pattern is moral suasion: a social planner for a population of RJ agents can further her objectives just by informing people about an option in their choice sets. More specifically, a planner who prefers more virtuous option p to less virtuous option q can design a binary signal about q that she would like every DM who prefers q to p to acquire. Surprisingly, some DMs will voluntarily acquire appropriate information even though they disagree with the planner on the original choice set. The planner benefits from providing this information even if she cannot require anyone to pay attention to it.

Proposition 7. *Fix a constraint distribution ν, lotteries $p, q \in int(\Delta(Z))$ such that $\nu(\mathcal{U}) < \nu(q \in \mathcal{M}(\{p,q\})) < 1$, and a social planner s such that $s(p) > s(q)$. There exists a signal (q_1, q_2, α) about q such that*

$$S_{(\mu,N)}\left(\alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right) > S_{(\mu,N)}\left(\left\{\delta_{\{p,q\}}, \alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right\}\right) > S_{(u,N)}\left(\delta_{\{p,q\}}\right)$$

for any preference distribution μ *such that* $supp(\mu) = \{u \in \mathcal{U} : u(q) \geq u(p)\}$ *and any self-knowledge* N.

We use Example 8 to illustrate. Consider a social planner who cares only about expected donations to disabled recipients, so $s(q_1) > s(q) > s(q_2) = s(p)$. Plausibly, the obligation to donate to a disabled person is stronger than the obligation to donate to a drug addict, so

$$\nu(p \notin \mathcal{M}(\{p, q_1\})) > \nu(p \notin \mathcal{M}(\{p, q\})) > \nu(p \notin \mathcal{M}(\{p, q_2\})).$$

It is easy to see that the social planner is better off if the DM learns the recipient's type. If q_2 is realized, the planner doesn't care what the DM does; if q_1 is realized, the planner benefits because the DM is more likely to make a transfer than he would if he were ignorant. To see why some DMs will learn the recipient's type voluntarily, consider a DM who has a slight preference for p over q_1, but a strong preference for q_1 over q_2. Since this DM is primarily concerned with avoiding a transfer to a drug addict, he prefers to become informed even if he must sacrifice some chance of keeping the money. By contrast, a DM who does not care about the recipient's type will actively avoid any information that reduces his chance of keeping the money.

The findings of Fong and Oberholzer-Gee (2011) are aligned with the above, although they use a more complex setup in which DMs can transfer any part of \$10 to the recipient. Expected transfers to disabled recipients do rise from the no-information treatment (\$3) to the full-information treatment (\$4.30). In a third treatment, DMs could pay \$1 to learn the recipient's type before making a transfer decision. This setup does not quite match Proposition 7, in which information is freely available. Fong and Oberholzer-Gee (2011) found that two-thirds of DMs declined to pay for information and transferred \$2 on average, while the remaining one-third paid for information and transferred \$4.50. As a result, average transfers to disabled recipients were \$2.80, slightly lower than in the no-information case. This was likely because DMs passed most of the information cost on to the recipient. Fong and Oberholzer-Gee (2011) estimated that having \$9 rather than \$10 reduced transfers by \$0.80 on average. Correcting for this raises average transfers to disabled recipients to \$3.10, slightly higher than the no-information case. This is probably an underestimate because more DMs would have acquired information (and raised their donations to disabled recipients) if information had been free.

Many other experiments, including Ehrich and Irwin (2005) on ethical consumption, Dana et al. (2007) and Grossman and Van Der Weele (2017) on sharing with others, Serra-Garcia and Szech (2019) on donations to charity, Kajackaite (2015) on donations to a lobbying organization, and Woolley and Risen (2018) on task choice, study the effects of information about a possible negative externality to an appealing action. The informational interventions in these papers encourage virtuous behavior: the externality is mitigated in full-information treatments relative to no-information

treatments.[8] As predicted in Proposition 7, a substantial fraction of subjects seem to anticipate the effects of information, and avoid it when they have the chance. This erodes the benefits of the informational intervention.

Generalizing from these findings, one might conclude that information is a force for virtue: people behave better when they are better informed about the consequences of their actions. However, it is important to remember that the informational interventions discussed above were *designed* to promote virtuous behavior and provoke avoidance. The next result, Proposition 8, shows that information can have exactly the opposite effect. There are signals that every DM with less-than-virtuous preferences would like to acquire, but every social planner with virtuous preferences would like to withhold. This type of information is not limited to contrived examples; it exists whenever virtuous people disagree. Example 9 illustrates the formal notion of disagreement, which is given in Definition 26.

Example 9. *This example is based on Norton et al. (2004). The DM must hire someone for a managerial role at a construction company. Candidates are evaluated on both education and experience. Let*

$$\begin{aligned} p &= \textit{male candidate} \\ q_1 &= \textit{female candidate with more education, less experience than } p \\ q_2 &= \textit{female candidate with more experience, less education than } p \\ q &= \frac{1}{2}q_1 + \frac{1}{2}q_2. \end{aligned}$$

Since women have been historically underrepresented in this field, the DM may believe that an unbiased person would give female candidates a slight edge. This makes it unjustifiable to choose p over q, but not to choose p over q_1 or q_2. An unbiased person who considers education more important than experience might well choose p over q_2, while one who considers experience more important than education might choose p over q_1.

Definition 26 (Disagreement)**.** *Constraint distribution ν exhibits disagreement about (q_1, q_2, p) if there exists $\mathcal{M} \in supp(\nu)$ such that the sets*

$$\{m \in \mathcal{M} : m(q_1) \geq m(p)\} \text{ and } \{m \in \mathcal{M} : m(q_2) \geq m(p)\}$$

are nonempty and disjoint.

Proposition 8. *If constraint distribution ν exhibits disagreement about lotteries (q_1, q_2, p), there*

[8] Kajackaite (2015) is an exception: she finds that subjects are very conservative when they do not know whether the externality has been realized, so they act like subjects who know there is a negative externality.

exist $\alpha \in (0,1)$ *and* $q \in \Delta(Z)$ *such that* (q_1, q_2, α) *is a signal about* q*, and*

$$S_{(u,N)}\left(\delta_{\{p,q\}}\right) > S_{(\mu,N)}\left(\left\{\delta_{\{p,q\}}, \alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right\}\right) = S_{(\mu,N)}\left(\alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right)$$

for any social planner s *such that* $s(p) > \max\{s(q_1), s(q_2)\}$*, any preference distribution* μ *such that* $\mu(\{u \in \mathcal{U} : \min\{u(q_1), u(q_2)\} > u(p)\}) = 1$*, and any self-knowledge* N*.*

In Example 9, it is not difficult to see why information might have a pernicious effect. A DM who wishes to hire a male candidate regardless of qualifications may choose q over p to avoid revealing his bias. However, he can justify choosing p over q_1 by arguing that experience is more important, and p over q_2 by arguing that education is more important. This is exactly what Norton et al. (2004) find in their study of hypothetical hiring decisions. Subjects picked the male candidate 66% of the time (75% when the male candidate was more educated, and 57% when the male candidate was more experienced). Norton et al. (2004) also asked subjects to explain their decisions. When the male candidate was more educated (and in a control condition without gender), half of subjects mentioned that they considered education more important than experience. This proportion dropped below one-quarter when the female candidate was more educated, suggesting that some subjects used experience as an excuse to choose their favored candidate.

6 Related models

The justification model with unknown primary preference (JU) is closely connected to the model of attention in Masatlioglu et al. (2012) (MNO), and to the model of rationalization in Cherepanov et al. (2013) (CFS). Both models use a two-tier structure in which a preference breaks ties on a consideration set. MNO interpret the consideration set as the subset of items to which the DM pays attention. The key restriction in their model is that removal of an item outside the consideration set does not change choice. This property holds in JU as well, so JU is a special case of MNO. CFS interpret the consideration set as the subset of items the DM can rationalize. The difference between CFS and JU is that CFS take rationales to be unstructured binary relations rather than preferences. Thus, JU is a special case of CFS. Since CFS can impose transitivity of rationales without loss of generality, completeness is really the distinguishing factor. While this may seem like a technical distinction, the requirement that justifications be preferences underpins the primary interpretation of JU: people justify their decisions by pretending to be better versions of themselves.

For a more formal comparison, consider the revealed preference relation from each model.

Definition 27 (Revealed preference in CFS, MNO)**.**

1. *If* $a \neq c(A \cup \{a\})$ *and, for some* $B \supseteq A$*,* $c(B) \neq c(B \cup \{a\})$*, then* $c(A \cup \{a\})\ R^{JU}\ a$*.*

2. *If* $a \neq c(A \cup \{a\})$ *and, for some* $B \supset A$, $a = c(B \cup \{a\})$, *then* $c(A \cup \{a\})\ R^{CFS}\ a$.

3. *If* $a \neq c(A \cup \{a\})$ *and* $c(A \cup \{a\}) \neq c(A)$, *then* $c(A) \cup \{a\})\ R^{MNO}\ a$.

The revealed preference relation has the same interpretation in all three models: (a, b) belongs to the transitive closure of the revealed preference relation if and only if (a, b) belongs to the primary preference relation in all representations for c. It is not difficult to see that $R^{JU} \supseteq R^{MNO}$ and $R^{JU} \supseteq R^{CFS}$. We claim that the inclusion is strict on any dataset that violates WARP. Since any such dataset contains a cycle or an almost-WARP set, it suffices to show that neither MNO nor CFS delivers full identification on either pattern. Consider a cycle:

$$a = c(\{a, b\}) \quad b = c(\{a, b, d\}) = c(\{b, d\}) \quad d = c(\{a, d\}).$$

The three revealed preference relations are given by

$$\begin{aligned} &a\ R^{CFS}\ b \\ &b\ R^{MNO}\ d \\ &a\ R^{JU}\ b\ R^{JU}\ d \end{aligned}$$

Now consider an almost-WARP set, where the items are indexed from pairwise-best (a_1) to pairwise-worst (a_n). Suppose that $a_i = c(\{a_1, \ldots, a_n\})$, where $i > 1$. Then, R^{CFS} and R^{MNO} are given by

$$\begin{aligned} & & a_j\ R^{CFS}\ a_i \text{ for all } j < i \\ \text{if } i \neq 2 & & a_i\ R^{MNO}\ a_j \text{ for all } j \neq i \\ \text{if } i = 2 & & a_2\ R^{MNO}\ a_j \text{ for all } j > 2 \end{aligned}$$

Since $R^{JU} \supseteq R^{MNO} \cup R^{CFS}$, the only case in which R^{JU} is acyclic is $i = 2$, in which case

$$a_1\ R^{JU} a_2\ \cdots\ a_{n-1}\ R^{JU}\ a_n$$

Thus, JU delivers strictly stronger identification on any dataset that is inconsistent with preference maximization.

In addition to providing stronger identification, JU places stronger restrictions on the data than CFS and MNO. JU is falsifiable with as few as three alternatives, but CFS and MNO are not. To see why, consider an almost-WARP set with three alternatives in which $a_3 = c(\{a_1, \ldots, a_3\})$. CFS and MNO give opposing interpretations of this case, so JU rules it out.

From this example, the reader may wonder whether JU is simply the intersection of CFS and MNO. This is not the case: on almost-WARP sets with more than three elements, JU delivers

stronger identification than CFS and MNO put together. Moreover, on choice domains with more than four elements, JU imposes stronger restrictions on the data than CFS and MNO put together. Thus, the collection of datasets consistent with JU is the intersection of the collections consistent with CFS and MNO for $|\mathcal{A}| \leq 4$, but is a strict subset of the intersection for larger domains.

There is an obvious connection between Corollary 3 of this paper and Proposition 3 of CFS. Both results provide unique "canonical" representations in which the constraints on the DM are minimized. Both state that the primary preference in the canonical representation is pinned down by pairwise choice when choice data is acyclic. The key distinction is that Proposition 3 of CFS *requires* choice data to be acyclic, while Corollary 3 applies to all choice data consistent with JU. As shown in Section 4.2, cycles play an important role in JU: they occur whenever one item is revealed unjustifiable in the presence of another.[9] Moreover, the leading empirical example in CFS is a cycle. For these reasons, the assumption of acyclicity is not innocuous.

JU is also a special case of the model of rationalization in Kalai et al. (2002). That model defines the consideration set in the same way as JU, but does not restrict choice from the consideration set. Kalai et al. (2002) show that this model has no empirical content. Thus, the empirical content of JU comes from the tiebreaking "primary" preference. Kalai et al. (2002) consider another way of introducing empirical content by restricting the number of rationales. That route is not taken in this paper, as there is nothing particularly implausible about a justification model with many justifications. First, wherever "good and reasonable people" exhibit subtle differences of opinion, many preferences will count as justifiable. Second, more justifications correspond to fewer constraints on the DM, so a representation with large $\mathcal{M}$ is arguably more parsimonious.

Several existing models are special cases of JU/JO. The dynamic choice model in Gul and Pesendorfer (2005) is JU with a single justification (or a set of justifications that can be collapsed into a single weak preference).[10] The model in Aizerman and Malishevski (1981) is JU with a constant tiebreaker, so $c(A) = \mathcal{M}(A)$. Aizerman and Malishevski (1981) offer a behavioral characterization of their model, which may be reinterpreted as a characterization of the justifiable sets in JU. The model of willpower in Masatlioglu et al. (2020) is a special case of JO. Like JO, the willpower model uses $(\succsim, c)$ as the primitive. Masatlioglu et al. (2020) show that their model is characterized by IUA and other conditions. Given the structural differences between the two models, this commonality is surprising. The consideration set in the willpower model is characterized by a temptation utility and a willpower cutoff, and it is not immediately obvious how to translate these into justifications.

Several other models of two-tier decision making are related to, but not nested with, JU. The model of sequential rationalization in Manzini and Mariotti (2007) uses one asymmetric binary relation to pin down a consideration set, and another to break ties. Manzini and Mariotti (2007) show that their model satisfies a property they call "Always Chosen:" a is chosen from A if it

[9] Almost-WARP sets occur when one item is revealed unjustifiable in the presence of a non-singleton *set* of items.

[10] The interpretation is the opposite of JU, though. Gul and Pesendorfer (2005) focus on situations in which the "justification" is shortsighted/tempted and the "primary preference" is forward-looking/rational.

pairwise-defeats every other item in A. JU violates this property when $|\mathcal{M}| > 1$ because different binary choices may appeal to different justifications.

The model of selfishness in Dillenberger and Sadowski (2012) pits the DM's primary preference against a subjective norm, but allows the DM to override the norm when his preference is strong enough. The model satisfies Always Chosen because the morally *best* item in the choice set is the only impediment to utility maximization. It does not matter whether the choice set contains many morally good items, or only one. Masatlioglu et al. (2020) satisfies Always Chosen for a similar reason.

Cunningham and de Quidt (2015) (CD) consider a domain in which each item is characterized by a bundle of binary attributes. They distinguish "explicit" preferences over the domain from "implicit" preferences over the attributes. Explicit preferences are independent of the choice set, while implicit preferences may be activated to varying degrees by different choice sets. Both types of preferences are aggregated in a linear utility function. Note that the explicit preference in CD is more like a justification than a primary preference: the DM departs from it more when he can conceal his motivations for doing so. In fact, there is no primary preference in CD. While the signs of the implicit preferences are fixed, the magnitudes are context-dependent. CD provide a series of tests that an analyst can use to recover the signs of the implicit preferences. These tests are similar in spirit to the patterns of choice highlighted in Section 4.2, but CD do not offer a characterization result analogous to Theorem 4.

The proliferation of deterministic models of two-tier decision making is not replicated on the stochastic side. Cattaneo et al. (2020) study a population of decision makers who pay attention to different sets of alternatives. Attention satisfies a monotonicity property: alternatives are more likely to be attended to in smaller sets than in larger sets. Justifiable sets in RJ satisfy monotonicity. The key distinction between RJ and Cattaneo et al. (2020) (as well as Manzini and Mariotti 2014 and Brady and Rehbeck 2016) is that RJ allows for preference heterogeneity as well as heterogeneity in justifications, and fully recovers both. These improvements do not come for free, though: they require the domain to be a set of lotteries, and for preferences to have an EU structure.

7 Conclusion

The family of models in this paper make several points that may be of interest to empirical researchers. First, the puzzling behaviors well documented in moral domains are likely to extend to some non-moral domains. In fact, the same patterns of behavior should be present in any choice setting that invites conflict between primary preferences and social image or self-image. Social image is determined by many factors besides altruism, including rationality, courage, self control, good taste, work ethic, and other virtues. To the author's knowledge, Woolley and Risen (2018) is the only paper to look for "wiggling" behavior in some of these domains.

Second, caution is warranted when interpreting "direct" measures of pro-sociality. A subject who reports a high willingness to donate to charity may be highly constrained rather than highly altruistic. The same is true of a subject who reports that fairness or generosity is central to her sense of self. Putting too much trust in these measures may lead to misleading conclusions. For instance, subjects might seek out information in hopes that it will free them from pressing obligations, not because they place a high value on doing good. A more reliable picture of subjects' pro-sociality can be obtained by offering opportunities to "wiggle" out of obligations. Even so, subjects with a particularly strong sense of obligation may resist these opportunities.

Third, little is known about the way justifiers respond to bundles of choices. Gneezy et al. (2019) and Haisley and Weber (2010) demonstrate that subjects exhibit consistency motives, at least when inconsistencies are sufficiently easy to spot. Subjects find it harder to inflict a bad outcome on someone else when they have already formed or expressed a negative opinion of that outcome. On the other hand, Exley (2016) demonstrates that significant inconsistencies can persist in within-subject data. It would be interesting to see how quickly consistency motives decay over time, and whether they are mitigated or eliminated by anonymity.

We conclude with a note on welfare implications. Although we have referred to the tiebreaker in the justification model as the "primary preference," we do not take it to be a reliable measure of welfare. One can always draw a distinction between choice behavior and welfare—even when choices maximize a standard preference relation—but the point seems particularly important when moral principles are involved. It is entirely plausible that people are better off when they choose in accordance with their principles rather than their baser inclinations. Even those who lack lofty moral principles may benefit from having a virtuous social image. The "primary preference" should therefore be interpreted as an interesting feature of a decision maker's psychology, not the preference of a benevolent agent acting on his behalf.

References

Aizerman, Mark and Andrew Malishevski, "General theory of best variants choice: Some aspects," *IEEE Transactions on Automatic Control*, 1981, *26* (5), 1030–1040.

Andreoni, James, Justin M Rao, and Hannah Trachtman, "Avoiding the ask: A field experiment on altruism, empathy, and charitable giving," *Journal of Political Economy*, 2017, *125* (3), 625–653.

Brady, Richard L and John Rehbeck, "Menu-dependent stochastic feasibility," *Econometrica*, 2016, *84* (3), 1203–1223.

Cattaneo, Matias D, Xinwei Ma, Yusufcan Masatlioglu, and Elchin Suleymanov, "A random attention model," *Journal of Political Economy*, 2020, *128* (7), 2796–2836.

Charness, Gary and Uri Gneezy, "What's in a name? Anonymity and social distance in dictator and ultimatum games," *Journal of Economic Behavior & Organization*, 2008, *68* (1), 29–35.

Cherepanov, Vadim, Timothy Feddersen, and Alvaro Sandroni, "Rationalization," *Theoretical Economics*, 2013, *8* (3), 775–800.

Clark, Stephen A, "The random utility model with an infinite choice space," *Economic Theory*, 1996, *7* (1), 179–189.

Cunningham, Tom and Jonathan de Quidt, "Implicit preferences inferred from choice," *Available at SSRN 2709914*, 2015.

d'Adda, Giovanna, Yu Gao, Russell Golman, and Massimo Tavoni, "It's so Hot in Here: Information Avoidance, Moral Wiggle Room, and High Air Conditioning Usage," Technical Report, Fondazione Eni Enrico Mattei 2018.

Dana, Jason, Daylian M Cain, and Robyn M Dawes, "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organizational Behavior and human decision Processes*, 2006, *100* (2), 193–201.

_ , Roberto A Weber, and Jason Xi Kuang, "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 2007, *33* (1), 67–80.

Dillenberger, David and Philipp Sadowski, "Ashamed to be selfish," *Theoretical Economics*, 2012, *7* (1), 99–124.

Ehrich, Kristine R and Julie R Irwin, "Willful ignorance in the request for product attribute information," *Journal of Marketing Research*, 2005, *42* (3), 266–277.

Exley, Christine L, "Excusing selfishness in charitable giving: The role of risk," *The Review of Economic Studies*, 2016, *83* (2), 587–628.

Falk, Armin, "Facing Yourself-A Note on Self-image," Technical Report, CESifo Working Paper Series 2017.

Fong, Christina M and Felix Oberholzer-Gee, "Truth in giving: Experimental evidence on the welfare effects of informed giving to the poor," *Journal of Public Economics*, 2011, *95* (5-6), 436–444.

Franzen, Axel and Sonja Pointner, "Anonymity in the dictator game revisited," *Journal of Economic Behavior & Organization*, 2012, *81* (1), 74–81.

Freddi, Eleonora, "Do people avoid morally relevant information? Evidence from the refugee crisis," Technical Report, CentER 2017.

Gneezy, Uri, Silvia Saccardo, and Roel Van Veldhuizen, "Bribery: Behavioral drivers of distorted decisions," *Journal of the European Economic Association*, 2019, *17* (3), 917–946.

_ , _ , Marta Serra-Garcia, and Roel van Veldhuizen, "Bribing the self," *Games and Economic Behavior*, 2020, *120*, 311–324.

Grossman, Zachary and Joel J Van Der Weele, "Self-image and willful ignorance in social decisions," *Journal of the European Economic Association*, 2017, *15* (1), 173–217.

Gul, Faruk and Wolfgang Pesendorfer, "The revealed preference theory of changing tastes," *The Review of Economic Studies*, 2005, *72* (2), 429–448.

_ and _ , "Random expected utility," *Econometrica*, 2006, *74* (1), 121–146.

Haisley, Emily C and Roberto A Weber, "Self-serving interpretations of ambiguity in other-regarding behavior," *Games and economic behavior*, 2010, *68* (2), 614–625.

Haley, Kevin J and Daniel MT Fessler, "Nobody's watching?: Subtle cues affect generosity in an anonymous economic game," *Evolution and Human behavior*, 2005, *26* (3), 245–256.

Hamman, John R, George Loewenstein, and Roberto A Weber, "Self-interest through delegation: An additional rationale for the principal-agent relationship," *American Economic Review*, 2010, *100* (4), 1826–46.

Herden, Gerhard and Andreas Pallack, "On the continuous analogue of the Szpilrajn Theorem I," *Mathematical social sciences*, 2002, *43* (2), 115–134.

Kajackaite, Agne, "If I close my eyes, nobody will get hurt: The effect of ignorance on performance in a real-effort experiment," *Journal of Economic Behavior & Organization*, 2015, *116*, 518–524.

Kalai, Gil, Ariel Rubinstein, and Ran Spiegler, "Rationalizing choice functions by multiple rationales," *Econometrica*, 2002, *70* (6), 2481–2488.

Loewenstein, George, Samuel Issacharoff, Colin Camerer, and Linda Babcock, "Self-serving assessments of fairness and pretrial bargaining," *The Journal of Legal Studies*, 1993, *22* (1), 135–159.

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

_ **and** _ , "Stochastic choice and consideration sets," *Econometrica*, 2014, *82* (3), 1153–1176.

Masatlioglu, Yusufcan, Daisuke Nakajima, and Emre Ozdenoren, "Willpower and compromise effect," *Theoretical Economics*, 2020, *15* (1), 279–317.

_ , _ , and Erkut Y Ozbay, "Revealed attention," *American Economic Review*, 2012, *102* (5), 2183–2205.

Norton, Michael I, Joseph A Vandello, and John M Darley, "Casuistry and social category bias.," *Journal of personality and social psychology*, 2004, *87* (6), 817.

Plott, Charles R, "Path independence, rationality, and social choice," *Econometrica: Journal of the Econometric Society*, 1973, pp. 1075–1091.

Rodriguez-Lara, Ismael and Luis Moreno-Garrido, "Self-interest and fairness: self-serving choices of justice principles," *Experimental Economics*, 2012, *15* (1), 158–175.

Serra-Garcia, Marta and Nora Szech, "The (in) elasticity of moral ignorance," Technical Report, KIT Working Paper Series in Economics 2019.

Strotz, Robert Henry, "Myopia and inconsistency in dynamic utility maximization," *The review of economic studies*, 1955, *23* (3), 165–180.

Woolley, Kaitlin and Jane L Risen, "Closing your eyes to follow your heart: Avoiding information to protect a strong intuitive preference.," *Journal of personality and social psychology*, 2018, *114* (2), 230.

A Proofs of results in text

A.1 Proof of Theorem 1

First, we show necessity. Necessity of Optimization follows because the items that maximize a preference over a set must all be indifferent. For IUA, fix A, a such that $a \in A$. Suppose $a \succsim c(A)$ and $a \notin c(A)$, and fix $B \supseteq A$. For all $\succ_m \in \mathcal{M}$, we have $a \not\succsim_m A$, so $a \not\succsim_m B$. To confirm that $c(B) = c(B \setminus \{a\})$, it suffices to show that $\mathcal{M}(B) = \mathcal{M}(B \setminus \{a\})$. Take any $b \in \mathcal{M}(B)$. Since $b \succsim_m B$ for some $\succ_m \in \mathcal{M}$, $b \succsim_m A$ for some $\succsim_m \in \mathcal{M}$, so $b \neq a$. Since $b \succsim_m B$ implies $b \succsim_m B \setminus \{a\}$, we have $b \in \mathcal{M}(B \setminus \{a\})$. Now take any $b \in \mathcal{M}(B \setminus \{a\})$. There exists $\succ_m \in \mathcal{M}$ such that $b \succsim_m B \setminus \{a\}$. Since it cannot be that $a \succsim_m b \succsim_m B \setminus \{a\}$, we have $b \succsim_m B$, so $b \in \mathcal{M}(B)$.

Now we show sufficiency. To define $\mathcal{M}$, we need the notion of exclusion from below.

Definition 28 (Exclusion from below ($\triangleright$))**.** *Say $X \in \mathcal{F}(\mathcal{A})$ excludes $y \notin X$ from below (written $X \triangleright y$) if $y \succsim X$ and $y \notin c(X \cup \{y\})$.*

Say that a strict preference $\succ_m$ respects exclusion from below if

$$X \triangleright y \quad \Longrightarrow \quad \exists x \in X \; x \succ_m y.$$

We take $\mathcal{M}$ to be the set of strict preferences on $\mathcal{A}$ that respect exclusion from below.

Fix $A \in \mathcal{F}(\mathcal{A})$ and $b \notin A$ such that $b \succsim c(A \cup \{b\})$ and $b \notin c(A \cup \{b\})$. We show that $b \notin \mathcal{M}(A \cup \{b\})$.

Lemma 1. *Fix $X \in \mathcal{F}(\mathcal{A})$ and $y \notin X$. If $y \succsim c(X \cup \{y\})$ and $y \notin c(X \cup \{y\})$, then*

$$c(X \cup \{y\}) \cup \{x \in X : y \succ x\} \triangleright y.$$

Proof. Take any $x \in X$ such that $x \succsim y$ and $x \notin c(X \cup \{y\})$. Since $y \succsim c(X \cup \{y\})$, we have $x \succsim c(X \cup \{y\})$. By IUA, $c(X \cup \{y\}) = c((X \cup \{y\}) \setminus \{x\})$, so $y \notin c((X \cup \{y\}) \setminus \{x\})$ and $y \succsim c((X \cup \{y\}) \setminus \{x\})$. Iterating this argument, we can remove every $x \in X$ such that $x \succsim y$ and $x \notin c(X \cup \{y\})$ without changing choice. We end up with

$$y \notin c(\{y\} \cup c(X \cup \{y\}) \cup \{x \in X : y \succ x\}).$$

This implies $c(X \cup \{y\}) \cup \{x \in X : y \succ x\} \triangleright y$. □

Lemma 1 implies that $c(A \cup \{b\}) \cup \{a \in A : b \succ a\} \triangleright b$. Since each $\succ_m \in \mathcal{M}$ respects exclusion from below, we have $b \not\succ_m c(A \cup \{b\}) \cup \{a \in A : b \succ a\}$ for all $\succ_m \in \mathcal{M}$. This implies $b \not\succ_m A$ for all $\succ_m \in \mathcal{M}$, so $b \notin \mathcal{M}(A \cup \{b\})$.

Now fix $A \in \mathcal{F}(\mathcal{A})$ and $b \notin A$ such that $b \in c(A \cup \{b\})$. We show that $b \in \mathcal{M}(A \cup \{b\})$: there is a strict preference $\succ_m$ that respects exclusion from below and has $b \succ_m A$.

We will construct an appropriate $\succ_m$ by extending $\triangleright$. We define several useful properties of $\triangleright$.

Definition 29 (Menu-item relation)**.** *A menu-item relation is a subset of $(\mathcal{F}(\mathcal{A}) \cup \{\emptyset\}) \times \mathcal{A}$.*

Definition 30 (Transitivity)**.** *A menu-item relation R is transitive if*

$$(X \; R \; x, Y \; R \; y \text{ and } x \in Y) \Longrightarrow (X \cup Y) \setminus \{x, y\} \; R \; y.$$

We denote the transitive closure of a menu-item relation R by $\mathrm{tr}(R)$.

To see why $\triangleright$ is transitive, suppose $X \triangleright x$, $Y \triangleright y$, and $x \in Y$. We have $x \succsim X$ and $y \succsim Y$. Since $x \in Y$, we have $y \succsim x \succsim X$. IUA implies y is irrelevant for choice on any superset of Y, so $y \notin c(X \cup Y \cup \{y\})$. IUA also implies x is irrelevant for choice on any superset of X, so $c(X \cup Y \cup \{y\}) = c((X \cup Y \cup \{y\}) \setminus \{x\})$. We have $y \notin c((X \cup Y \cup \{y\}) \setminus \{x\})$ as well as $y \succsim X \cup Y \setminus \{x, y\}$, so $X \cup Y \setminus \{x, y\} \triangleright y$.

Definition 31 (Properness)**.** *A menu-item relation R is proper if $X\ R\ x \implies X \neq \emptyset$.*

Definition 32 (Irreflexivity)**.** *A menu-item relation R is irreflexive if $X\ R\ x \implies x \notin X$.*

Definition 33 (Consistency with $b \succ A$)**.** *A menu-item relation R is consistent with $b \succ A$ if it is not the case that $A'\ R\ b$ for any $A' \in \mathcal{F}(A)$.*

To see why $\triangleright$ is consistent with $b \succ A$, suppose $A' \triangleright b$ for some $A' \subset A$. By IUA, b is irrelevant for choice on any superset of A', including A. This contradicts $b \notin c(A \cup \{b\})$.

The following two lemmas will be useful for extending $\triangleright$.

Lemma 2. *Fix an irreflexive, transitive and proper menu-item relation R. Fix distinct $x, y \in \mathcal{A}$ such that $\neg(\{y\}\ R\ x)$. Then, $tr(R \cup (\{x\}, y))$ is irreflexive and proper.*

Proof. Let $R^0 := R$. For $i > 0$, let R^i be the extension of R^{i-1} obtained by imposing

$$\left(\bigcup_{j=1}^{k} X_j \cup \{y_{k+1}, \ldots, y_n\}\right) \setminus \{y\}\ R^i\ y$$

whenever

$$\{y_1, \ldots, y_n\}\ R^0\ y \text{ and, for all } j \leq k,\ X_j\ R^{i-1}\ y_j.$$

Then, the transitive closure of R is $\bigcup_{i=0}^{\infty} R^i$. This is a standard result about the transitive closure. The usual proof goes through with the version of transitivity used here.

Repeated applications of transitivity will not lead to a violation of irreflexivity, so we only need to check whether $\text{tr}(R \cup (\{x\}, y))$ is proper. To keep track of repeated applications of transitivity, we introduce the notion of a tree. To simplify notation, we write z^k instead of $(z_0, \ldots, z_k)$ and $\{z^k\}$ instead of $\{z_0, \ldots, z_k\}$.

Definition 34 (Q-tree)**.** *For a menu-item relation Q, a Q-tree from $W \in (\mathcal{F}(\mathcal{A}) \cup \{\emptyset\})$ to $w \in \mathcal{A}$ is inductively defined as follows:*

- *The level-0 node $z_0 := w$ is mapped to a parent set $Z_1(z_0)$ such that $Z_1(z_0)\ Q\ z_0$. A generic member of of $Z_1(z_0)$ is denoted $z_1(z_0)$.*
- *For $k > 0$: each level-k node $z_k(z^{k-1}) \notin W \cup \{z^{k-1}\}$ is mapped to a parent set $Z_{k+1}(z^k)$ such that $Z_{k+1}(z^k)\ Q\ z_k$. A generic member of $Z_{k+1}(z^k)$ is denoted $z_k(z^{k-1})$.*
- *For some finite $K > 0$: each level-K node $z_K(z^{K-1})$ belongs to $W \cup \{z^{K-1}\}$.*

We refer to nodes that do not have parents as top nodes. A branch of a tree is a sequence $(z_0, z_1(z_0), z_2(z^1), \ldots, z_k(z^{k-1}))$ where $z_k(z^{k-1})$ is a top node. We refer to $(z_0, \ldots, z_{i-1}(z^{i-2}))$ descendants of $z_i(z^{i-1})$, and $(z_{i+1}(z^i), \ldots, z_k(z^{k-1}))$ as ancestors of z_i.

It is not difficult to see that (W, w) belongs to $\mathrm{tr}(Q)$ if and only if there is a Q-tree from W to w. Suppose that $\mathrm{tr}(R \cup (\{x\}, y))$ is improper, so there is a $R \cup (\{x\}, y)$-tree from $\emptyset$ to w for some $w \in \mathcal{A}$. Notice that there must be at least one point in the tree in which x is the sole parent of y. Otherwise, there would be an R-tree from $\emptyset$ to w, contradicting properness of R.

Construct a new tree by removing all the ancestors of y wherever x is the sole parent of y. The result is an R-tree. Let V be the set of items that descend from any instance of y that had x as its sole parent in the original tree. Fix any $v \in V$, and drop all the items from the R-tree that are not ancestors of v. The result is an R-tree from $\{y\}$ to v, so it must be that $\{y\}\ R\ v$. Now return to the original tree. Take any point in the tree where x is the sole parent of y. Construct a new tree by removing everything except this instance of x and its ancestors. The result is an R-tree from a subset of $V \cup \{y\}$ to x. To see why, recall that every top node in the original tree is a duplicate of one of its descendants. Fix any top node z_k of the new tree, and consider the branch of the original tree running through it: $(w, \ldots, y, x, \ldots, z_k)$. If z_k is not duplicated in $(x, \ldots, z_{k-1})$, it must be duplicated in $(w, \ldots, y)$—so it must belong to $V \cup \{y\}$. Since R is transitive, we have $V'\ R\ x$ for some $V' \subseteq V \cup \{y\}$. Since $\{y\}\ R\ v$ for all $v \in V$, applying transitivity once more gives $\{y\}\ R\ x$—contradiction. □

Lemma 3. *Fix an irreflexive, transitive, proper and (b, A)-consistent menu-item relation R. For any $a \in A$, $tr(R \cup (\{b\}, a))$ is consistent with $b \succ A$.*

Proof. Suppose that $\mathrm{tr}(R \cup (\{b\}, a))$ is inconsistent with $b \succ A$, so $(A', b) \in \mathrm{tr}(R \cup (\{b\}, a))$ for some $A' \in \mathcal{F}(A)$. Then, there must be an $R \cup (\{b\}, a)$ tree from A' to b. Construct a new tree by removing all the ancestors of a wherever b is the sole parent of a. The result is an R-tree from a subset of $A' \cup \{a\}$ to b. Since R is transitive, we have $A''\ R\ b$ for some $A'' \subseteq A' \cup \{a\} \subseteq A$. This contradicts consistency of R with $b \succ A$. □

Let $A = \{a_1, \ldots, a_n\}$. Let $\triangleright^0 := \triangleright$. For $i \in \{1, \ldots, n\}$, let $\triangleright^i = \mathrm{tr}(\triangleright_{i-1} \cup (\{b\}, a_i))$. Since $\triangleright$ is irreflexive, proper, transitive, and consistent with $b \succ A$, we can use Lemmas 2 and 3 to show that the same is true of each $\triangleright^i$. Notice that $\{b\} \triangleright^n a$ for all $a \in A$.

Now we use Lemma 2 to show that $\triangleright^n$ can be extended to an irreflexive, proper and transitive relation $\triangleright^+$ such that, for all distinct $x, y \in \mathcal{A}$, $\{x\} \triangleright^+ y$ or $\{y\} \triangleright^+ x$. The proof is similar to that of the Szpilrajn Extension Theorem. Consider the set of irreflexive, proper and transitive relations that extend $\triangleright^n$, ordered by set inclusion. Take any chain in the partially ordered set. The union of its elements is clearly irreflexive, proper and transitive, so it is an upper bound for the chain. By Zorn's Lemma, the partially ordered set must have a maximal element $\triangleright^+$. Suppose that, for some distinct x, y, neither $\{x\} \triangleright^+ y$ nor $\{x\} \triangleright^+ y$. By Lemma 2, $\triangleright^+$ can be extended to another irreflexive, proper and transitive relation containing $(\{x\}, y)$. Then $\triangleright^+$ cannot be maximal, a contradiction.

Moreover, for each $X \in \mathcal{F}(\mathcal{A})$ and $y \in \mathcal{A}$, $\triangleright^+$ must satisfy

$$X \triangleright y \Longrightarrow \left(\exists x \in X \text{ s.t. } \{x\} \triangleright^+ y\right). \tag{5}$$

Suppose not. Then $\{y\} \triangleright^+ x$ for all $x \in X$, as well as $X \triangleright^+ y$. Since $\triangleright^+$ is transitive, $\emptyset \triangleright^+ y$. Since $\triangleright^+$ is proper, this is a contradiction. Similarly, suppose that $\{x\} \triangleright^+ y$ and $\{y\} \triangleright^+ x$. By transitivity, $\emptyset \triangleright^+ x$, a contradiction.

We can use $\triangleright^+$ to define a strict preference $\succ_m$:

$$\{x\} \triangleright^+ y \quad \Longleftrightarrow \quad x \succ_m y.$$

It is easy to see that $\succ_m$ is antisymmetric, complete and transitive. It respects exclusion from below because of (5), so it is indeed in $\mathcal{M}$. It also satisfies $b \succ_m A$ because $\triangleright^+$ extends $\triangleright^n$, and $\{b\} \triangleright^n a$ for all $a \in A$.

A.2 Proof of Proposition 1

First, we show necessity of ISA. Fix $B \in \mathcal{F}(\mathcal{A})$ and $A \subseteq S(B)$. For all $\succ_m \in \mathcal{M}$, we have $a \not\succsim_m A$, so $a \not\succsim_m B$. To confirm that $c(B) = c(B \setminus A)$, it suffices to show that $\mathcal{M}(B) = \mathcal{M}(B \setminus A)$. Take any $b \in \mathcal{M}(B)$. Since there exists $\succ_m \in \mathcal{M}$ such that $b \succsim_m B$, it cannot be that $b \in A$. Since $b \succsim_m B$ implies $b \succsim_m B \setminus A$, we have $b \in \mathcal{M}(B \setminus A)$. Now suppose $b \in \mathcal{M}(B \setminus A)$, so $b \succsim_m B \setminus A$ for some $\succ_m \in B \setminus A$. Suppose that $a \succsim_m b \succsim_m B \setminus A$ for some $a \in A$. For the $\succ_m$-best such a, we must have $a \succsim_m B$—contradiction. Conclude that $b \succsim_m B$, so $b \in \mathcal{M}(B)$.

Now we show sufficiency. The proof is similar to that of Theorem 1. Let D be the menu-item relation such that

$$\{x\} \; D \; y \quad \Longleftrightarrow \quad x \succ_D y.$$

For any $y \in \mathcal{A}$ and $X \in \mathcal{F}(\mathcal{A})$, let $D(y, X)$ be the subset of items in X that are not dominated by y:

$$D(y, X) := X \setminus \{x \in X : y \succ_D x\}.$$

Define D-exclusion from below as follows:

$$X \triangleright_D y \quad \Longleftrightarrow \quad \left(X \triangleright y \text{ and } X = D(y, X)\right).$$

We will take $\mathcal{M}$ to be the set of D-monotone strict preferences that respect $\triangleright_D$. Formally, a D-monotone strict preference $\succ_m$ belongs to $\mathcal{M}$ if and only if

$$X \triangleright_D y \quad \Longrightarrow \quad \exists x \in X \; x \succ_m y.$$

Fix $A \in \mathcal{F}(\mathcal{A})$ and $b \notin A$ such that $b \succsim c(A \cup \{b\})$ and $b \notin c(A \cup \{b\})$. We show that $b \notin \mathcal{M}(A \cup \{b\})$.

Lemma 4. *Fix $X \in \mathcal{F}(\mathcal{A})$ and $y \notin X$. If $y \succsim c(X \cup \{y\})$ and $y \notin c(X \cup \{y\})$, then*

$$D(y, c(X \cup \{y\}) \cup \{x \in X : y \succ x\}) \triangleright_D y.$$

Proof. By Lemma 1, $c(X \cup \{y\}) \cup \{x \in X : y \succ x\} \triangleright y$. This implies $y \notin c(\{y\} \cup c(X \cup \{y\}) \cup \{x \in X : y \succ x\})$. By ISA, every item in $c(X \cup \{y\}) \cup \{x \in X : y \succ x\}$ that is dominated by y can be removed without changing choice. We get

$$y \notin c(\{y\} \cup D(y, c(X \cup \{y\}) \cup \{x \in X : y \succ x\})).$$

This implies

$$D(y, c(X \cup \{y\}) \cup \{x \in X : y \succ x\}) \triangleright y,$$

so

$$D(y, c(X \cup \{y\}) \cup \{x \in X : y \succ x\}) \triangleright_D y,$$

□

Lemma 4 implies that $D(b, c(A \cup \{b\}) \cup \{a \in A : b \succ a\}) \triangleright_D b$. Since all the preferences in $\mathcal{M}$ respect D-exclusion from below, we have $b \not\succ_m D(b, c(A \cup \{b\}) \cup \{a \in A : b \succ a\})$ for all $\succ_m \in \mathcal{M}$. This implies $b \not\succ_m A$ for all $\succ_m \in \mathcal{M}$, so $b \notin \mathcal{M}(A \cup \{b\})$.

Now fix $A \in \mathcal{F}(\mathcal{A})$ and $b \notin A$ such that $b \in c(A \cup \{b\})$. We show that $b \in \mathcal{M}(A \cup \{b\})$: there is a strict preference $\succ_m$ that respects D-exclusion from below and has $b \succ_m A$.

We will construct an appropriate $\succ_m$ by extending $\triangleright_D$. First, we define two useful properties of menu-item relations.

Definition 35 (D-transitivity)**.** *A menu-item relation R is D-transitive if*

$$(X \; R \; x, Y \; R \; y \text{ and } x \in Y) \Longrightarrow D(y, X \cup Y \setminus \{x, y\}) \; R \; y.$$

For any menu-item relation R, let $D\text{-tr}(R)$ denote the D-transitive closure of R.

Definition 36 (D-monotonicity)**.** *A menu-item relation R is D-monotone if R extends D and*

$$X \; R \; y \quad \Longrightarrow \quad X = D(y, X).$$

Lemma 5. *$D\text{-tr}(\triangleright_D \cup D)$ is irreflexive, D-monotone and proper, and consistent with $b \succ A$.*

Proof. Since $\triangleright_D \cup D$ is irreflexive and D-monotone, and application of D-transitivity preserves irreflexivity and D-monotonicity, it is clear that $D\text{-tr}(\triangleright_D \cup D)$ is irreflexive and D-monotone. As in

the proof of Lemma 2, we check properness via a tree. The relevant construction is very similar to Definition 34. The only difference is as follows. In Definition 34, each top node is in W or identical to one of its descendants. Now, each top node is in W or identical to *or dominated by* one of its descendants.

Suppose there is a $(\triangleright_D \cup D)$-tree from $\emptyset$ to w. Consider a menu Z that consists of all the items in the tree. Take any $z \in c(Z)$. By assumption, there exists $Z' \subset Z$ such that $Z' \triangleright_D z$ or $Z'\ D\ z$. By ISA, $c(Z) = c(Z \setminus \{z\})$, which contradicts the assumption that $z \in c(Z)$.

Suppose that there is a $(\triangleright_D \cup D)$-tree from $A' \in \mathcal{F}(A)$ to b. Consider a menu Z that consists of all the items in the tree as well as $A \setminus A'$. By assumption, for each $z \in Z \setminus A$, there exists $Z' \subset Z$ such that $Z' \triangleright_D z$ or $Z'\ D\ z$. By ISA, $b \notin c(Z)$. Also by ISA, $c(Z) = c(A \cup \{b\})$. This contradicts the assumption that $b \in c(A \cup \{b\})$. □

Lemma 6. *Fix an irreflexive, proper, D-transitive and D-monotone menu-item relation R. Fix distinct $x, y \in \mathcal{A}$ such that $\neg(\{y\}\ R\ x)$. The D-transitive closure of $R \cup (\{x\}, y)$ is irreflexive, D-monotone and proper.*

Proof. The proof is very similar to that of Lemma 2, but using the modified notion of a tree from the proof of Lemma 5. A pair (W, w) belongs to $D\text{-tr}(R \cup (\{x\}, y))$ if and only if there is an $R \cup (\{x\}, y)$-tree from W to w. Repeated application of D-transitivity will not cause a violation of irreflexivity or D-monotonicity, so we only need to check properness.

Suppose there is an $R \cup (\{x\}, y)$-tree from $\emptyset$ to w. Construct a new tree by removing all the ancestors of y wherever x is the sole parent of y. Let V be the set of items that descend from any instance of y that had x as its sole parent in the original tree. Just as in the proof of Lemma 2, we have $\{y\}\ R\ v$ for each $v \in V$. Let W be the set of items in the original tree that are dominated by a member of $\{y\} \cup V$. Since R is D-monotone and D-transitive, $\{y\}\ R\ w$ for each $w \in W$.

Now return to the original tree. Take any point in the tree where x is the sole parent of y. Construct a new tree by removing everything except this instance of x and its ancestors. The result is an R-tree from a subset of $W \cup V \cup \{y\}$ to x. To see why, notice that every top node in the original tree is a duplicate of one of its descendants or dominated by one of its descendants. Fix any top node z_k of the new tree, and consider the branch of the original tree running through it: $(w, \ldots, y, x, \ldots, z_k)$. If z_k is not duplicated in, or dominated by something in, $(x, \ldots, z_{k-1})$, then it must be duplicated in, or dominated by something in, $(w, \ldots, y)$. Since each item in $(w, \ldots, y)$ belongs to $V \cup \{y\}$, z_k belongs to $W \cup V \cup \{y\}$. Since R is D-transitive, we have $X\ R\ x$ for some $X \subseteq W \cup V \cup \{y\}$. Since $\{y\}\ R\ v$ for all $v \in V$, and $\{y\}\ R\ w$ for all $w \in W$, applying D-transitivity once more gives $\{y\}\ R\ x$—contradiction. □

Lemma 7. *Fix an irreflexive, proper, D-transitive, D-monotone menu-item relation R consistent with $b \succ A$. For any $a \in A$, $D\text{-tr}(R \cup (\{b\}, a))$ is consistent with $b \succ A$.*

Proof. The proof is exactly the same as that of Lemma 3, but using D-transitivity in place of transitivity and using the modified notion of a tree from the proof of Lemma 5. □

We can define $\triangleright_D^i$ for $i \in \{0, \ldots, n\}$ exactly as in the proof of Theorem 1, but using D-tr$(\triangleright_D \cup D)$ in place of $\triangleright$, and D-tr instead of tr. For each i, $\triangleright_D^i$ will be irreflexive, proper, D-transitive, D-monotone, and consistent with $b \succ A$. We will have $\{b\} \triangleright_D^n a$ for all $a \in A$.

As in the proof of Theorem 1, we can use Lemma 6 to extend $\triangleright_D^n$ to an irreflexive, proper, D-transitive and D-monotone relation $\triangleright_D^+$ such that, for all distinct $x, y \in \mathcal{A}$, $\{x\} \triangleright_D^+ y$ or $\{y\} \triangleright_D^+ x$. We will have

$$X \triangleright_D y \Longrightarrow \left(\exists x \in X \text{ s.t. } \{x\} \triangleright_D^+ y\right). \tag{6}$$

We can use $\triangleright_D^+$ to define a strict preference $\succ_m$:

$$\{x\} \triangleright_D^+ y \quad \Longleftrightarrow \quad x \succ_m y.$$

It is easy to see that $\succ_m$ is antisymmetric, complete and transitive. It extends $\succ_D$ because $\triangleright_D^+$ extends D. It respects exclusion from below because of (6). Finally, it satisfies $b \succ_m A$ because $\triangleright_D^+$ extends $\triangleright^n$, and $\{b\} \triangleright^n a$ for all $a \in A$.

A.3 Proof of Theorem 2

Lemma 8. *C-IUA implies IUA.*

Proof. We first show that $B \subset \bar{W}(B)$ for all $B \in \mathcal{F}(\mathcal{A})$. Take any $b \in B$. By Improvability, we can find a sequence $B_i \to b$ such that $b \in W(B_i)$ for all i. Let $\hat{B}_i := B_i \cup B \setminus \{b\}$. We have $\hat{B}_i \to B$, and $b \in W(\hat{B}_i)$ for all i, so $b \in \bar{W}(B)$.

Take any $A \in \mathcal{F}(\mathcal{A})$, and index the items in A from best (1) to worst ($|A|$). Break ties arbitrarily, with one exception: everything in $\{a \in A : a \sim c(A), a \notin c(A)\}$ must have a lower index than everything in $c(A)$.

If $a_1 \in c(A)$, then there is no $a \in A$ such that $a \succsim c(A)$ and $a \notin c(A)$, so IUA has no bite. Suppose that $a_1 \notin c(A)$. We have $a_1 \succsim A \setminus \{a_1\}$, so $a_1 \in W(A \setminus \{a_1\})$, so $a_1 \in W(A)$. We showed above that $W(A) \in \mathcal{F}(\bar{W}(A))$, so C-IUA implies $c(A) = c(A \setminus \{a_1\})$. Now suppose $a_2 \succsim c(A)$ but $a_2 \notin c(A)$, so $a_2 \notin c(A \setminus \{a_1\})$. We have $a_2 \succsim A \setminus \{a_1, a_2\}$, so $a_2 \in W(A \setminus \{a_1, a_2\})$, so $a_2 \in W(A)$. C-IUA implies $c(A) = c(A \setminus \{a_2\}) = c(A \setminus \{a_1, a_2\})$. Iterating the argument, we get $c(A) = c(A \setminus \{a_i\})$ for all a_i such that $a_i \succsim c(A)$ and $a_i \notin c(A)$. This is IUA. □

Definition 37 (Respects exclusion)**.** *$m : \mathcal{A} \to \mathbb{R}$ respects exclusion if, for all $(A, b) \in \mathcal{F}(\mathcal{A}) \times \mathcal{A}$ such that $b \in W(A)$, there exists $a \in A$ such that $m(a) > m(b)$.*

Let

$$\mathcal{M} := \{m \in C(\mathcal{A}, \mathbb{R}) : m \text{ respects exclusion } \}.$$

Later, we will confirm that $\mathcal{M}$ is nonempty.

For any $A \in \mathcal{F}(\mathcal{A})$, let

$$\mathcal{M}(A) := \bigcup_{m \in \mathcal{M}} \arg\max_{a \in A} m(a).$$

We show that, for any $A \in \mathcal{F}(\mathcal{A})$, $a \succsim c(A)$ and $a \notin c(A)$ implies $a \notin \mathcal{M}(A)$. By Lemma 1, $a \succsim c(A)$ and $a \notin c(A)$ implies $c(A) \cup \{a \in A : c(A) \succ a\} \triangleright a$. By definition of W, $a \in W(c(A) \cup \{a \in A : c(A) \succ a\})$, so $a \in W(A)$. For any $m : \mathcal{A} \to \mathbb{R}$ that respects exclusion, $a \notin \arg\max_{\tilde{a} \in A} m(\tilde{a})$. This implies $a \notin \mathcal{M}(A)$.

The remainder of the proof establishes that, for any $A \in \mathcal{A}$, $c(A) \subseteq \mathcal{M}(A)$. To this end, take any $A, b \in \mathcal{F}(\mathcal{A}) \times \mathcal{A}$ such that $b \in c(A \cup \{b\})$. We will show that there exists $m^* \in \mathcal{M}$ such that $b \in \arg\max_{a \in A} m^*(a)$, so $b \in \mathcal{M}(A)$.

If there is no $(X, y) \in \mathcal{F}(\mathcal{A}) \times \mathcal{A}$ such that $y \in W(X)$, then $\mathcal{M} = C(A, \mathbb{R})$. This implies $\mathcal{M}(X) = X$ for all $X \in \mathcal{F}(\mathcal{A})$. We can therefore assume non-triviality when constructing m^*.

Definition 38 (Non-triviality)**.** *There exists $(X, y) \in \mathcal{F}(\mathcal{A}) \times \mathcal{A}$ such that $y \in W(X)$.*

Under non-triviality, we can apply Theorem 3.4 in Herden and Pallack (2002) (HP). The theorem uses the following definition, adapted to our notation.

Definition 39 (HP-system)**.** *Let R be a binary relation on $\mathcal{F}(\mathcal{A})$. A family $\{E_i\}_{i=0}^{\infty}$ of open subsets of $\mathcal{F}(\mathcal{A})$ is an HP-system for R if it satisfies the following conditions:*

1. *There exist $E, E' \in \{E_i\}_{i=0}^{\infty}$ such that $cl(E) \subset E'$.*
2. *For all $E, E' \in \{E_i\}_{i=0}^{\infty}$, $E \subseteq E'$ or $E' \subseteq E$.*
3. *For all $E, E' \in \{E_i\}_{i=0}^{\infty}$ such that $cl(E) \subset E'$, there exists some $E'' \in \{E_i\}_{i=0}^{\infty}$ such that $cl(E) \subset E'' \subset cl(E'') \subset E'$.*
4. *For any $X, Y \in \mathcal{F}(\mathcal{A})$ and any $E \in \{E_i\}_{i=0}^{\infty}$: if $X \in E$ and $X \; R \; Y$, then $Y \in E$.*
5. *For each $X, Y \in \mathcal{F}(\mathcal{A})$ such that $X \; R$; Y and $\neg(Y \; R \; X)$, there exist $E, E' \in \{E_i\}_{i=0}^{\infty}$ such that $cl(E) \subset E'$, $Y \in E$ and $X \notin E'$.*

We will construct an HP system $\{E_i\}_{i=0}^{\infty}$ for the relation R given by

$$X \; R \; Y \quad \iff \quad Y \subset W(X). \tag{7}$$

The following Lemma will be used repeatedly.

Lemma 9. *For any $z \in \mathcal{A}$ and any $X, Y \in \mathcal{F}(\mathcal{A})$: if $z \in W(Y)$ and $Y \subset \bar{W}(X)$, or if $z \in \bar{W}(Y)$ and $Y \subset W(X)$, then $z \in W(X)$.*

Proof. Suppose $z \in W(Y)$ and $Y \subset \bar{W}(X)$. By definition of W, we have $Y' \subseteq Y$ such that $z \succsim Y'$ and $z \notin c(\{z\} \cup Y')$. By definition of $\bar{W}$, we have $X_i \to X$ and $y_i \to y$ such that $y_i \in W(X_i)$ for all i. $y_i \in W(X_i)$ means there is some $X_i(y) \subseteq X_i$ such that $y_i \succsim X_i(y)$ and $y_i \notin c(\{y_i\} \cup X_i(y))$. Passing to a subsequence if necessary, let $X(y) := \lim_{i\to\infty} X_i(y)$. Let $X' := \bigcup_{y\in Y'} X(y)$. We have $X_i(y) \cup X' \setminus X(y) \to X'$, Since $y_i \in W(X_i(y))$ for all i, we also have $y_i \in W(X_i(y) \cup X' \setminus X(y))$ for all i. We conclude that $Y' \subset \bar{W}(X')$. By C-IUA, $z \notin c(\{z\} \cup X')$. Since $y_i \succsim X_i(y)$ for all i and all $y \in Y'$, and since $\succsim$ is continuous, $y \succsim X(y)$ for all $y \in Y'$. Since Y' is finite, there must be some $y \in Y'$ such that $y \succsim X'$. Since $z \succsim Y'$, we have $z \succsim X'$. We conclude that $z \in W(X')$, so $z \in W(X)$.

A parallel argument covers the second case: $z \in \bar{W}(Y)$ and $Y \subset W(X)$. Suppose $z \in \bar{W}(Y)$. Just as above, we can find y' such that $z \in \bar{W}(Y')$ and $z \succsim Y'$. For each $y \in Y'$, we can find $X(y)$ such that $y \in W(X(y))$ and $y \succsim X(y)$. Letting $X' := \bigcup_{y\in Y'} X(y)$, we get $Y' \subset W(X')$ and $y \succsim X'$ for some $y \in Y'$. Since $z \succsim Y'$, $z \succsim X'$. C-IUA gives $z \notin c(\{z\} \cup X')$. We conclude that $z \in W(X')$, so $z \in W(X)$. □

Now we construct $\{E_i\}_{i=0}^{\infty}$.

1. Let
$$E_0 := \mathcal{F}(W(\{b\} \cup A)).$$
Continuity says that $W(\{b\} \cup A)$ is open, so $\mathcal{F}(W(\{b\} \cup A))$ is open. By IUA, $b \notin E_0$.

   Let
$$\bar{E}_0 := \mathcal{F}(\bar{W}(\{b\} \cup A)).$$
We show that $\bar{E}_0$ is closed. Take $\{X_i \in \mathcal{F}(\bar{W}(\{b\} \cup A))\}_{i=1}^{\infty}$ such that $X_i \to X \in \mathcal{F}(\mathcal{A})$. For each i, there exist $\{X_{ij} \in \mathcal{F}(\mathcal{A})\}_{i=1}^{\infty} \to X_i$ and $\{Y_{ij} \in \mathcal{F}(\mathcal{A})\}_{i=1}^{\infty} \to \{b\} \cup A$ such that $X_{ij} \in W(Y_{ij})$ for all j. Notice that $X_{ii} \to X$ and $Y_{ii} \to \{b\} \cup A$. Since $X_{ii} \in W(Y_{ii})$ for all i, $X \in \mathcal{F}(\bar{W}(\{b\} \cup A))$.

   By the definitions of W and $\bar{W}$, $E_0 \subset \bar{E}_0$. Since $\bar{E}_0$ is closed, $\mathrm{cl}(E_0) \subset \bar{E}_0$. We showed in the proof of Lemma 8 that, for any $X \in \mathcal{F}(\mathcal{A})$, $X \subset \bar{W}(X)$. Thus, $\{b\} \cup A \subset \bar{E}_0$. Since $b \notin E_0$, $b \in \bar{E}_0 \setminus E_0$.

2. The inductive hypothesis is as follows. For each $j \in \{0, \ldots, i-1\}$, suppose that we have already constructed $E_i, \bar{E}_i \in \mathcal{F}(\mathcal{A})$ such that
$$E_j = \mathcal{F}(W(Z(E_j))) \tag{8}$$
$$\bar{E}_j = \mathcal{F}(\bar{W}(Z(E_j))) \tag{9}$$
for some $Z(E_j) \in \mathcal{F}(\mathcal{A})$. Suppose that, for each j, E_j is open and $\bar{E}_j$ is closed, and that

$\{z_j\} \in \bar{E}_j \setminus E_j$. Finally, suppose that there is a permutation π of $\{0, \ldots, i-1\}$ such that

$$E_{\pi(0)} \subset \bar{E}_{\pi(0)} \subset E_{\pi(1)} \subset \bar{E}_{\pi(1)} \subset \cdots \subset E_{\pi(n-1)} \subset \bar{E}_{\pi(n-1)}. \tag{10}$$

(a) Suppose that $\{z_i\} \in \bar{E} \setminus E$ for some $E \in \{E_1, \ldots, E_{i-1}\}$. Set $Z(E_i) := Z(E)$, and set

$$E_i = \mathcal{F}(W(Z(E_i))) = E \tag{11}$$
$$\bar{E}_i = \mathcal{F}(\bar{W}(Z(E_i))) = \bar{E}. \tag{12}$$

(b) Suppose that $\{z_i\} \in \cap_{j=0}^{i-1} E_j$. Set $Z(E_i) := \{z_i\}$, and set

$$E_i := \mathcal{F}(W(Z(E_i)))$$
$$\bar{E}_i := \mathcal{F}(\bar{W}(Z(E_i))).$$

Since $z_i \in \bar{W}(z_i) \setminus W(z_i)$, $\{z_i\} \in \bar{E}_i \setminus E_i$.

Suppose that E is the smallest member of $\{E_j\}_{j=0}^{i-1}$. We show that $\bar{E}_i \subset E$. Since $\{z_i\} \in E$, $z_i \in W(Z(E))$. Fix any $X \in \mathcal{F}(\bar{W}(z_i))$. By Lemma 9, $X \in \mathcal{F}(W(Z(E))) = E$.

(c) Suppose that $\{z_i\} \notin \cup_{j=0}^{i-1} \bar{E}_j$. Suppose that E is the largest member of $\{E_k\}_{j=0}^{i-1}$. By Improvability, we can find $Z \in \mathcal{F}(\mathcal{Z})$ arbitrarily close to $Z(E)$ such that $Z(E) \subset W(Z)$. If we choose Z sufficiently close to $Z(E)$, we will have $z_i \notin W(Z)$. Suppose not. Then we have a sequence $\{Z_j \in \mathcal{F}(\mathcal{A})\}_{j=1}^{\infty} \to Z(E)$ such that $z_i \in W(Z_j)$ for all j. This implies $z_i \in \bar{W}(Z(E))$—contradiction.

Let $Z(E_i) := Z \cup \{z_i\}$, and set $E_i, \bar{E}_i$ according to (11). Since $z_i \in \bar{W}(Z \cup \{z_i\})$, and since $z_i \notin W(Z)$, $\{z_i\} \in \bar{E}_i \setminus E_i$. We show that $\bar{E} \subset E_i$. Suppose $X \in \bar{E}$, so $X \in \mathcal{F}(\bar{W}(Z(E)))$. Since $Z(E) \in \mathcal{F}(W(Z))$, Lemma 9 implies $X \in \mathcal{F}(W(Z))$, so $X \in \mathcal{F}(W(Z \cup \{z_i\})) = E_i$.

(d) Suppose that $z_i \in E_j$ for some $j \in \{0, \ldots, i-1\}$ and that $z_i \notin \bar{E}_{j'}$ for some $j' \in \{0, \ldots, i-1\}$. Suppose that E is the largest element of $\{E_j : \{z_i\} \notin \bar{E}_j\}_{j=0}^{i-1}$, and that E' is the smallest element of $\{E_j : \{z_i\} \in E_j\}_{j=0}^{i-1}$. We must have $\bar{E} \subset E'$.

As in Step 2c, we can find $Z \in \mathcal{F}(\mathcal{Z})$ arbitrarily close to $Z(E)$ such that $Z(E) \subset W(Z)$ and $z_i \notin W(Z)$. Since $Z(E) \in \bar{E} \subset E'$, and since E' is open, we can ensure $Z \in E'$ by choosing Z sufficiently close to $Z(E)$. Let $Z(E_i) := Z \cup \{z_i\}$, and set $E_i, \bar{E}_i$ according to (11).

As in Step 2c, $\{z_1\} \in \bar{E}_i \setminus E_i$, and $\bar{E} \subset E_i$. We show that $\bar{E}_i \subset E'$. Suppose that $X \in \bar{E}_i$, so $X \in \mathcal{F}(\bar{W}(Z \cup \{z_i\}))$. Since $\{z_i\} \cup Z' \in E'$, $\{z_i\} \cup Z' \in \mathcal{F}(W(Z(E')))$. By Lemma 9, $X \in \mathcal{F}(W(Z(E'))) = E'$.

Rather than proving that $\{E_i\}_{i=0}^{\infty}$ is an HP-system for R given in (7), we will prove a stronger

result. Define a new relation S by

$$X \; S \; Y \quad \iff \quad \nexists E, E' \in \{E_i\}_{i=0}^{\infty} \text{ s.t. } X \in E, \bar{E} \subset E', Y \notin E'.$$

Notice that $b \; S \; A$ because $\{b\} \in \bar{E}_0 \setminus E_0$ and $A \in \bar{E}_0$. We show that

$$X \; R \; Y \quad \implies \quad X \; S \; Y \text{ and } \neg(Y \; S \; X).$$

Suppose that $X \; R \; Y$, so $Y \in \mathcal{F}(W(X))$. Suppose that $\neg(X \; S \; Y)$, so there exist $E, E' \in \{E_i\}_{i=0}^{\infty}$ such that $X \in E$, $\bar{E} \subset E', Y \notin E'$. We have $X \in \mathcal{F}(W(Z(E))) = E$. By Lemma 9, $Y \in \mathcal{F}(W(Z(E))) = E$, which contradicts $Y \notin E' \supset E$. Thus, $X \; R \; Y$ implies $X \; S \; Y$.

Now we show that $\neg(Y \; S \; X)$. We need to find $E, E' \in \{E_i\}_{i=0}^{\infty}$ such that $Y \in E$, $\bar{E} \subset E'$, and $X \notin E'$. By Improvability, we can find $Z \in \mathcal{F}(\mathcal{Z})$ arbitrarily close to Y such that $Y \subset W(Z)$. Since $Y \subset W(X)$, we ensure $Z \subset W(X)$ by choosing Z sufficiently close to Y. Since $Z \subset \mathcal{Z}$, for each $z \in Z$, there is some $E(z) \in \{E_i\}_{i=0}^{\infty}$ such that $z \in \bar{E}(z) \setminus E(z)$. Choose the $z \in Z$ for which $E(z)$ is largest; call it z^* and write E^* instead of $E(z^*)$. We have $Z \subset \bar{E}^*$. Suppose that $X \in \bar{E}^* = \mathcal{F}(\bar{W}(Z(E^*)))$. Since $z^* \in W(X)$, Lemma 9 implies $z^* \in W(Z(E^*))$, so $\{z^*\} \in E^*$. This contradicts the definition of E^*. Conclude that $X \notin \bar{E}^*$. Since $Y \subset W(Z)$ and $Z \subset \bar{E}^* = \mathcal{F}(\bar{W}(Z(E^*)))$, Lemma 9 implies $Y \in \mathcal{F}(W(Z(E^*))) = E^*$. We can repeat the same arguments with $Z(E^*)$ in place of Y to get $Z \subset \mathcal{F}(\mathcal{Z})$ such that $Z(E^*) \subset E(Z)$ and $X \notin \bar{E}(Z)$. $Z(E^*) \subset E(Z)$ implies $\bar{E}^* \subset E(Z)$. Putting everything together, we have $Y \in E^*$, $\bar{E}^* \subset E(Z)$, and $X \notin E(Z)$ as desired. Conclude that S extends R.

Now we show that $\{E_i\}_{i=0}^{\infty}$ is indeed an HP-system for S. We check the requirements of Definition 39 in order.

1. This is implied by requirement 4. By non-triviality, there is some $(X, Y) \in \mathcal{F}(\mathcal{A})^2$ such that $Y \subset W(X)$, i.e. $X \; R \; Y$. Notice that $Y \subset W(X)$ implies $\neg(X \subset W(Y))$. Otherwise, we would have $Y \subset W(Y)$ by Lemma 9. Applying IUA gives $c(Y) = \emptyset$—contradiction. Thus, we have (X, Y) such that $(X \; R \; Y)$ and $\neg(Y \; R \; X)$. Under this condition, requirement 4 says that there exist $E, E' \in \{E_i\}_{i=1}^{\infty}$ such that $\text{cl}(E) \subset E'$.

2. From (10), it is clear that $E_i \subseteq E_j$ or $E_i \subseteq E_j$ for all $E_i, E_j \in \{E_i\}_{i=0}^{\infty}$.

3. For any E_i, E_j such that $\text{cl}(E_i) \subset E_j$, we need to find E_k such that $\text{cl}(E_i) \subset E_k \subset \text{cl}(E_k) \subset E_j$. We first show there is E_k such that $\bar{E}_i \subset E_k \subset \bar{E}_k \subset E_j$. Since E_j is open and $\bar{E}_i$ is closed, there must be some $z_k \in \mathcal{Z}$ such that $k > \max_{i,j}$ and $\{z_k\} \in E_j \setminus \bar{E}_i$. We constructed the $\{E_j\}_{j=0}^{\infty}$ so that $\{z_k\} \notin \bar{E}_i$ implies $\bar{E}_i \subset E_k$ and $\{z_k\} \in E_j$ implies $\bar{E}_k \subset E_j$. To see why, notice that $k > i$ implies $\bar{E}_i$ was already present when E_k was defined. Since $\{z_k\} \notin \bar{E}_i$, step 2d requires $\bar{E}_i \subset E_k$. Similarly, $k > j$ implies E_j was already present when $\bar{E}_k$ was defined. Since $\{z_k\} \in E_j$, step 2d requires $\bar{E}_k \subset E_j$.

Since $\text{cl}(E_i) \subseteq \bar{E}_i$, $\text{cl}(E_i) \subset E_k$ follows from $\bar{E}_i \subset E_k$. Similarly, $\text{cl}(E_k) \subset E_j$ follows from $\bar{E}_k \subset E_j$.

4. Take any $(X, y) \in \mathcal{F}(\mathcal{A}) \times \mathcal{A}$ such that $X \; S \; \{y\}$. Suppose that $X \in E_i$ for some $E_i \in \{E_i\}_{i=0}^{\infty}$. We show that $\{y\} \in E_i$. By definition of S, there cannot be $E, E' \in \{E_i\}_{i=0}^{\infty}$ such that $X \in E$, $\bar{E} \subset E'$, and $\{y\} \notin E'$. It suffices to show that, for all $E \in \{E_i\}_{i=0}^{\infty}$, $X \in E$ implies that there exists $E' \in \{E_i\}_{i=0}^{\infty}$ such that $X \in E' \subset \bar{E}' \subset E$.

   The argument is very similar to the one we used to show that S extends R. $X \in E$ means $X \in \mathcal{F}(W(Z(E)))$. By Improvability, we can find $Z \in \mathcal{F}(\mathcal{Z})$ arbitrarily close to X such that $X \in W(Z)$. By choosing Z sufficiently close to X, we will have $Z \in \mathcal{F}(W(Z(E))) = E$. Since $Z \subset \mathcal{Z}$, for each $z \in Z$, there is some $E(z) \in \{E_i\}_{i=0}^{\infty}$ such that $z \in \bar{E}(z) \setminus E(z)$. Choose the $z \in Z$ for which $E(z)$ is largest; call it z^* and write E^* instead of $E(z^*)$. We have $Z \subset \bar{E}^*$. Suppose that $\bar{E} \subset \bar{E}^*$, so $Z(E) \subset \bar{W}(Z(E^*))$. Since $z^* \in Z \subset W(Z(E))$, Lemma 9 implies $z^* \in W(Z(E^*))$, so $\{z^*\} \in E^*$—contradiction. Conclude that $\bar{E}$ is a strict superset of $\bar{E}^*$, so $\bar{E}^* \subset E$. Since $X \subset W(Z)$ and $Z \subset \bar{E}^* = \mathcal{F}(\bar{W}(Z(E^*)))$, Lemma 9 implies $X \in \mathcal{F}(W(Z(E^*))) = E^*$. Putting everything together, $X \in E^* \subset \bar{E}^* \subset E$ as desired.

5. Take any $(X, Y) \in \mathcal{F}(\mathcal{A})^2$ such that $X \; S \; Y$ and $\neg(Y \; S \; X)$. $\neg(Y \; S \; X)$ implies that there exist $E, E' \in \{E_i\}_{i=0}^{\infty}$ such that $Y \in E$, $\bar{E} \subset E'$, and $X \notin E'$.

By HP Theorem 3.4 (and the fact that $\mathcal{A}$ is a separable metric space), there exists a continuous $f : \mathcal{F}(\mathcal{A}) \to \mathbb{R}$ such that, for any $(X, Y) \in \mathcal{F}(\mathcal{A})^2$,

$$\begin{aligned} X \; S \; Y \quad &\implies \quad f(X) \geq f(Y) \\ X \; S \; Y \text{ and } \neg(Y \; S \; X) \quad &\implies \quad f(X) > f(Y). \end{aligned}$$

Since $Y \in \mathcal{F}(W(X))$ implies $X \; S \; Y$ and $\neg(Y \; S \; X)$, $Y \in \mathcal{F}(W(X))$ implies $f(X) > f(Y)$. Since $b \; S \; A$, $f(\{b\}) \geq f(A)$.

Take any $X \in \mathcal{F}(\mathcal{A})$. It is easy to see that $X \; S \; \{x\}$ for all $x \in X$. For some $x \in X$, we will also have $\{x\} \; S \; X$. To see why, notice that there must be some $x \in X$, denoted x^*, such that $\{x^*\} \in E$ implies $X \in E$ for all $E \in \{E_i\}_{i=0}^{\infty}$. Suppose that $\neg(\{x^*\} \; S \; X)$, so there exist $E, E' \in \{E_i\}_{i=0}^{\infty}$ such that $X \notin E'$, $\bar{E} \subset E'$, and $\{x^*\} \in E$. By definition of x^*, $\{x^*\} \in E$ implies $X \in E$, so $X \in E'$. We conclude that $\{x^*\} \; S \; X$. We must have $f(\{x^*\}) \geq f(X) \geq \max_{x \in X} f(\{x\})$. This implies $f(X) = \max_{x \in X} f(\{x\})$. Define $m^* : \mathcal{A} \to \mathbb{R}$ by

$$m^*(\cdot) := f(\{\cdot\}).$$

Since $f(\{b\}) \geq f(A) = \max_{a \in A} f(\{a\})$, we have $m^*(b) \geq \max_{a \in A} m(a)$. Since, for any $(X, y) \in \mathcal{F}(\mathcal{A}) \times \mathcal{A}$, $y \in W(X)$ implies $\max_{x \in X} f(\{x\}) = f(X) > f(Y) = \max_{y \in Y} f(\{y\})$, m^* respects

exclusion. Since m^* inherits continuity from f, $m^* \in \mathcal{M}$ as desired. This completes the construction of m^*.

Since (A, b) was chosen arbitrarily from $\{\mathcal{F}(\mathcal{A}) \times \mathcal{A} : b \in c(\{b\} \cup A)\}$, we conclude that $c(A) \subseteq \mathcal{M}(A)$.

Since $\succsim$ is continuous, there is some $u \in C(A, \mathbb{R})$ such that u represents $\succsim$. Since $c(A) \subseteq \mathcal{M}(A)$ and since

$$a \succsim c(A) \text{ and } a \notin c(A) \quad \Longrightarrow \quad a \notin \mathcal{M}(A),$$

$(u, \mathcal{M})$ represents $(\succsim, c)$. We show that $(u, \mathcal{M})$ satisfies local non-satiation, recoverability, and closedness. For local non-satiation: take any $a \in \mathcal{A}$. By Improvability, we can find $Z \in \mathcal{F}(\mathcal{Z})$ arbitrarily close to a such that $a \in W(Z)$. Since the preferences in $\mathcal{M}$ respect exclusion, Z is strictly preferred to a by $\mathcal{M}$.

For recoverability: take any $(B, a) \in \mathcal{F}(\mathcal{A}) \times \mathcal{A}$ such that $\max_{b \in B} m(b) > m(a)$ for all $m \in \mathcal{M}$. We show that $\max_{b \in B : u(b) \leq u(a)} m(b) > m(a)$ for all $m \in \mathcal{M}$. It suffices to show that

$$\exists m \in \mathcal{M} \text{ s.t. } m(a) \geq \max_{b \in B : u(b) \leq u(a)} m(b) \quad \Longrightarrow \quad \exists m \in \mathcal{M} \text{ s.t. } m(a) \geq \max_{b \in B} m(b).$$

Let $B' := \{b \in B : u(b) \leq u(a)\}$, and suppose $m(a) \geq \max_{b \in B'} m(b)$ for some $m \in \mathcal{M}$. Since $(u, \mathcal{M})$ represents $(\succsim, c)$, $a \succsim B'$ and $a \in c(\{a\} \cup B')$, so $a \notin W(B')$. Since $B \setminus B' \succ a$, $a \notin W(B)$. We can construct $\hat{m} \in \mathcal{M}$ such that $\hat{m}(a) \geq \max_{b \in B} \hat{m}(b)$ in exactly the same way that we constructed m^* above. (Notice that the construction does not require $a \in c(\{a\} \cup B)$, which might not hold; it only requires $a \notin W(B)$, which we have just shown.)

For closedness: since $W(B)$ is open for all $B \in \mathcal{F}(\mathcal{A})$, it suffices to show that

$$\bigcap_{m \in \mathcal{M}} \{a \in \mathcal{A} : m(a) < \max_{b \in B} m(b)\} = W(B).$$

Since each $m \in \mathcal{M}$ respects exclusion, $a \in W(B)$ implies $m(a) < \max_{b \in B} m(B)$. Now suppose $m(a) < \max_{b \in B} m(b)$ for all $m \in \mathcal{M}$. Since recoverability holds, we have $m(a) < \max_{b \in B'} m(b)$ where $B' := \{b \in B : u(a) \geq u(b)\}$ for all $m \in \mathcal{M}$. Since $(u, \mathcal{M})$ represents $(\succsim, c)$, $a \succ B'$ and $a \notin c(\{a\} \cup B')$. By definition of W, $a \in W(B')$, so $a \in W(B)$.

A.4 Proof of Theorem 3

The following piece of notation is useful for both necessity and sufficiency. Recall the definition of $B(p)$ in (1). Let

$$NB(p) := \{q \in \Delta(Z) : p \succsim q\} \setminus B(p) = \{q \in \Delta(Z) : p \succsim q \text{ and } p \in \mathcal{M}(\{p, q\})\}.$$

First, we show necessity. It is well known that the first parts of Continuity and Independence are necessary (and sufficient) for $\succsim$ to have an EU representation. The second part of Continuity says that $NB(p)$ is closed. Take any convergent sequence (q_n) such that each $q_n \in NB(p)$. By definition, $q_n \succsim p$ and $p \in c(\{q_n, p\})$ for all n. Since $\succsim$ is continuous, $q \succsim p$. For each n, we have $(m_n)'p \geq (m_n)'q$ for some $m_n \in \mathcal{M}$. Since $\mathcal{M}$ is compact, some subsequence of m_n has a limit $m \in \mathcal{M}$. We will have $m'p \geq m'q$, so $p \in c(\{q, p\})$. We conclude that $q \in NB(p)$, so $NB(p)$ is closed.

Now consider Convexity. Suppose that A excludes p. Let $\underline{A} := c(A \cup \{p\}) \cup \{a \in A : p \succ a\}$. By Lemma 1, $p \notin c(\underline{A} \cup \{p\})$. For each $m \in \mathcal{M}$, we have some $a \in \underline{A}$ such that $m'p < m'a$. It is without loss to assume that $m'p = 0$ for all $m \in \mathcal{M}$. We want to find a set of weights α such that $p \succsim \sum_{a \in \underline{A}} \alpha(a)\delta_a$ and

$$\sum_{a \in \underline{A}} \alpha(a) m'a > 0$$

for all $m \in \mathcal{M}$. The first part is easy—it will hold for any α since $\succsim$ is EU—so we focus on the second. For each $m \in \mathcal{M}$, let $m_{\underline{A}} := (m'a)_{a \in \underline{A}}$. Let $\mathcal{M}_{\underline{A}}$ be the set of the $m_{\underline{A}}$. Like $\mathcal{M}$, $\mathcal{M}_{\underline{A}}$ is nonempty, compact and convex. Let $N := \mathbb{R}_{-}^{|\underline{A}|}$, which is nonempty, closed and convex. Notice that no element of $\mathcal{M}_{\underline{A}}$ can be weakly negative (otherwise, some $m \in \mathcal{M}$ would rank p weakly higher than each member of $\underline{A}$). Thus, $\mathcal{M}_{\underline{A}}$ and N are disjoint, and we can apply the Separating Hyperplane Theorem. This delivers a nonzero $\alpha \in \mathbb{R}^{|\underline{A}|}$ and $c \in \mathbb{R}$ such that $\alpha' n < c < \alpha' m_A$ for all $n \in N, m_A \in \mathcal{M}_{\underline{A}}$. Since the zero vector belongs to N, we must have $c > 0$. Suppose the ith element of α is strictly negative. By choosing n with a sufficiently negative number in ith position and zeros elsewhere, we get $\alpha' n > c$, a contradiction. Thus, each element of α is weakly positive. If we rescale α to a unit sum, we still have $\alpha' m_A > 0$ for all $m_A \in \mathcal{M}_{\underline{A}}$. We can rewrite this as $\sum_{a \in \underline{A}} \alpha(a) m'a > 0$ for all $m \in \mathcal{M}$, which is exactly what we needed.

For the other direction of Convexity, suppose $p \in c(A \cup \{p\})$. For some $m \in \mathcal{M}$, we have $m'p \geq m'a$ for all $a \in A$. For this m, we clearly have $m'p \geq m'a$ for all $a \in \text{co}(A)$. No $a \in \text{co}(A)$ can belong to $B(p)$.

Consider the second part of Monotonicity. Suppose that $p >_{FOSD} q$ and that $A \supset \{p, q\}$. First, we show that $q \notin c(A)$. Suppose otherwise, so $m'q = \max_{a \in A} m'a$ for some $m \in \mathcal{M}$. Since each $m \in \mathcal{M}$ is weakly FOSD-monotone, we have $m'p = \max_{a \in A} m'a$, so $p \in \mathcal{M}(A)$. Since $\succsim$ is strictly FOSD-monotone and $p >_{FOSD} q$, it cannot be that $q \in c(A)$. To confirm that $c(A) = c(A \setminus \{q\})$, it now suffices to show that $\mathcal{M}(A) \setminus \{q\} = \mathcal{M}(A \setminus \{q\})$. Take any $r \in \mathcal{M}(A \setminus \{q\})$, so $m'r = \max_{a \in A \setminus \{q\}} m'a$. Since $p \in A \setminus \{q\}$ and each $m \in \mathcal{M}$ is weakly FOSD-monotone, we have $m'r = \max_{a \in A} m'a$, so $r \in \mathcal{M}(A)$. Now take any $r \in \mathcal{M}(A) \setminus \{q\}$, so $m'r = \max_{a \in A} m'a \geq \max_{a \in A \setminus \{q\}} m'a$. Since $r \neq q$, we have $r \in \mathcal{M}(A \setminus \{q\})$.

Now we show sufficiency.

Lemma 10. *$B(p)$ is a convex cone.*

Proof. First, suppose that $q \in B(p)$, so $p \succsim q$ and $\{q\} = c(\{p, q\})$. By $\succsim$-Independence, $p \succsim \alpha p + (1-\alpha)q$ for all $\alpha \in (0,1)$. By c-Independence, $\{\alpha p + (1-\alpha)q\} = c(\{p, \alpha p + (1-\alpha)q\})$, so $\alpha p + (1-\alpha)q \in B(p)$. Similarly, suppose that $\alpha p + (1-\alpha)q \in B(p)$, so $p \succsim \alpha p + (1-\alpha)q$ and $\{\alpha p + (1-\alpha)q\} = c(\{\alpha p + (1-\alpha)q, p\})$. By $\succsim$-Independence, $p \succsim q$. By c-Independence, $\{q\} = c(\{q, p\})$, so $q \in B(p)$. Now suppose that $q, r \in B(p)$, so $p \succsim q, r$ and $\{q\} = c(\{p, q\})$, $\{r\} = c(\{p, r\})$. By $\succsim$-Independence, $p \succsim \alpha q + (1-\alpha)r$ for all $\alpha \in (0,1)$. By $\succsim$- and c-Independence,

$$\begin{aligned} \alpha p + (1-\alpha)r &\succsim \alpha q + (1-\alpha)r \\ \{\alpha q + (1-\alpha)r\} &= c(\{\alpha q + (1-\alpha)r, \alpha p + (1-\alpha)r\}), \end{aligned}$$

so $\alpha q + (1-\alpha)r$ excludes $\alpha p + (1-\alpha)r$ from below. Also by $\succsim$- and c-Independence,

$$\begin{aligned} p &\succsim \alpha p + (1-\alpha)r \\ \{\alpha p + (1-\alpha)r\} &= c(\{p, \alpha p + (1-\alpha)r\}), \end{aligned}$$

so $\alpha p + (1-\alpha)r$ excludes p from below. By IUA, we can add p to a set containing $\alpha p + (1-\alpha)r$ without affecting choice, so

$$\{\alpha q + (1-\alpha)r\} = c(\{\alpha p + (1-\alpha)r, \alpha q + (1-\alpha)r, p\}).$$

Also by IUA, we can remove $\alpha p + (1-\alpha)r$ from a set containing $\alpha q + (1-\alpha)r$ without affecting choice, so

$$\{\alpha q + (1-\alpha)r\} = c(\{\alpha q + (1-\alpha)r, p\})$$

so $\alpha q + (1-\alpha)r \in B(p)$. □

Lemma 11. *Fix $p, p', q, q' \in \Delta(Z)$ such that $q - p = q' - p'$. If $q \in B(p)$, then $q' \in B(p')$.*

Proof. We have

$$\frac{1}{2}p + \frac{1}{2}q' = \frac{1}{2}p' + \frac{1}{2}q.$$

Since $q \in B(p)$, $p \succsim q$ and $\{q\} = c(\{p, q\})$. By c-Independence,

$$\begin{aligned} \left\{\frac{1}{2}q + \frac{1}{2}p'\right\} &= c\left(\left\{\frac{1}{2}q + \frac{1}{2}p', \frac{1}{2}p + \frac{1}{2}p'\right\}\right) \\ \left\{\frac{1}{2}p + \frac{1}{2}q'\right\} &= c\left(\left\{\frac{1}{2}p + \frac{1}{2}q', \frac{1}{2}p + \frac{1}{2}p'\right\}\right) \\ \{q'\} &= c(\{p', q'\}). \end{aligned}$$

Similarly, by $\succsim$-Independence,

$$\frac{1}{2}p + \frac{1}{2}p' \succsim \frac{1}{2}q + \frac{1}{2}p'$$
$$\frac{1}{2}p + \frac{1}{2}p' \succsim \frac{1}{2}p + \frac{1}{2}q'$$
$$p' \succsim q'.$$

We conclude that $q' \in B(p')$. □

We are now ready to define $\mathcal{M}$. Take p in the interior of $\Delta(Z)$. Take any supporting hyperplane H of $B(p)$ that passes through some boundary point b of $B(p)$ with $p \succ b$. Since $B(p)$ is a cone with vertex p, H will also pass through p. Since $B(p)$ is open in $\{q \in \Delta(Z) : p \succsim q\}$, H cannot include any point in $B(p)$. Let $\mathcal{H}$ be the set with generic member H. For each H, take a unit-norm $m \in \mathbb{R}^{|Z|}$ such that $m'h = 0$ for all $h \in H$ (including p) and $m'q > 0$ for all $q \in B(p)$. Collect these m, and take the closed convex hull. This is $\mathcal{M}$.

We show that everything in $\mathcal{M}$ is weakly D-monotone. It suffices to show that any preference with an indifference curve in $\mathcal{H}$ that has $B(p) \succ p$ is weakly D-monotone. (All the other preferences are combinations and/or limits of these, so will inherit weak D-monotonicity.) Suppose that some preference $\succsim_h$ with an indifference curve in $\mathcal{H}$ has $r \succ_h q$ even though $q >_{FOSD} r$. By definition of $\mathcal{H}$, we have $b \sim_h p$ for some boundary point b of $B(p)$ such that $p \succ b$. It is without loss to assume that b is interior, so there exists $\lambda > 0$ such that $b + \lambda(q - r) \in \Delta(Z)$. Since $p \succ b$, we will have $p \succ b + \lambda(q - r)$ for λ small enough. We also have $b \succ_h b + \lambda(q - r)$ and $b + \lambda(q - r) >_{FOSD} b$. Since b is a boundary point of $B(p)$, we can find $\tilde{b}$ arbitrarily close to b such that $\tilde{b} \in B(p)$. Since $p \sim_h b \succ_h b + \lambda(q - r)$, we can ensure $p \succ_h \tilde{b} + \lambda(q - r)$ by choosing $\tilde{b}$ sufficiently close to b. This implies $\tilde{b} + \lambda(q - r) \notin B(p)$. Since $p \succ b + \lambda(q - r)$, we can also ensure $p \succ \tilde{b} + \lambda(q - r)$ by choosing $\tilde{b}$ sufficiently close to b. This implies $\tilde{b} + \lambda(q - r) \in NB(p)$. Now we can derive a contradiction. To simplify the notation, let $b^* = \tilde{b} + \lambda(q - r)$. Notice that $b^* >_{FOSD} \tilde{b}$, so $c(\{p, b^*, \tilde{b}\}) = c(\{p, b^*\})$ by the second part of Monotonicity. Since $b^* \notin NB(p)$, we have $p \in c(\{p, b^*\})$, so $p \in c(\{p, b^*, \tilde{b}\})$. But since $\tilde{b} \in B(p)$, IUA implies $p \notin c(\{p, b^*, \tilde{b}\})$. Conclude that $\succsim_h$ is weakly FOSD-monotone, so every preference in $\mathcal{M}$ is weakly D-monotone.

We now show

$$B(p) = \bigcap_{m \in \mathcal{M}} \{q \in \Delta(Z) : m'q > m'p\} \cap \{q \in \Delta(Z) : p \succsim q\}. \tag{13}$$

Suppose $b \notin B(p)$, but $b \in \bigcap_{m \in \mathcal{M}} \{q \in \Delta(Z) : m'q > m'p\}$ and $p \succsim b$. By construction of $\mathcal{M}$, b must be a boundary point of $B(p)$ that is not a limit of points in $NB(p)$ that are strictly worse than p. Clearly, $p \sim b$. Moreover, for any $q \prec p$, there must be some α sufficiently close to 1 such

that

$$\alpha b + (1-\alpha)q \in B(p).$$

(If this were not the case, then b could be written as a limit of points in $NB(p)$ that are strictly worse than p.) Take r such that $p >_{FOSD} r$. (p is interior, so some such r must exist.) Since the primary preference is strictly FOSD-monotone, we have $p \succ r$. We also have $\alpha \in (0,1)$ such that

$$\alpha b + (1-\alpha)r \in B(p).$$

That is, $p \notin c(\{p, \alpha b + (1-\alpha)r\})$ even though $p \succsim \alpha b + (1-\alpha)r$. By the second part of Convexity, $p \notin c(\{p, b, r\})$. By the second part of Monotonicity and $p >_{FOSD} r$, $c(\{p,b,r\}) = c(\{p,b\})$, so $p \notin c(\{p,b\})$. Since $p \sim b$, this implies $b \in B(p)$—contradiction.

Now suppose $B(p) \nsubseteq \bigcap_{m\in\mathcal{M}} \{q \in \Delta(Z) : m'q > m'p\} \cap \{q \in \Delta(Z) : p \succsim q\}$. By construction of $\mathcal{M}$, this can only happen if $b \sim p$, and there is a sequence $\{H\}_{n=1}^{\infty}$ of hyperplanes in $\mathcal{H}$ converging to $\{q \in \Delta(Z) : p \sim q\}$. Recall that each hyperplane in $\mathcal{H}$ passes through some boundary point of $B(p)$ that is strictly worse than p. Take the sequence of such points corresponding to $\{H\}_{n=1}^{\infty}$. Passing to a subsequence if necessary, let b be the limit of this sequence of points. For any sufficiently small perturbation $\tilde{b}$ of b with $\tilde{b} \prec p$, we must have $\tilde{b} \in B(p)$. (Otherwise, $\{q \in \Delta(Z) : p \sim q\}$ could not be the limit of $\{H\}_{n=1}^{\infty}$.) We can now apply the argument in the previous paragraph. Take any r such that $p >_{FOSD} r$. We must have $\alpha b + (1-\alpha)r \in B(p)$ for some $\alpha \in (0,1)$. Applying the second part of Convexity, $p \notin c(\{p,b,r\})$. Applying the second part of Monotonicity, $c(\{p,b,r\}) = c(\{p,b\})$, so $p \notin c(\{p,b\})$, so $b \notin NB(p)$. Since b is a limit point of $NB(p)$, this contradicts the second part of Continuity.

Notice that it does not matter which p we use to define $\mathcal{M}$, since Lemma 11 ensures that we will get the same set of utilities (up to an irrelevant additive constant) for any interior p. To finish the proof, we have to show that $\mathcal{M}$ satisfies two conditions. First, no utility in $\mathcal{M}$ would justify choosing an item that the DM doesn't choose, but likes as much as anything he does choose. Second, for any item the DM chooses, some utility in $\mathcal{M}$ justifies it.

Consider the first part. Suppose $q \succsim c(A \cup \{q\})$ and $q \notin c(A \cup \{q\})$. Let $\underline{A} = c(A \cup \{q\}) \cup \{a \in A : q \succ a\}$. By Lemma 1, $q \notin c(\underline{A} \cup \{q\})$. By Convexity, we can find $a^* \in \text{co}(\underline{A})$ such that $q \notin c(\{q, a^*\})$. Since $q \succsim a^*$, $a^* \in B(q)$. By (13), $m'a^* > m'q$ for all $m \in \mathcal{M}$. For each $m \in \mathcal{M}$, we must have $a \in \underline{A}$ such that $m'a > m'q$. This is exactly what we needed.

For the second part, suppose $q \in c(A \cup \{q\})$. To start, suppose $q \succsim A$. Suppose that we cannot find $m \in \mathcal{M}$ so that $m'q \geq m'a$ for all $a \in A$. Recall the argument we used to show necessity of Convexity. Since $\mathcal{M}$ is compact and convex, we can use the same argument to find an $a^* \in \text{co}(A)$ such that

$$a^* \in \bigcap_{m\in\mathcal{M}} \{r \in \Delta(Z) : m'r > m'q\} \cap \{r \in \Delta(Z) : q \succsim r.\}$$

By (13), $a^* \in \text{co}(A) \cap B(q)$. By Convexity, $q \notin c(A \cup \{q\})$, a contradiction.

Now we relax the assumption that $q \succsim A$. Let $\underline{A} = \{a \in A : q \succsim a\}$. We know $q \in c(\underline{A} \cup \{q\})$. (Suppose not. Then $q \notin c(A \cup \{q\})$ by IUA, a contradiction.) By the previous argument, we can find $m^* \in \mathcal{M}$ such that $(m^*)'q \geq \max_{a \in \underline{A}}(m^*)'a$. Now take any item $\bar{a} \in A \setminus \underline{A}$. Since $q \in c(\{q\} \cup A)$ and $\bar{a} \succ q$, $\bar{a} \notin c(\{q\} \cup A)$. By Lemma 1, $\bar{a} \notin c(\{\bar{a}, q\} \cup \underline{A})$ even though $\bar{a} \succ \{q\} \cup \underline{A}$. We already showed that there is no $m \in \mathcal{M}$ such that $m'\bar{a} \geq \max_{x \in \underline{A} \cup \{q\}} m'x$. In particular, it cannot be that $(m^*)'\bar{a} \geq (m^*)'q \geq max_{a \in \underline{A}}(m^*)'a$, so $(m^*)'q \geq \max_{a \in \underline{A} \cup \bar{a}}(m^*)'a$. Since $\bar{a}$ was an arbitrary selection from $A \setminus \underline{A}$, we have $(m^*)'q \geq \max_{a \in A} m'a$.

A.5 Proof of Corollary 1

For the minimal set, recall the construction in the proof of Theorem 3. We start with p in the interior of $\Delta(Z)$. Then we take the supporting hyperplanes of $B(p)$ that pass through some boundary point of $B(p)$ that is strictly worse than p. $\mathcal{H}$ is the set of such hyperplanes. Let $\bar{co}(\mathcal{H})$ be the closed convex hull of $\mathcal{H}$. Let $\mathcal{M}_\succ$ denote the set of preferences with representations in $\mathcal{M}$. $\mathcal{M}_\succ$ is precisely the set of EU preferences that have indifference curves in $\bar{co}(\mathcal{H})$ and that strictly prefer $B(p)$ to p.

We show that $\mathcal{M}_\succ$ is minimal. Take any boundary point b of $B(p)$ such that $p \succ b$. Since $B(p)$ is open in $\{q \in \Delta(Z) : p \succsim q\}$, $b \notin B(p)$, so $p \in c(\{p, b\})$. Clearly, every set of justifiable preferences in every representation must contain some preference $\succsim_m$ such that $p \succsim_m b$. Take any closed, convex proper subset of $\mathcal{M}_\succ$, and call it $\mathcal{N}_\succ$. Recall that $\mathcal{M}_\succ$ is the closed, convex hull of the set of EU preferences that are indifferent between p and a boundary point of $B(p)$ that is strictly worse than p. Suppose that, for each boundary point b of $B(p)$ such that $b \succ p$, there exists $\succsim_n \in \mathcal{N}_\succ$ such that $p \sim_n b$. Since $\mathcal{N}_\succ$ is convex and closed, this implies $\mathcal{N}_\succ \supseteq \mathcal{M}_\succ$—contradiction.

Now we turn to the maximal set. We will modify the construction in the proof of Theorem 3. As before, take p in the interior of $\Delta(Z)$. Let $\mathcal{M}_\succ^{max}$ be the set of weakly D-monotone EU preferences that strictly prefer $B(p)$ to p. If $\mathcal{M}_\succ^{max}$ is indeed a representation, it is obviously maximal. It is easy to see that $\mathcal{M}_\succ$ is convex. Suppose that it is not closed, so there exists a point $b \in B(p)$ such that $p \succsim_m b$ for some limit of D-monotone preferences that strictly prefer $B(p)$ to p. The argument will be familiar from the proof of Theorem 3. Notice that b must be on the boundary of $B(p)$. Also, since $b \notin NB(p)$ and $NB(p)$ is closed, b cannot be a limit of points in $NB(p)$. In this situation, only one preference with $p \succsim_m b$ could possibly be in $\mathcal{M}_\succ^{max}$: the preference exactly opposite $\succsim$. But this preference is not weakly D-monotone, so it is not in $\mathcal{M}_\succ^{max}$.

We can define a compact, convex set of utility representations for the maximal set of EU preferences just as we did for the minimal set. Call the result $\mathcal{M}^{max}$. To confirm that (13) holds, suppose that $b \notin B(p)$ but $p \succsim b$ and $m'b \succ m'p$ for all $m \in \mathcal{M}^{\max}$. That is, every D-monotone EU preference that strictly prefers $B(p)$ to p also prefers b to p. We already know that the preferences in the minimal set are D-monotone and prefer $B(p)$ to p, so they must prefer b to p as well. But

since the minimal set of EU preferences represents $(\succsim, c)$, we must have $\{b\} = c(\{p, b\})$, which contradicts $b \notin B(p)$. The rest of the proof of Theorem 3 relies only on (13), so it goes through as before.

A.6 Proof of Corollary 2

Let $\mathcal{M}^1_\succ$ be the maximal set of EU preferences corresponding to $(\succsim, c_1)$. Recall that $\mathcal{M}^2_\succ$ is the set of D-monotone EU preferences that strictly prefer $B^2(p)$ to p. If $B_1(p) \supset B_2(p)$, then $B^1(p) \succ_m p$ implies $B^2(p) \succ_m p$. Thus, $\mathcal{M}^2_\succ$ includes all the D-monotone preferences that strictly prefer $B^2(p)$ to p, so $\mathcal{M}^2_\succ \supseteq \mathcal{M}^1_\succ$. To confirm that the inclusion is strict, take any point $b \in B_1(p)$ such that $b \notin B_2(p)$. Since $b \in B_1(p)$, we have $p \notin c_1(\{p, b\})$, so there does not exist $\succsim_m \in \mathcal{M}^1_\succ$ such that $p \succsim_m b$. Since $b \notin B_2(p)$, we have $p \in c_2(\{p, b\})$, so there exists $\succsim_m \in \mathcal{M}^2_\succ$ such that $p \succsim_m b$.

For the converse, suppose $\mathcal{M}^1_\succ \subset \mathcal{M}^2_\succ$. Suppose that there exists $b \in B^2(p)$ such that $b \notin B^1(p)$. We must have $p \notin c_2(\{p, b\})$, so we must not have any $\succsim_m \in \mathcal{M}^2_\succ$ such that $p \succsim_m b$. On the other hand, we must have $p \in c_1(\{p, b\})$, so we must have $\succsim_m \in \mathcal{M}^1_\succ$ such that $p \succsim_m b$. Since this preference cannot belong to $\mathcal{M}^2_\succ$, we have a contradiction. Conclude that $B^2(p) \subseteq B^1(p)$. To confirm that the inclusion is strict, suppose that $B^2(p) = B^1(p)$. Since the maximal set of EU preferences is precisely the set of D-monotone EU preferences that prefer $B(p)$ to p, we have $\mathcal{M}^1_\succ = \mathcal{M}^2_\succ$—contradiction.

A.7 Proof of Corollary 2

If the revealed preference relation for c is acyclic, then there exists a strict preference on $\succ$ that extends it. For any such $\succ$, $(\succ, c)$ satisfies IUA. To see why, suppose that $b \succ c(\{b\} \cup A)$. We need to show that $c(\{b\} \cup B) = c(B)$ for any $B \supseteq A$. Suppose that $c(\{b\} \cup B) \neq c(B)$ for some $B \supseteq A$. Then, $c(A \cup \{b\})$ is revealed preferred to b, so $c(A \cup \{b\}) \succ b$—contradiction. By Theorem 1, $(\succ, c)$ has a justification representation $\mathcal{M}$—so c has a justification representation $(\succ, \mathcal{M})$.

Suppose that c has a justification representation $(\succ, \mathcal{M})$, so $(\succ, c)$ has a justification representation $\mathcal{M}$. Fix any A, b such that $b \neq c(A \cup \{b\})$ and, for some $B \supseteq A$, $c(B) \neq c(B \cup \{b\})$. Since IUA is necessary, it must be that $c(A \cup \{b\}) \succ b$. That is, $\succ$ must extend revealed preference, so revealed preference must be acyclic.

A.8 Proof of Proposition 3

This proof builds on that of Theorem 4.

Suppose that no justifiable preference in any representation has $a \succ_m B$. Consider the $\mathcal{M}$ constructed in the proof of Theorem 4. In that proof, we showed that there exists $\succ_m$ such that $a \succ_m B$ unless there is a revealed-exclusion-tree from a subset of A to b. Suppose there is a revealed-exclusion-tree from $A' \subset A$ to b.

Consider the $\succ$ constructed in the proof of Theorem 4. By Lemma 13, $y \succ X$ whenever y is revealed excluded by X. Thus, each node in a revealed-exclusion-tree must be strictly $\succ$-better than all its parents. Since we have a revealed-exclusion-tree from A' to b, and since $\succ$ is transitive, we must have $b \succ A'$. As shown in the proof of Theorem 4, we must also have $b \neq c(A' \cup \{b\})$.

Take any $A'' \subset A'$ such that $b \neq c(A'' \cup \{b\})$ but $b = c(A''' \cup \{b\})$ for any strict subset A'' of A'. Since $b \neq c(A' \cup \{b\})$, there must be some such A''. We show that b is revealed excluded by A''.

Suppose that A'' is not a singleton, and choice on the proper subsets of $A'' \cup \{b\}$ violates WARP. Then, a subset of $A'' \cup \{b\}$ is a cycle or almost-WARP set. We conclude that some item in $A'' \cup \{b\}$ is revealed excluded by a subset of $A'' \cup \{b\}$. Since $b = c(\{b\} \cup A''')$ for every strict subset A''' of A'', b cannot be revealed excluded. Suppose $a \neq b$ is revealed excluded. By IEA, $b \neq c(\{b\} \cup A'') = c(\{b\} \cup A'' \setminus \{a\}) = b$—contradiction. Conclude that choice on the proper subsets of $A'' \cup \{b\}$ satisfies WARP. Since b is chosen over every proper subset of A'', but not over A'' itself, choice on A'' violates WARP. Conclude that A'' is almost-WARP, and that b is revealed excluded by A''.

Now suppose that a is revealed excluded by B. Suppose $b = \{b\}$, so there is a chain from a to b, and $b = c(\{a, b\})$. Recall that $a \succ b \succ d$ for any cycle (a, b, d) and any $\succ$ in any representation. This implies $x_1 \succ x_2 \succ \cdots x_n$ for any chain $(x_1, \ldots, x_n)$ and any $\succ$ in any representation. Since there is a chain from a to b, we have $a \succ b$ for any $\succ$ in any representation. Since $b = c(\{a, b\})$, we have $b \notin \mathcal{M}(\{a, b\})$ for any $\mathcal{M}$ in any representation.

Now suppose $|B| > 1$, so $\{a\} \cup B$ is an almost-WARP set, and $a = c(\{a, b\})$ for all $b \in B$. We show that every $\succ$ in every representation agrees with pairwise choice on $\{a\} \cup B$. Index the items in $\{a\} \cup B$ from pairwise-best to pairwise-worst: $x_1, \ldots, x_n$, where $x_1 = a$. Now suppose there is a representation in which $x_j \succ x_i$, for $j > i$. Since $x_i = c(\{x_i, x_j\})$, $x_i \succ_m x_j$ for all $\succ_m \in \mathcal{M}$. This implies $c(\{a\} \cup B) = c(\{a\} \cup B \setminus \{x_j\})$. Since $\{a\} \cup B$ is almost-WARP, $a \neq c(\{a\} \cup B)$ but $a = c(\{a\} \cup B')$ for all proper subsets B' of B. Thus, $c(\{a\} \cup B) = c(\{a\} \cup B \setminus \{x_j\})$ holds only if $x_j = a$. But $a = x_1$ and $j > i \geq 1$—contradiction. Conclude that every $\succ$ in every representation agrees with pairwise choice on $\{a\} \cup B$. In particular, every $\succ$ has $a \succ B$. This implies $a \notin \mathcal{M}(\{a\} \cup B)$ for every $\mathcal{M}$ in every representation.

A.9 Proof of Theorem 4

Write $x \; C \; y$ if there is some z such that (x, y, z) is a cycle, or (z, x, y) is a cycle.

Lemma 12. *Define a binary relation $\succ$ by $a \succ b$ if (1) $(a, b) \in tr(C)$ or (2) $(a, b) \notin tr(C)$, $(a, b) \notin tr(C)$, and $a = c(\{a, b\})$. Then, $\succ$ is a strict preference.*

Proof. Clearly, $\succ$ is complete. Suppose it contains a cycle: $x_1 \succ x_2 \succ x_n$ where $x_1 = x_n$. For each adjacent pair (x_i, x_{i+1}), either (1) $(x_i, x_{i+1}) \in \text{tr}(C)$, or (2) $(x_i, x_{i+1}) \notin \text{tr}(C)$, $(x_{i+1}, x_i) \notin \text{tr}(C)$, and $x_i = c(\{x_i, x_{i+1}\})$.

We show that $(x_i, x_{i+1}) \in \text{tr}(C)$ for some i. Suppose not. Then we have $x_i = c(\{x_i, x_{i+1}\})$ for each i. If $x_1 = c(\{x_{n-2}, x_1\})$, then one of the following is a cycle: (x_{n-2}, x_{n-1}, x_1), (x_{n-1}, x_1, x_{n-2}), (x_1, x_{n-2}, x_{n-1}). Then, $x_{n-2}\ C\ x_{n-1}$ or $x_{n-1}\ C\ x_1$, which contradicts the assumption that no adjacent pair is in $\text{tr}(C)$. Conclude that $x_{n-2} = c(\{x_{n-2}, x_1\})$. But then we can remove x_{n-1} without breaking the cycle in $\succ$. We can iterate this argument, removing one item at each step, until we end up with $x_1 = c(\{x_1, x_2\})$, $x_2 = c(\{x_2, x_3\})$, and $x_3 = c(\{x_1, x_3\})$. One of the following must be a cycle: (x_1, x_2, x_3), (x_2, x_3, x_1), (x_3, x_1, x_2). In any case, $x_1\ C\ x_2$ or $x_2\ C\ x_3$, which contradicts the assumption that no adjacent pair is in $\text{tr}(C)$.

Suppose there is some $(x_{i-1}, x_i) \notin \text{tr}(C)$. By the previous step, there is some $(x_{i-1}, x_i) \notin \text{tr}(C)$ such that $(x_i, x_{i+1}) \in \text{tr}(C)$. We can temporarily expand the cycle by adding $(y_1, \ldots, y_k)$ between x_i, x_{i+1}, where $x_i\ C\ y_1\ C\ \ldots\ y_k\ C\ x_{i+1}$. Since $(x_{i-1}, x_i) \notin \text{tr}(C)$, we must have $x_{i-1} = c(\{x_{i-1}, x_i\})$. Since $x_i\ C\ y_1$, we must have $x_i = c(\{x_i, y_1\})$. Now suppose $y_1 = c(\{y_1, x_{i-1}\})$. One of the following must be a cycle: (x_i, y_1, x_{i-1}), (y_1, x_{i-1}, x_i), (x_{i-1}, x_i, y_1). The second and third cases are ruled out because they contradict $(x_{i-1}, x_i) \notin \text{tr}(C)$, so we must have $x_i\ C\ y_1\ C\ x_{i-1}$. But then $(x_i, x_{i-1}) \in \text{tr}(C)$, which contradicts $x_{i-1} \succ x_i$. Conclude that $x_{i-1} = c(\{y_1, x_{i-1}\})$. If $y_1 \succ x_{i-1}$, it must be that $(y_1, x_{i-1}) \in \text{tr}(C)$. Since $x_i\ C\ y_1$, we must have $(x_i, x_{i-1}) \in \text{tr}(C)$, which is a contradiction. Conclude that $x_{i-1} \succ y_1$. This means we can remove x_i while preserving the cycle. Suppose $(x_{i-1}, y_1) \in \text{tr}(C)$. Since $(y_1, x_{i+1}) \in \text{tr}(C)$, we must have $(x_{i-1}, x_{i+1}) \in \text{tr}(C)$. Conclude that $x_1 \succ \ldots \succ x_{i-1} \succ x_{i+1} \succ \ldots x_n$. We have shortened the original cycle by one item. Now suppose $(x_{i-1}, y_1) \notin \text{tr}(C)$. We can repeat the argument above to remove y_1 while preserving the cycle. If $(x_{i-1}, y_2) \in \text{tr}(C)$, we have again shortened the cycle by one item. If $(x_{i-1}, y_2) \notin \text{tr}(C)$, we can repeat the argument once more. We can keep repeating it until we have either shortened the original cycle by one item, or all the ys have been removed. In that case, we will have $x_{i-1} \succ x_{i+1}$—so we will still have shortened the cycle by one item.

We can repeat the procedure above until we have removed all the $(x_{i-1}, x_i) \notin \text{tr}(C)$. Re-indexing the elements, we now have a cycle in which each $(x_{i-1}, x_i) \in \text{tr}(C)$. For each (x_{i-1}, x_i), we can find a finite sequence $(y_1, \ldots, y_k)$ such that $x_i\ C\ y_1\ C\ \ldots\ y_k\ C\ x_{i+1}$. Re-indexing the elements, we now have a cycle in which $x_{i-1}\ C\ x_i$ for each i. This implies both (1) $(x_i, x_{i-1}) \in \text{tr}(C)$ for each i, and (2) $x_{i-1} = c(\{x_{i-1}, x_i\})$. Putting (1) and (2) together, x_i is revealed excluded by x_{i-1} for each i. By IEA, $x_i \notin c(\{x_1, \ldots, x_{n-1}\})$ for all i. This is a contradiction. Conclude that $\succ$ does not contain a cycle. □

We let $\mathcal{M}$ be the set of strict preferences that respect revealed exclusion. That is, $\succ_m$ belongs to $\mathcal{M}$ if and only if

$$a \text{ is revealed excluded by } B \quad \Longrightarrow \quad b \succ_m a \text{ for some } b \in B.$$

It remains to show that $(\succ, \mathcal{M})$ deliver the correct predictions. First, suppose that $b = c(\{b\} \cup A)$.

We show that $b \succ_m A$ for some $\succ_m \in \mathcal{M}$. We construct an appropriate $\succ_m$ as follows. Let

$$B_0 := A$$
$$B_i = B_{i-1} \cup \bigcup_{B' \in \mathcal{F}(B_i)} \{a \in \mathcal{A} : a \text{ is revealed excluded by } B'\} \quad \text{for } i > 0$$
$$B := \bigcup_{i \geq 0}^{\infty} B_i$$
$$T = \mathcal{A} \setminus B$$

For any distinct $x, y \in \mathcal{A}$, impose $x \succ_m y$ if (1) $\{x, y\} \subset B$ or $\{x, y\} \subset T$ and $y \succ x$, or (2) $x \in T$ and $y \in B$. Notice that $\succ_m$ is a strict preference.

We show that $b \in T$, so $b \succ_m A$. Suppose not. Then (using the definition of a tree from the proof of Theorem 1), there is a revealed-exclusion-tree from a subset of A to b. Consider the menu Z that consists of everything in the tree along with any other items in A. Since b is revealed excluded in Z, $b \neq c(Z)$. Since every item in $Z \setminus A$ is revealed excluded in Z, $c(Z) = c(A \cup \{b\})$, so $b \neq c(A \cup \{b\})$—contradiction.

Now we show that $\succ_m$ respects revealed exclusion.

Lemma 13. *If y is revealed excluded by X, then $y \succ X$.*

Proof. Suppose $X = \{x\}$. Then there is a chain from y to x, so $y \succ x$. Now suppose $|X| > 1$. Then $X \cup \{y\}$ is almost-WARP, so $y \neq c(X)$ but $y = c(X' \cup \{y\})$ for any strict subset X' of X. Suppose that there is a chain from x to y for some $x \in X$. Since $y = c(\{x, y\})$, x is revealed excluded by y. Since IEA is necessary, $c(\{y\} \cup X' \setminus \{x\}) = c(\{y\} \cup X)$—contradiction. Conclude that there is no chain from x to y for any $x \in X$. Since $y = c(\{x, y\})$ for each $x \in X$, we have $y \succ X$. □

Suppose y is revealed excluded by X, but $y \succ_m X$. We can write $X = T' \cup B'$ where $B' \subseteq B$ and $T' \subseteq T$. Suppose $y \in B$. If T' is nonempty, we have $t \succ_m y$ for all $t \in T'$. Thus, T' must be empty. It must be that $y \succ_m B'$ but y is revealed excluded by B'. By Lemma 13, we have $y \succ B'$. Since $\{y\} \cup B' \subseteq B$, we have $B' \succ_m y$ by definition of $\mathcal{M}$—contradiction. Now suppose $y \in T$. If T' is empty, then y is revealed excluded by $B' \subseteq B$, which contradicts the assumption that $y \in T$. Thus, $T' \neq \emptyset$. By Lemma 13, we have $y \succ B' \cup T'$. Since $\{y\} \cup T' \subseteq T$, we have $T' \succ_m y$—contradiction. Conclude that $\succ_m$ respects revealed exclusion. There is a preference $\succ_m \in \mathcal{M}$ such that $b \succ_m A$.

Now suppose that $b \succ c(\{b\} \cup A)$. We show that there is no $\succ_m \in \mathcal{M}$ such that $b \succ_m A$.

Lemma 14. *If $x \in X$ is not revealed excluded by any subset of X, and if $x \neq c(X)$, then $c(X) \succ x$.*

Proof. Let

$$X^* := \{x \in X : x \text{ is not revealed excluded by any } X' \subset X\}.$$

By assumption, $x \in X^*$. By IEA, $c(X) = c(X^*)$, so $x \neq c(X^*)$. Take any $X' \subseteq X^*$ such that $|X'| = 3$. If choice on X' violates WARP, then X' is an almost-WARP set or a cycle, so something in X' is revealed excluded. This contradicts the definition of X^*. By induction on the size of X', we can show that choice on X^* satisfies WARP. Since $x \in X^*$, we must have $c(X^*) = c(\{x, c(X^*)\})$. This implies $c(X^*) \succ x$ unless there is a chain from $x \to c(X^*)$. In that case, x is revealed excluded by $c(X^*)$, which contradicts $x \in X^*$. Conclude that $c(X^*) \succ x$. Since $c(X^*) = c(X)$, we have $c(X) \succ x$. □

Suppose that b is not revealed excluded by a subset of A. Lemma 14 implies $c(\{b\} \cup A) \succ b$, which contradicts the assumption that $b \succ c(\{b\} \cup A)$. Conclude that b is revealed excluded by a subset of A, so there is no $\succ_m \in \mathcal{M}$ such that $b \succ_m A$.

A.10 Proof of Corollary 3

The proof of Theorem 4 constructs precisely this representation. Suppose the $\mathcal{M}$ constructed in that proof is not maximal. Then, there is some $\succ_m$ in some representation that does not respect revealed exclusion. That is, $b \succ_m A$ for some menu A and item b such that b is revealed excluded by A. Proposition 3 says that this cannot be the case.

For uniqueness of $\succ$, consider some representation $(\succ', \mathcal{M})$ such that $a \succ b$ but $b \succ' a$. Suppose that there is a chain from a to b. Then, a is revealed preferred to b, which contradicts $b \succ' a$. Now suppose that there is a chain from b to a. This implies $b \succ a$, which contradicts $a \succ b$. Finally, suppose that a and b are not linked by a chain. By definition of $\succ$, we have $a = c(\{a, b\})$. Since $b \succ' a$ and $(\succ', \mathcal{M})$ is a representation, it must be that $a \succ_m b$ for all $\succ_m \in \mathcal{M}$. Consider a preference $\succ_{bad}$ that is exactly opposite $\succ$. We show that $\succ_{bad} \in \mathcal{M}$. Recall from the proof of Proposition 3 that $y \succ X$, so $X \succ_{bad} y$, whenever y is revealed excluded by X. Conclude that $\succ_{bad}$ respects revealed exclusion, so $\succ_{bad} \in \mathcal{M}$. Since $a \succ b$, we have $b \succ_{bad} a$. This contradicts the assumption that $a \succ_m b$ for all $\succ_m \in \mathcal{M}$. Conclude that $(\succ', \mathcal{M})$ is not a representation.

A.11 Proof of Proposition 4

We show sufficiency. Since c_L satisfies IEA, we construct $\succ$ in accordance with Lemma 12. We then let $\mathcal{M}^L$ be the set of strict preferences consistent with revealed exclusion in L. That is, $\succ_m \in \mathcal{M}^L$ if and only if

$$a \text{ is revealed excluded by } B \text{ in } L \quad \Longrightarrow \quad b \succ_m a \text{ for some } b \in B.$$

By Theorem 4, $(\succ, \mathcal{M}^L)$ represents c_L.

For $\mathcal{M}^H$, we need to define a new relation R that captures replacement as well as revealed exclusion.

Definition 40 (Relation R)**.** *Say that $Z\ R\ z$ if either of the following holds:*

1. *z is revealed excluded in L by Z.*
2. *z is replaced in $Z \cup \{z\}$, and no item in Z is revealed excluded in L by any subset of $Z \cup \{z\}$.*

Let $\mathcal{M}^H$ be the set of strict preferences that respect R. That is, $\succ_m \in \mathcal{M}^H$ if and only if

$$B\ R\ a \quad \Longrightarrow \quad b \succ_m a \text{ for some } b \in B.$$

Since R extends revealed exclusion in L, $\mathcal{M}^H \subseteq \mathcal{M}^L$.

Lemma 15. *If a is replaced in $B \cup \{a\}$, then a is replaced in*

$$B^* = \{b \in B : b \textit{ is not revealed excluded in} L \textit{by any} B' \subset B\} \cup \{a\},$$

so $B^*\ R\ a$.

Proof. Suppose $a = c_L(B \cup \{a\}) \neq c_H(B \cup \{a\})$. Since c_L satisfies IEA and c_H satisfies IREA, we have $c_L(B \cup \{a\}) = c_L(B^* \cup \{a\})$ and $c_H(B \cup \{a\}) = c_H(B^* \cup \{a\})$. Thus, $a = c_L(B^* \cup \{a\}) \neq c_H(B^* \cup \{a\})$, so a is replaced in $B^* \cup \{a\}$. □

Lemma 15 implies that each $\succ_m \in \mathcal{M}^H$ respects replacement:

$$a \text{ is replaced in } B \cup \{a\} \quad \Longrightarrow \quad b \succ_m a \text{ for some } b \in B \setminus \{a\}.$$

It remains to show that $(\succ, \mathcal{M}^H)$ delivers the correct predictions. First, suppose that $b = c_H(\{b\} \cup A)$. We show that $b \succ_m A$ for some $\succ_m \in \mathcal{M}_H$. We can construct an appropriate $\succ_m$ following the approach of Theorem 4. The only difference is that we use R instead of revealed exclusion, so

$$B_i = B_{i-1} \cup \bigcup_{B' \in \mathcal{F}(B_i)} \{a \in \mathcal{A} : B'\ R\ a\} \quad \text{for } i > 0.$$

We can then define $\succ_m$ as before. To use the argument that $\succ_m \in \mathcal{M}^H$, we need to show that $Z\ R\ z$ implies $z \succ Z$. From Lemma 13, we know that $z \succ Z$ if z is revealed excluded in L by Z. Suppose instead that z is replaced in $Z \cup \{z\}$, so $z = c_L(Z \cup \{z\}) \neq c_H(Z \cup \{z\})$. Suppose further that no item in Z is revealed excluded in L by any subset of $Z \cup \{z\}$. Since $z = c_L(Z \cup \{z\})$, z is not revealed excluded in L by any subset of Z.

Toward a contradiction, suppose $z' \succ z$ for some $z' \in Z$. There are two possibilities: (1) $z = c_L(\{z', z\})$ and z' comes before z in a chain in L, or (2) $z' = c_L(\{z', z\})$ and z, z' are not linked by a chain in L. In case (1), z' is revealed excluded in L by z, which contradicts our assumption about Z. To rule out case (2), we show that $c_L(\{z, z'\}) = c_L(Z \cup \{z\}) = z$. In the proof of Theorem

4, we showed that the restriction of c to a set in which nothing is revealed excluded satisfies WARP. Since nothing in $Z \cup \{z\}$ is revealed excluded in L, c_L satisfies WARP on $Z \cup \{z\}$. WARP and $z = c_L(Z \cup \{z\})$ imply $z = c_L(\{z, z'\})$. This completes the proof that $Z \; R \; z$ implies $z \succ Z$. We can now use the arguments from Theorem 4, with R in place of revealed exclusion, to show that $\succ_m$ respects R.

Now suppose that $b \succ c_H(\{b\} \cup A)$. We show that there is no $\succ_m \in \mathcal{M}_H$ such that $b \succ_m A$. Toward a contradiction, suppose that $\neg(A' \; R \; b)$ for all $A' \subseteq A$. Let

$$A^* := \{a \in A : \neg(A'' \; R \; a) \text{ for all } A'' \subseteq A\}.$$

Suppose $c_L(A^* \cup \{b\}) \neq c_H(A^* \cup \{b\})$, so $c_L(A^* \cup \{b\})$ is replaced in $A^* \cup \{b\}$. By Lemma 15, there exists $X \subset A^* \cup \{b\}$ such that $X \; R \; c_L(A^* \cup \{b\})$. By definition of A^*, $c_L(A^* \cup \{b\}) \notin A^* \cup \{b\}$—contradiction. Conclude that $c_L(A^* \cup \{b\}) = c_H(A^* \cup \{b\})$. Since c_H satisfies IREA, we have $c_H(A^* \cup \{b\}) = c_H(A \cup \{b\})$. Putting both equalities together, we have $b \succ c_L(A^* \cup \{b\})$. But since b is not revealed excluded in L by any subset of $A^* \cup \{b\}$, Lemma 14 says that $c_L(A^* \cup \{b\})$—contradiction. Conclude that $A' \; R \; b$ for some $A' \subseteq A$, so $\neg(b \succ_m A)$ for all $\succ_m$ that respect R.

A.12 Lemmas used in proofs for RJ

For any $A \in \mathcal{F}(\Delta(Z))$ and any $\mathcal{M} \in \text{supp}(\nu)$, let

$$W_{\mathcal{M}}(A) := \bigcap_{\succsim \in \mathcal{M}} \{x \in \Delta(Z) : \exists a \in A \text{ s.t. } a \succ x\}.$$

For brevity, we typically write $\nu(x \in W(A))$ instead of $\nu(\{\mathcal{M} : x \in W_{\mathcal{M}}(A)\})$.

Lemma 16.

1. *For any $\mathcal{M}, \mathcal{N} \in supp(\nu)$ such that $\mathcal{M} \subset \mathcal{N}$ and any $p \in \Delta(Z)$, $cl(W_{\mathcal{N}}(p)) \setminus \{p\} \subset W_{\mathcal{M}}(p)$.*
2. *For any $p, q \in \Delta(Z)$ such that $\nu(q \notin W(p)) \in (\nu(\mathcal{U}), 1)$, we can find $\mathcal{M} \in supp(\nu)$ such that q is on the boundary of $W_{\mathcal{M}}(p)$.*

Proof.

1. Fix any $q \in \text{cl}(W_{\mathcal{N}})(p) \setminus \{p\}$. We have $n'p \geq n'q$ for all $n \in \mathcal{N}_R$. Thus, any $n \in \mathcal{N}_R$ such that $n'p = n'q$ is on the boundary of $\mathcal{N}_R$. By the first property of ν, we have $\mathcal{M} \subset \text{int}(\mathcal{N})$, so $\mathcal{M}$ cannot contain any boundary point of $\mathcal{N}$. In particular, $m'p > m'q$ for all $m \in \mathcal{M}$. Thus, $q \in W_{\mathcal{M}}(p)$.

2. Let $\mathcal{M}^*$ be the element of $\text{supp}(\nu)$ such that

$$\nu(\{\mathcal{M} : \mathcal{M} \subseteq \mathcal{M}^*\}) = \nu(q \in W(p)).$$

By the second property of ν, $\mathcal{M}^*$ exists and is unique.

We show that q is on the boundary of $W_{\mathcal{M}^*}(p)$. Suppose $q \notin \text{cl}(W_{\mathcal{M}^*}(p))$. Let $\underline{\mathcal{M}}$ be the largest element of $\text{supp}(\nu)$ such that $q \in \text{cl}(W_{\underline{\mathcal{M}}}(p))$. ($\underline{\mathcal{M}}$ must exist because $\text{supp}(\nu)$ is closed.) By the first property of ν, there exists $\tilde{\mathcal{M}} \in \text{supp}(\nu)$ such that $\underline{\mathcal{M}} \subset \text{int}(\tilde{\mathcal{M}}) \subset \tilde{\mathcal{M}} \subset \text{int}(\mathcal{M}^*)$. By definition of $\underline{M}$, we must have $q \notin \text{cl}(W_{\tilde{\mathcal{M}}})$, so

$$\nu(q \in W(p)) < \nu(\{\mathcal{M} : \mathcal{M} \subseteq \tilde{\mathcal{M}}\}) < \nu(\{\mathcal{M} : \mathcal{M} \subseteq \mathcal{M}^*\}),$$

which contradicts the definition of $\mathcal{M}^*$.

Now suppose $q \in W_{\mathcal{M}^*}(p)$. Let $\bar{\mathcal{M}}$ be the smallest element of $\text{supp}(\nu)$ such that $q \notin W_{\bar{\mathcal{M}}}(p)$. (Again, $\bar{\mathcal{M}}$ must exist because $\text{supp}(\nu)$ is closed.) By the first property of ν, there exists $\tilde{\mathcal{M}} \in \text{supp}(\nu)$ such that $\mathcal{M}^* \subset \text{int}(\tilde{\mathcal{M}}) \subset \tilde{\mathcal{M}} \subset \text{int}(\bar{\mathcal{M}})$. Since $q \in W_{\tilde{\mathcal{M}}}(p)$, we have

$$\nu(q \in W(p)) \geq \nu(\{\mathcal{M} : \mathcal{M} \subseteq \tilde{\mathcal{M}}\}) > \nu(\{\mathcal{M} : \mathcal{M} \subseteq \mathcal{M}^*\}),$$

which contradicts the definition of $\mathcal{M}^*$.

□

Lemma 17. *For any $p, q, r \in \Delta(Z)$:*

1. $\nu(p \in W(r)) \geq \min\{\nu(p \in W(q)), \nu(q \in W(r))\}$.
2. *If $\nu(p \in W(q)) \neq \nu(q \in W(r))$ and $\nu(p \in W(r)) > 0$, then*

$$\nu(p \in W(r)) > \min\{\nu(p \in W(q)), \nu(q \in W(r))\}.$$

Proof.

1. Since $\text{supp}(\nu)$ can be ordered by set inclusion,

$$\nu(p \in W(q) \text{ and } q \in W(r)) = \min\{\nu(p \in W(q)), \nu(q \in W(r))\}.$$

$p \in W(q)$ and $q \in W(r)$ implies $p \in W(r)$, so

$$\nu(p \in W(r)) \geq \min\{\nu(p \in W(q)), \nu(q \in W(r))\}.$$

2. The non-trivial case is $\min\{\nu(p \in W(q)), \nu(q \in W(r))\} > 0$. Suppose $\nu(p \in W(q)) > \nu(q \in W(r))$. If $\nu(p \in W(q)) < \max_{x \in \Delta(Z)} \nu(x \in W(q))$, we can use the second part of Lemma 16 to obtain $\mathcal{M}_1$ such that p is on the boundary of $W_{\mathcal{M}_1}(q)$. If $\nu(p \in W(q)) = \max_{x \in \Delta(Z)} \nu(x \in W(q))$, set $\mathcal{M}_1 = \mathcal{U}$. Since the first part of Lemma 16 says $\text{cl}(W_{\mathcal{M}_1}(q)) \setminus \{q\} \subset W_{\mathcal{M}}(q)$ for all $\mathcal{M} \subset \mathcal{M}_1$, we have $p \in W_{\mathcal{M}}(q)$ for all $\mathcal{M} \subset \mathcal{M}_1$. By assumption, $\nu(q \in W(r)) < \max_{x \in \Delta(Z)} \nu(x \in W(r))$, so we can find $\mathcal{M}_2$ such that q is on the boundary of $W_{\mathcal{M}_2}(r)$. We have $r \succsim q$ for all $\succsim \in \mathcal{M}_2$. We also have $\mathcal{M}_2 \subset \mathcal{M}_1$, so $p \in W_{\mathcal{M}_2}(q)$. That is, $q \succ p$ for all $\succsim \in \mathcal{M}_2$. By transitivity, $r \succ p$ for all $\succsim \in \mathcal{M}_2$, so $p \in W_{\mathcal{M}_2}(r)$. We can find $\mathcal{M}_3 \supset \mathcal{M}_2$ such that $p \in W_{\mathcal{M}_3}(r)$, so $\nu(p \in W(r)) > \nu(q \in W(r))$. A parallel argument covers the case $\nu(p \in W(q)) < \nu(q \in W(r))$.

□

Lemma 18.

1. *Suppose that $\nu(\mathcal{U}) < 1$. For any $p \in int(\Delta(Z))$, the sets*

$$\{x \in \Delta(Z) : \nu(p \in W(x)) > 0\} \text{ and } \{x \in \Delta(Z) : \nu(x \in W(p)) > 0\}$$

are disjoint, convex, open and nonempty.

2. *For any $p, q, r \in \Delta(Z)$: if $\nu(p \in W(q)) > 0$ and $\nu(r \in W(p)) > 0$, then $\nu(p \in W(x)) = \nu(x \in W(p)) = 0$ for some $x \in co(\{q, r\})$.*

Proof.

1. Since ρ does not have an REU representation, $\nu(\mathcal{U}) < 1$, so there exist $x, y \in \Delta(Z)$ such that $\nu(y \in W(x)) > 0$. By Independence, $\nu(p + \lambda(y - x) \in W(p)) > 0$ for any $\lambda > 0$ such that $p + \lambda(y - x) \in \Delta(Z)$, and $\nu(p \in W(p + \lambda(x - y))) > 0$ for any $\lambda > 0$ such that $p + \lambda(x - y) \in \Delta(Z)$. Since p is interior, $p + \lambda(y - x)$ and $p + \lambda(x - y)$ will belong to $\Delta(Z)$ for λ small enough.

   Suppose that $\nu(p \in W(q)) > 0$ and $\nu(q \in W(p)) > 0$. If $p \in W(q)$, then $W(p) \subset W(q)$. If in addition $q \in W(p)$, then $q \in W(q)$. Since $\text{supp}(\nu)$ can be ordered by set inclusion,

$$\nu(p \in W(q) \text{ and } q \in W(p)) = \min\{\nu(p \in W(q)), \nu(q \in W(p))\} > 0$$

   so $\nu(q \in W(q)) > 0$. This implies $\rho(q|\{q\}) < 1$—contradiction. Conclude that $\nu(q \in W(p)) = 0$ whenever $\nu(p \in W(q)) > 0$.

   For convexity, suppose that $\nu(q \in W(p)) \geq \nu(r \in W(p)) > 0$. We have $\nu(\{q, r\} \subset W(p)) = \nu(r \in W(p))$. Since each realization of $W(p)$ is convex, $\nu(\text{co}(\{q, r\}) \subset W(p)) = \nu(r \in$

$W(p)) > 0$. A parallel argument establishes that $\nu(q \in B(p)) \geq \nu(r \in B(p)) > 0$ implies $\nu(\text{co}(\{q, r\}) \subset B(p)) > 0$.

For any $\mathcal{M} \in \text{supp}(\mu)$, the sets $W_{\mathcal{M}}(p)$ and $\{x \in \Delta(Z) : p \in W_{\mathcal{M}}(x)\}$ are open because $\mathcal{M}$ is closed. If $\nu(x \in W(p)) > 0$, then $x \in W_{\mathcal{M}}(p)$ for some $\mathcal{M} \neq \underline{\mathcal{M}} := \min(\text{supp}(\nu), \supset)$. That implies $x \in W_{\underline{\mathcal{M}}}(p)$. Conversely, if $x \in W_{\underline{\mathcal{M}}}(p)$, $x \in W_{\mathcal{M}}(p)$ for some $\mathcal{M} \neq \underline{\mathcal{M}}$, so $\nu(x \in W(p)) > 0$. Thus, $\nu(x \in W(p)) > 0$ if and only if $x \in W_{\underline{\mathcal{M}}}(p)$. We have

$$\{x \in \Delta(Z) : \nu(x \in W(p)) > 0\} = W_{\underline{\mathcal{M}}}(p).$$

A similar argument establishes that $\nu(p \in W(x)) > 0$ if and only if $p \in W_{\underline{\mathcal{M}}}(x)$, so

$$\{x \in \Delta(Z) : \nu(p \in W(x)) > 0\} = \{x \in \Delta(Z) : p \in W_{\underline{\mathcal{M}}}(x)\}.$$

2. Since $\{x \in \Delta(Z) : \nu(p \in W(x)) > 0\} \cap \text{co}(\{q, r\})$ and $\{x \in \Delta(Z) : \nu(x \in W(p)) > 0\} \cap \text{co}(\{q, r\})$ are disjoint, nonempty, and open in $\text{co}(\{q, r\})$, the set $\{x \in \Delta(Z) : \nu(p \in W(x)) = \nu(x \in W(p)) = 0\} \cap \text{co}(\{q, r\})$ must be nonempty.

□

Lemma 19. *For any $p \in \Delta(Z)$ and any $a \in \mathcal{F}(\Delta(Z))$,*

$$\nu(p \in W(A)) = \max_{a \in co(A)} \nu(p \in W(a)).$$

Proof. Fix any $\mathcal{M} \in \text{supp}(\nu) \setminus \{\mathcal{U}\}$. We show that $p \in \text{cl}(W_{\mathcal{M}}(A))$ if and only if $p \in \text{cl}(W_{\mathcal{M}}(a))$ for some $a \in \text{co}(A)$. $p \in \text{cl}(W_{\mathcal{M}}(A))$ means that for each $\succsim \in \mathcal{A}$, there exists $a \in A$ such that $a \succsim p$. Let

$$\mathcal{M}_R = \bigcup_{\succsim \in \mathcal{M}} \{(m(a_1), \ldots, m(a_{|A|})) : m \text{ represents } \succsim \text{ and } m(p) = 0\}$$
$$\mathcal{N}_R = \mathbb{R}^{|A|}_{--}$$

Since $\mathcal{M}$ is convex and nonempty, so is $\mathcal{M}_R$. Clearly, $\mathcal{N}_R$ is convex and nonempty, and $\mathcal{N}_R \cap \mathcal{M}_R = \emptyset$. By the Separating Hyperplane Theorem, there exist nonzero $v \in \mathbb{R}^{|A|}$ and c such that $v'm \geq c$ for all $m \in \mathcal{M}_R$ and $v'n \leq c$ for all $n \in \mathcal{N}$. Suppose that $v'n = c$ for some n. Since $\mathcal{N}_R$ is open and v is nonzero, we can find $\tilde{n} \in \mathcal{N}_R$ such that $v'n > c$—contradiction. Thus, $v'n < c$ for all $n \in \mathcal{N}_R$. Suppose that $c < 0$. By choosing $n \in \mathcal{N}_R$ sufficiently close to 0, we get $v'n > c$—contradiction. Thus, $c \geq 0$. Suppose $c > 0$. By choosing $m \in \mathcal{M}_R$ sufficiently close to 0, we get $v'm < c$—contradiction. Thus, $c = 0$. Suppose $v(i) < 0$ for some i. By choosing $n(i)$ sufficiently negative and $n(j)$ sufficiently close to 0 for $j \neq i$, we get $v'n > 0$—contradiction. Thus, $v(i) \geq 0$ for all i, and

(since v is nonzero) $\sum_i v(i) > 0$. Let

$$a^* = \sum_i \frac{v(i)}{\sum_j v(j)} a_i.$$

Since $\sum_j v(j) > 0$, we have

$$\sum_i \frac{v(i)}{\sum_j v(j)} m(i) \geq 0$$

for all $m \in \mathcal{M}_R$. This implies $m(a^*) \geq m(p)$ for every m that represents some $\succsim \in \mathcal{M}$, so $a^* \succsim p$ for all $\succsim \in \mathcal{M}$.

For each $\mathcal{M} \in \text{supp}(\nu)$ such that $p \in \text{cl}(W_{\mathcal{M}}(A))$, let

$$A_{\mathcal{M}} := \{a \in \text{co}(A) : p \in \text{cl}(W_{\mathcal{M}}(a))\}.$$

We have just shown that $A_{\mathcal{M}}$ is nonempty; we now show that it is closed. Take any convergent sequence belonging to $A_{\mathcal{M}}$. Since

$$\text{cl}(W_{\mathcal{M}}(x)) = (x - y) + \text{cl}(W_{\mathcal{M}}(y))$$

for any $x, y \in \Delta(Z)$, we have $\text{cl}(W_{\mathcal{M}}(a_i)) \to \text{cl}(W_{\mathcal{M}}(\lim_i a_i))$. Since $p \in \text{cl}(W_{\mathcal{M}}(a_i))$ for all i, $p \in \text{cl}(W_{\mathcal{M}}(\lim_i a_i))$. Thus, $\lim_i a_i \in A_{\mathcal{M}}$ as desired.

Note that $A_{\mathcal{M}} \supset A_{\mathcal{M}'}$ if $\mathcal{M} \subset \mathcal{M}'$. Since $\text{supp}(\nu)$ can be ordered by set inclusion, so can the $A_{\mathcal{M}}$.

Since each $A_{\mathcal{M}}$ is closed and the $A_{\mathcal{M}}$ can be ordered by set inclusion,

$$\bigcap_{\{\mathcal{M} \in \text{supp}(\nu) : p \in \text{cl}(W_{\mathcal{M}}(A))\}} A_{\mathcal{M}}$$

is nonempty. For any a^* belonging to this set, $p \in \text{cl}(W_{\mathcal{M}}(a^*))$ for all $\mathcal{M} \in \text{supp}(\nu)$ such that $p \in \text{cl}(W_{\mathcal{M}}(A))$. That is,

$$p \in \text{cl}(W_{\mathcal{M}}(a^*)) \quad \iff \quad p \in \text{cl}(W_{\mathcal{M}}(A)).$$

Fix any $\mathcal{M}$ such that $p \in W_{\mathcal{M}}(A)$. Suppose $p \notin W_{\mathcal{M}}(a^*)$. We can find $\mathcal{N} \supset \mathcal{M}$ such that $p \in W_{\mathcal{N}}(A)$. Since $\text{cl}(W_{\mathcal{N}}(a^*)) \setminus \{a^*\} \subset W_{\mathcal{M}}(a^*)$, $p \notin \text{cl}(W_{\mathcal{N}}(a^*))$. This contradicts the definition of a^*. We conclude that

$$p \in W_{\mathcal{M}}(a^*) \quad \iff \quad p \in W_{\mathcal{M}}(A).$$

This implies

$$\nu(p \in W(a^*)) = \nu(p \in W(A)).$$

Since $W(a) \subset W(A)$ for all $a \in \text{co}(A)$, we must have $\nu(p \in W(a)) \leq \nu(p \in W(A))$ for all $a \in \text{co}(A)$. We conclude that

$$\nu(p \in W(a^*)) = \max_{a \in \text{co}(A)} \nu(p \in W(a)).$$

□

Lemma 20. *For any $p, q \in \Delta(Z)$ and any $A \in \mathcal{F}(\Delta(Z))$:*

1. *$\nu(p \in W(A)) = \nu(p \in W(A \setminus \{p\}))$.*
2. *If $\nu(p \in W(A)) \leq \nu(q \in W(A))$, then $\nu(p \in W(A)) = \nu(p \in W(A \setminus \{q\}))$.*

Proof.

1. By Lemma 19, there exists some $a \in \text{co}(A)$ such that $p \in W(A)$ if and only if $p \in W(a)$. Suppose that $a = \lambda p + (1-\lambda)a'$ for some $a' \in \text{co}(A \setminus \{p\})$ and some $\lambda \in (0, 1)$. By Independence, $p \in W(\lambda p + (1-\lambda)a')$ if and only if $p \in W(a')$. Thus, $\nu(p \in W(a')) = \nu(p \in W(A))$, so

   $$\nu(p \in W(A \setminus \{p\})) \geq \nu(p \in W(a')) = \nu(p \in W(A)).$$

   Since $\text{co}(A \setminus \{p\}) \subseteq \text{co}(A)$, it cannot be that $\nu(p \in W(A \setminus \{p\})) > \nu(p \in W(A))$, so we have $\nu(p \in W(A \setminus \{p\})) = \nu(p \in W(A))$ as desired.

2. Suppose that $\nu(p \in W(A)) \leq \nu(q \in W(A))$. Since $\text{supp}(\nu)$ can be ordered by set inclusion, $p \in W(A)$ implies $q \in W(A)$. By the first part of this Lemma, $q \in W(A)$ implies $q \in W(A \setminus \{q\})$. By Independence, $q \in W(A \setminus \{q\})$ implies $\lambda q + (1-\lambda)a \in W(\lambda A \setminus \{q\} + (1-\lambda)a)$ for all $\lambda \in (0, 1]$ and all $a \in \text{co}(A \setminus \{q\})$. Since $\lambda A \setminus \{q\} + (1-\lambda)a \subset A \setminus \{q\}$, $\lambda q + (1-\lambda)a \in W(\lambda A \setminus \{q\})$ implies $\lambda q + (1-\lambda)a \in W(A \setminus \{q\})$.

   Suppose that $p \in W(\lambda q + (1-\lambda)a)$ for some $\lambda \in (0, 1]$ and $a \in \text{co}(A \setminus \{q\})$. Since $\lambda q + (1-\lambda)a \in \text{co}(A)$, $p \in W(A)$. We have just seen that this implies $\lambda q + (1-\lambda)a \in W(A \setminus \{q\})$, which in turn implies $W(\lambda q + (1-\lambda)a) \subset W(A \setminus \{q\})$. Since $p \in W(\lambda q + (1-\lambda)a)$ implies $q \in W(A \setminus \{q\})$, it cannot be that

   $$\nu(p \in W(\lambda q + (1-\lambda)a)) > \nu(p \in W(A \setminus \{q\})).$$

   This implies

   $$\nu(p \in W(A)) = \max_{x \in \text{co}(A)} \nu(p \in W(x)) = \nu(p \in W(A \setminus \{q\})).$$

□

Lemma 21. *For any $p \in \text{int}(\Delta(Z))$:*

1. *The set $D(p) := \arg\max_{x \in \Delta(Z)} \nu(x \in W(p))$ is nonempty.*

2. *For any $d \in D(p)$, $\nu(d \in W(p)) \geq \nu(q \subset W(A))$ for any $q \in \Delta(Z)$ and $A \in \mathcal{F}(\Delta(Z))$.*

3. *For any $q \in \Delta(Z) \setminus D(p)$ such that $\nu(q \in W(p)) > 0$ and any $\lambda \in [0, 1)$:*

$$\begin{aligned} d \in D(p) \quad &\implies \quad \nu(q \in W(\lambda p + (1-\lambda)d)) < \nu(q \in W(p)) \\ q \in D(b) \quad &\implies \quad \nu(\lambda b + (1-\lambda)q \in W(p)) < \nu(q \in W(p)) \end{aligned}$$

Proof.

1. Notice that $D(p) = \bigcap_{\mathcal{M} \in \text{supp}(\nu) \setminus \{\mathcal{U}\}} W_{\mathcal{M}}(p)$. To see why $D(p)$ is nonempty, fix any increasing sequence $\{\mathcal{M}_i \in \text{supp}(\nu) : \mathcal{M}_i \neq \mathcal{U}\}_{i=1}^{\infty}$ such that, for each $t \in (\nu(\mathcal{U}), 1)$,

$$\exists i \in \{1, 2, \ldots\} \text{ s.t. } \nu(\{\mathcal{M} : \mathcal{M} \supset \mathcal{M}_i\}) < t.$$

   Fix ϵ such that $p + \epsilon(q - r) \in \Delta(Z)$ for any $q, r \in \Delta(Z)$. For each i, choose a point w_i on the boundary of $W_{\mathcal{M}_i}(p) \cap (B_\epsilon(p) \setminus B_{\epsilon/2}(p))$. Pass to a convergent subsequence if necessary, and let $w = \lim_i w_i$. Since $w \notin B_{\epsilon/2}(p)$, $w \neq p$. Since the $W_{\mathcal{M}_j}(p)$ are decreasing, we must have $\{w_j, w_{j+1}, \ldots\} \subset \text{cl}(W_{\mathcal{M}_j})$ for all j, so $w \in \text{cl}(W_{\mathcal{M}_j})$.

   We show that $w \in \bigcap_i W_{\mathcal{M}_i}(p)$. Suppose not. Then, there must be some i such that $w \notin W_{\mathcal{M}_i}(p)$. Since $\text{cl}(W_{\mathcal{M}_{i+1}}) \setminus \{p\} \subset W_{\mathcal{M}_i}(p)$, and since $w \neq p$, $w \notin \text{cl}(W_{\mathcal{M}_{i+1}}(p))$—contradiction.

   Finally, we show that $w \in D(p)$. Suppose $w \notin W_{\mathcal{M}^*}$ where $\mathcal{M}^* \in \text{supp}(\nu) \setminus \mathcal{U}$. We can find $\mathcal{M}_i$ such that

$$\nu(\{\mathcal{M} : \mathcal{M} \supset \mathcal{M}^*\}) > \nu(\{\mathcal{M} : \mathcal{M} \supset \mathcal{M}_i\}).$$

   That is, $\mathcal{M}_i \supset \mathcal{M}^*$. Since $w \notin W_{\mathcal{M}}(p)$, $w \notin W_{\mathcal{M}_i}(p)$—contradiction.

2. Since $\nu(q \subset W(A)) = \max_{x \in \text{co}(A)} \nu(q \in W(x))$ by Lemma 19, it suffices to show that $\nu(d \in W(p)) \geq \nu(q \in W(r))$ for any $q, r \in \Delta(Z)$. Suppose $\nu(d \in W(p)) < \nu(q \in W(r))$. By Independence, $\nu(p + \epsilon(q - r) \in W(p)) > \nu(d \in W(p))$ for any ϵ small enough that $p + \epsilon(q - r) \in W(p)$. But then $\nu(d \in W(p)) < \max_{x \in \Delta(Z)} \nu(x \in W(p))$—contradiction.

3. For the first part, suppose $\nu(q \in W(\lambda p + (1-\lambda)d)) = \nu(d \in W(p))$. Since $\nu(d \in W(p)) = \nu(\lambda p + (1-\lambda)d \in W(p))$, we have $\nu(q \in W(p)) \geq \nu(d \in W(p))$ by the first part of Lemma 17. Since $d \in D(p)$, we have $q \in D(p)$, which contradicts the assumption about q. Conclude that $\nu(q \in W(\lambda p + (1-\lambda)d)) < \nu(\lambda p + (1-\lambda)d \in W(p))$. By the second part of Lemma 17, $\nu(q \in W(p)) > \nu(q \in W(\lambda p + (1-\lambda)d))$.

   For the second part, suppose $\nu(\lambda b + (1-\lambda)q \in W(p)) = \nu(q \in W(b))$. Since $\nu(q \in W(b)) = \nu(q \in W(\lambda b + (1-\lambda)q))$, we have $\nu(q \in W(p)) \geq \nu(q \in W(b))$ by the first part of Lemma

17. Since $q \in W(b)$, we have $q \in D(p)$, which contradicts the assumption about q. Conclude that $\nu(\lambda b + (1-\lambda)q \in W(p)) < \nu(q \in W(\lambda b + (1-\lambda)q))$. By the second part of Lemma 17, $\nu(q \in W(p)) > \nu(q \in W(\lambda b + (1-\lambda)d))$.

□

Lemma 22. *For any $p \in \Delta(Z)$ and any $A, B \in \mathcal{F}(\Delta(Z))$: if $B_\epsilon(p) \cap co(A \cup \{p\}) = B_\epsilon(p) \cap co(B \cup \{p\})$ for some $\epsilon > 0$, then $\nu(p \in W(A)) = \nu(p \in W(B))$.*

Proof. By Lemma 19, there exists $a \in \text{co}(A)$ such that $p \in W(A)$ if and only if $p \in W(a)$. For any $\lambda \in [0,1)$, $p \in W(a)$ if and only if $p \in W(\lambda p + (1-\lambda)a)$. Since

$$\nu(p \in W(\lambda p + (1-\lambda)a)) = \nu(p \in W(A)) = \max_{\tilde{a} \in \text{co}(A \cup \{p\})} \nu(p \in W(\tilde{a})),$$

$\lambda p + (1-\lambda)a \in \arg\max_{\tilde{a} \in \text{co}(A \cup \{p\})} \nu(p \in W(\tilde{a}))$. Notice that $\lambda p + (1-\lambda)a \in \bar{B}_\epsilon(p)$ for λ small enough.

Now take any menu B such that $\bar{B}_\epsilon(p) \cap \text{co}(A \cup \{p\}) = \bar{B}_\epsilon(p) \cap \text{co}(B \cup \{p\})$. Clearly, $\lambda p + (1-\lambda)a \in B_\epsilon(p) \cap \text{co}(B \cup \{p\})$. We have

$$\begin{aligned}
\nu(p \in W(B)) &= \max_{\tilde{b} \in \text{co}(B \cup \{p\})} \nu(p \in W(\tilde{b})) \\
&\geq \max_{\tilde{b} \in \bar{B}_\epsilon(p) \cap \text{co}(B \cup \{p\})} \nu(p \in W(\tilde{b})) \\
&= \max_{\tilde{a} \in \bar{B}_\epsilon(p) \cap \text{co}(A \cup \{p\})} \nu(p \in W(\tilde{a})) \\
&= \nu(p \in W(a)) \\
&= \nu(p \in W(A)).
\end{aligned}$$

Now suppose $\nu(p \in W(B)) > \nu(p \in W(A))$, so $\nu(p \in W(b)) > \nu(p \in W(a))$ for some $b \in B$. For any $\lambda \in [0,1)$, $p \in W(b)$ if and only if $p \in W(\lambda p + (1-\lambda)b)$. For λ small enough, $\lambda p + (1-\lambda)b \in B_\epsilon(p) \cap \text{co}(B \cup \{p\})$, so $\lambda p + (1-\lambda)b \in A$. But then

$$\nu(p \in W(a)) < \nu(p \in W(\lambda p + (1-\lambda)b)) \leq \max_{\tilde{a} \in \text{co}(A)} \nu(p \in W(\tilde{a})) = \nu(p \in W(A)),$$

which contradicts the definition of A. Conclude that $\nu(p \in W(B)) = \nu(p \in W(A))$. □

Lemma 23. *For any $q, r \in \Delta(Z)$ and any $A \in \mathcal{F}(\Delta(Z))$: if $\nu(A \subset W(r)) > 0$ and $\nu(r \in W(q)) = 0$, then $\nu(r \in W(A \cup \{q\})) = 0$.*

Proof. Suppose $\nu(r \in W(A \cup \{q\})) > 0$. By Lemma 19, there exists $x \in \text{co}(A \cup \{q\})$ such that

$r \in W(x)$ whenever $r \in W(A \cup \{q\})$. Suppose

$$x = \lambda q + (1-\lambda)a$$

for some $a \in \text{co}(A)$ and $\lambda \in (0,1]$. Since $\nu(r \in W(q)) = 0$, we have

$$\nu(\{\mathcal{M} : \exists \succsim \in \mathcal{M} \text{ s.t. } r \succsim q\}) = 1.$$

Since $\nu(A \subset W(r)) > 0$, we also have

$$\nu(\{\mathcal{M} : \forall \succsim \in \mathcal{M}\ r \succ A\}) > 0.$$

Any preference in $\mathcal{U}$ that has $r \succsim q$ and $r \succ A$ must have $r \succ x$, so

$$\nu(\{\mathcal{M} : \exists \succsim \in \mathcal{M} \text{ s.t. } r \succ x) > 0.$$

This says $\nu(x \in W(r)) > 0$, so $\nu(r \in W(x)) = 0$—contradiction. □

A.13 Proof of Proposition 5

A.13.1 First part

Fix $p, q \in \text{int}(\Delta(Z))$ such that $\nu(p \in W(q)) > 0$ but $p \notin D(q)$. We restrict attention to ϵ small enough that $p + \epsilon(x-y), q + \epsilon(x-y) \in \Delta(Z)$ for all $x, y \in \Delta(Z)$.

By Lemma 21, $D(q) := \arg\max_{x \in \Delta(Z)} \nu(x \in W(q))$ is nonempty. Since

$$d \in D(q) \quad \Longrightarrow \quad \lambda d + (1-\lambda)q \in D(q),$$

the set contains points arbitrarily close to q.

By Independence, $\nu(q \in W(q + \lambda(q-p))) = \nu(p \in W(q)) > 0$ for all $\lambda > 0$. Fix λ such that $q + \lambda(q-p) \in B_{\epsilon/2}(q)$. For any x close enough to q, we have $x + \lambda(x-p) \in B_\epsilon(q)$, $\nu(q \in W(x+\lambda(x-p))) > 0$, and $\nu(p \in W(x)) > 0$. Choose some $x \in D(q)$ that satisfies these requirements, and call it x^*.

Suppose

$$\nu(p \in W(\alpha q + (1-\alpha)(x^* + \lambda(x^* - p)))) \geq \nu(p \in W(q))$$

for some $\alpha \in [0,1)$. By Independence,

$$\nu\left(p \in W\left(\frac{\alpha}{1+\lambda(1-\alpha)}q + \frac{(1-\alpha)(1+\lambda)}{1+\lambda(1-\alpha)}x^*\right)\right) \geq \nu(p \in W(q)).$$

But Lemma 19 and the third part of Lemma 21 imply

$$\nu(p \in W(q)) = \nu(p \in W(\{q, x^*\})) = \max_{x \in \text{co}(\{q, x^*\})} \nu(p \in W(x)),$$

so we have a contradiction. Conclude that

$$\nu(p \in W(q)) = \max_{x \in \text{co}(\{q, x^* + \lambda(x^* - p)\})} \nu(p \in W(x)) = \nu(p \in W(\{q, x^* + \lambda(x^* - p)\})).$$

Suppose $\nu(q \in W(\{p, x^* + \lambda(x^* - p)\})) \geq \nu(p \in W(\{q, x^* + \lambda(x^* - p)\}))$. By Lemma 20,

$$\nu(p \in W(\{q, x^* + \lambda(x^* - p)\})) = \nu(p \in W(x^* + \lambda(x^* - p))).$$

But we have just shown that

$$\nu(p \in W(\{q, x^* + \lambda(x^* - p)\})) = \nu(p \in W(q)) > \nu(p \in W(x^* + \lambda(x^* - p))).$$

Conclude that

$$\nu(q \in W(\{p, x^* + \lambda(x^* - p)\})) < \nu(p \in W(\{q, x^* + \lambda(x^* - p)\})).$$

Notice that $\nu(p \in W(x^* + \lambda(x^* - p))) = \nu(p \in W(x^*)) > 0$, and recall that $\nu(q \in W(x^* + \lambda(x^* - p))) > 0$. By Lemma 23, $\nu(x^* + \lambda(x^* - p) \in W(\{p, q\})) = 0$. We are now ready to compute $\rho(q|\{p, q, x^* + \lambda(x^* - p)\})$. To simplify notation, let

$$\tilde{q} := x^* + \lambda(x^* - p).$$

Two groups of DMs may choose q: those who have $q \succ p, \tilde{q}$, and those who have $p \succ q \succ \tilde{q}$, but $p \in W(q)$. We have

$$\begin{aligned}\rho(q|\{p, q, x^* + \lambda(x^* - p)\}) = \mu(q \succ p, \tilde{q})\nu(q \notin W(\tilde{q})) \\ + \mu(p \succ q \succ \tilde{q}) \left[\nu(q \notin W(\tilde{q})) - \nu(p \notin W(q))\right].\end{aligned}$$

Since $\nu(\{p, q\} \subset W(\tilde{q})) > 0$, we can find $\alpha \in (0, 1)$ such that $\nu(\{p, q\} \subset W(\alpha\tilde{q} + (1 - \alpha)x^*)) > 0$. Since $\nu(p \in W(x^*)) > 0$, and since $\alpha\tilde{q} + (1 - \alpha)x^*$ is a combination of p and x^*, we have $\nu(\alpha\tilde{q} + (1 - \alpha)x^* \in W(\tilde{q})) > 0$. By the second part of Lemma 18, we can find $\beta \in (0, 1)$ such that

$$\nu(\beta\tilde{q} + (1 - \beta)q \in W(\alpha\tilde{q} + (1 - \alpha)x^*)) = \nu(\alpha\tilde{q} + (1 - \alpha)x^* \in W(\beta\tilde{q} + (1 - \beta)q)) = 0.$$

By Lemma 23,

$$\nu(\alpha\tilde{q} + (1-\alpha)x^* \in W(\{p, q, \beta\tilde{q} + (1-\beta)q\})) = 0.$$

Since $\{y \in \Delta(Z) : \nu(p \in W(y)) > 0\}$ is convex by the first part of Lemma 18 and since $\min\{\nu(p \in W(q)), \nu(p \in W(\tilde{q}))\} > 0$, $\nu(p \in W(\beta\tilde{q} + (1-\beta)q)) > 0$. Since $\nu(q \in W(\tilde{q})) > 0$, $\nu(q \in W(\beta\tilde{q} + (1-\beta)q)) > 0$ as well. By Lemma 23,

$$\nu(\beta\tilde{q} + (1-\beta)q \in W(\{p, q, \alpha\tilde{q} + (1-\alpha)x^*\})) = 0.$$

By Lemma 22,

$$\nu(p \in W(\{q, \alpha\tilde{q} + (1-\alpha)x^*, \beta\tilde{q} + (1-\beta)q\})) = \nu(p \in W(\{q, \tilde{q}\})) = \nu(p \in W(q))$$
$$\nu(q \in W(\{p, \alpha\tilde{q} + (1-\alpha)x^*, \beta\tilde{q} + (1-\beta)q\})) = \nu(q \in W(\{p, \tilde{q}\})) = \nu(q \in W(\tilde{q})).$$

We are now ready to compute $\rho(q|\{p, q, \alpha\tilde{q} + (1-\alpha)x^*, \beta\tilde{q} + (1-\beta)q\})$. Two groups of DMs may choose q from this menu: those who have $q \succ p, \tilde{q}$, and those who have $p \succ q \succ \alpha\tilde{q} + (1-\alpha)x^*$, but $p \in W(q)$. We have

$$\rho(q|\alpha\tilde{q} + (1-\alpha)x^*, \beta\tilde{q} + (1-\beta)q\}) = \mu(q \succ p, \tilde{q})\nu(q \notin W(\tilde{q})) + \mu(p \succ q \succ \alpha\tilde{q} + (1-\alpha)x^*)\left[\nu(q \notin W(\tilde{q})) - \nu(p \notin W(q))\right].$$

To show that

$$\rho(q|\alpha\tilde{q} + (1-\alpha)x^*, \beta\tilde{q} + (1-\beta)q\}) < \rho(q|\{p, q, \tilde{q}\}),$$

it suffices to confirm that

$$\mu(p \succ q \succ \alpha\tilde{q} + (1-\alpha)x^*) < \mu(p \succ q \succ \tilde{q}).$$

Since x^* is a combination of p and $\tilde{q}$, $p \succ q \succ \alpha\tilde{q} + (1-\alpha)x^*$ implies $p \succ q \succ \tilde{q}$. The converse does not hold: a DM who likes q slightly more than $\tilde{q}$ and p a lot more than q will have $\tilde{q} + (1-\alpha)x^* \succ q$. Since μ has full support, the inequality holds.

Now we cover the case $p \in D(q)$. The arguments are very similar, but the construction is slightly different. Take any $\tilde{q} \in B_\epsilon(q)$ such that $\nu(q \in W(\tilde{q})) > 0$ but $q \notin D(\tilde{q})$. As above, we have

$$\nu(p \in W(q)) = \nu(p \in W(\{q, \tilde{q}\})) > \nu(q \in W(\tilde{q})) = \nu(q \in W(\{p, \tilde{q}\})) > \nu(\tilde{q} \in W(\{p, q\})) = 0,$$

so

$$\rho(q|\{p, q, \tilde{q}\}) = \mu(q \succ p, \tilde{q})\nu(q \notin W(\tilde{q})) + \mu(p \succ q \succ \tilde{q})[\nu(q \notin W(\tilde{q})) - \nu(p \notin W(q))].$$

As in the first case, we can find $\alpha \in (0,1)$ such that $\nu(\{p,q\} \subset W(\alpha\tilde{q} + (1-\alpha)p)) > 0$ and $\beta \in (0,1)$ such that

$$\nu(\beta\tilde{q} + (1-\beta)q \in W(\alpha\tilde{q} + (1-\alpha)p)) = \nu(\alpha\tilde{q} + (1-\alpha)p \in W(\beta\tilde{q} + (1-\beta)q)) = 0.$$

As in the first case, we have

$$\begin{aligned}
&\nu(\alpha\tilde{q} + (1-\alpha)p \in W(\{p,q,\beta\tilde{q} + (1-\beta)p\})) = 0\\
&\nu(\beta\tilde{q} + (1-\beta)q \in W(\{p,q,\alpha\tilde{q} + (1-\beta)q\})) = 0\\
&\nu(q \in W(\{p,\alpha\tilde{q} + (1-\alpha)p,\beta\tilde{q} + (1-\beta)p\})) = \nu(q \in W(\{p,\tilde{q}\})) = \nu(q \in W(\tilde{q}))\\
&\nu(p \in W(\{q,\alpha\tilde{q} + (1-\alpha)p,\beta\tilde{q} + (1-\beta)p\})) = \nu(p \in W(\{q,\tilde{q}\})) = \nu(p \in W(q)).
\end{aligned}$$

This delivers

$$\begin{aligned}
\rho(q|\alpha\tilde{q} + (1-\alpha)p,\beta\tilde{q} + (1-\beta)q\}) = \mu(q \succ p,\tilde{q})\nu(q \notin W(\tilde{q}))+\\
\mu(p \succ q \succ \alpha\tilde{q} + (1-\alpha)p)\left[\nu(q \notin W(\tilde{q})) - \nu(p \notin W(q))\right].
\end{aligned}$$

Since $\mu(p \succ q \succ \alpha\tilde{q} + (1-\alpha)p) < \mu(p \succ q \succ \tilde{q})$, there is a regularity violation as above.

A.13.2 Second part

Fix any $p,q \in \text{int}(\Delta(Z))$ such that $\nu(p \in W(q)) = 0$.

Suppose there exists $\bar{\epsilon} > 0$ such that $\nu(p \in W(\tilde{q})) = 0$ for all $\tilde{q} \in B_{\bar{\epsilon}}(q)$. We show that (p,q) cannot be anomalous for ϵ sufficiently small.

For any plane $P \in \Delta(Z)$ and any $x \in P$, let

$$D_P(x) := \max_{y \in P} \nu(y \in W(x)).$$

If we restrict attention to items in P, Lemma 21 will still apply with D_P in place of D.

Fix any plane P containing p and q. To begin, suppose $q \in D_P(p)$. By the second part of Lemma 21,

$$\nu(q \in W(\{p,\tilde{q}\})) = \nu(q \in W(p)) \geq \max\{\nu(\tilde{q} \in W(\{p,q\})), \nu(p \in W(\{q,\tilde{q}\}))\}.$$

Thus,

$$\rho(q|\{p,q,\tilde{q}\}) = \mu(q \succ p,\tilde{q})\nu(q \notin W(p)).$$

Similarly, for any $\beta_1, \beta_2, \delta_1, \delta_2 > 0$ such that $\beta_1 + \beta_2 = \delta_1 + \delta_2 = 0$, we have

$$\begin{aligned}\nu(q \in W(p)) \geq \max\{&\nu(p \in W(\{\beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p, q\})),\\ &\nu(\beta_1\tilde{q} + \beta_2 q \in W(\{\delta_1\tilde{q} + \delta_2 p, p, q\})),\\ &\nu(\delta_1\tilde{q} + \delta_2 p \in W(\{\{\beta_1\tilde{q} + \beta_2 q, p, q\}}))\},\end{aligned}$$

so

$$\begin{aligned}\rho(q|\{p, q, \beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\}) &= \mu(q \succ p, \beta_1\tilde{q} + \beta_2 q)\nu(q \notin W(p))\\ &= \mu(q \succ p, \tilde{q})\nu(q \notin W(p))\\ &= \rho(q|\{p, q, \tilde{q}\}).\end{aligned}$$

From now on, we assume $q \notin D_P(p)$. For ϵ small enough, any $\tilde{q} \in B_\epsilon(q) \cap P$ can be written in one of two ways. The first is

$$\tilde{q} = \alpha_1 q + \alpha_2 b + (1 - \alpha_1 - \alpha_2)p$$

where $q \in D_P(b)$ and $\alpha_1, \alpha_2 \geq 0$ (but $\alpha_1 + \alpha_2$ may exceed 1).

For ϵ small enough, b can be taken to be sufficiently close to q that

$$\nu(p \in W(x)) = 0$$

for all $x \in \text{co}(\{b, q\})$. By Independence, the same is true for all $x \in \text{co}(\{b, q, p\})$. Since $\text{co}(q, \tilde{q}) \subset \text{co}(\{b, q, p\})$, we have

$$\nu(p \in W(\{q, \tilde{q}\})) = 0$$

by Lemma 19.

We show that $\nu(\tilde{q} \in W(\{p, q\})) \geq \nu(q \in W(\{p, \tilde{q}\}))$.

$$\begin{aligned}\tilde{q} \in W(\{p, q\}) &= \max_{\lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1} \nu(\alpha_1 q + \alpha_2 b + \alpha_3 p \in W(\lambda_1 q + \lambda_2 p))\\ &= \max\left\{\max_{\lambda_1 \geq \alpha_1, \lambda_2 \geq \alpha_3} \nu\left(b \in W\left(\frac{\lambda_1 - \alpha_1}{\alpha_2} q + \frac{\lambda_2 - \alpha_3}{\alpha_2} p\right)\right),\right.\\ &\quad \max_{\lambda_1 > \alpha_1, \lambda_2 \leq \alpha_3} \nu\left(\frac{\alpha_2}{\lambda_1 - \alpha_1} b + \frac{\alpha_3 - \lambda_2}{\lambda_1 - \alpha_1} p \in W(q)\right)\\ &\quad \left.\max_{\lambda_1 \leq \alpha_1, \lambda_2 > \alpha_3} \nu\left(\frac{\alpha_1 - \lambda_1}{\lambda_2 - \alpha_3} q + \frac{\alpha_2}{\lambda_2 - \alpha_3} b \in W(p)\right)\right\}\end{aligned}$$

We work through the terms in the max. First, suppose that $\nu(b \in W(x)) > \nu(q \in W(p))$ for some $x \in \text{co}(\{p, q\})$. Since b was chosen so that $\nu(q \in W(b)) \geq \nu(b \in W(x))$, the first part of Lemma 17 implies $\nu(q \in W(x)) \geq \nu(b \in W(x)) > \nu(q \in W(p))$. But since $x \in \text{co}(\{p, q\})$, Independence implies

$\nu(q \in W(x)) = \nu(q \in W(p))$–contradiction. Conclude that $\nu(b \in W(x)) \leq \nu(q \in W(p))$ for all $x \in \text{co}(\{p, q\})$, so the first term cannot exceed $\nu(q \in W(p))$. Now consider the second term. Since $\nu(p \in W(q)) = 0$ and $\nu(q \in W(b)) > 0$, Lemma 23 gives $\nu(x \in W(q)) = 0$ for all $x \in \text{co}(\{b, p\})$. Thus, the second term is 0. Finally, consider the third term. Suppose that $\nu(x \in W(p)) > \nu(q \in W(p))$ for some $x \in \text{co}(\{b, q\})$. Independence implies $\nu(q \in W(x)) = \nu(q \in W(b))$, and b was chosen so that $\nu(q \in W(b)) \geq \nu(x \in W(p))$. We have $\nu(q \in W(x)) \geq \nu(x \in W(p))$. By the first part of Lemma 17, $\nu(q \in W(p)) \geq \nu(x \in W(p)) > \nu(q \in W(p))$—contradiction. Conclude that $\nu(x \in W(p)) \leq \nu(q \in W(p))$ for all $x \in \text{co}(\{b, q\})$. Thus, the third term cannot exceed $\nu(q \in W(p))$. Putting all three terms together, we have

$$\nu(\tilde{q} \in W(\{p, q\})) \leq \nu(q \in W(p)) \leq \nu(q \in W(\{p, \tilde{q}\})).$$

Since

$$\nu(q \in W(\{p, q\})) \geq \max\{\nu(\tilde{q} \in W(\{p, q\})), \nu(p \in W(\{q, \tilde{q}\}))\},$$

we have

$$\rho(q|\{p, q, \tilde{q}\}) = \mu(q \succ p, \tilde{q})\nu(q \notin W(\{p, \tilde{q}\})).$$

Now we show that

$$\nu(\beta_1\tilde{q} + \beta_2 q \in W(\{p, q, \delta_1\tilde{q} + \delta_2 p\})) \leq \nu(q \in W(\{p, \beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\})) \tag{14}$$

for any $\beta_1, \beta_2, \lambda_1, \lambda_2 > 0$ and $\beta_1 + \beta_2 = \lambda_1 + \lambda_2 = 1$. By Lemmas 19 and 22,

$$\begin{aligned}
\nu(\beta_1\tilde{q} + \beta_2 q \in W(\{p, q, \delta_1\tilde{q} + \delta_2 p\})) &= \nu(\beta_1\tilde{q} + \beta_2 q \in W(\{q, \delta_1\tilde{q} + \delta_2 p\})) \\
&= \max_{\lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1} \nu(\beta_1\tilde{q} + \beta_2 q \in W(\lambda_1 q + \lambda_2(\delta_1\tilde{q} + \delta_2 p))) \\
&= \max\left\{ \max_{\lambda_1 \leq \beta_2, \lambda_2 \leq \beta_1/\delta_1} \nu\left(\frac{\beta_1 - \lambda_2\delta_1}{\delta_2\lambda_2}\tilde{q} + \frac{\beta_2 - \lambda_1}{\delta_2\lambda_2} q \in W(p)\right), \right. \\
&\quad \max_{\lambda_1 < \beta_2, \lambda_2 > \beta_1/\delta_1} \nu\left(q \in W\left(\frac{\lambda_2\delta_1 - \beta_1}{\beta_2 - \lambda_1}\tilde{q} + \frac{\lambda_2\delta_2}{\beta_2 - \lambda_1} p\right)\right), \\
&\quad \left. \max_{\lambda_1 > \beta_2, \lambda_2 < \beta_1/\delta_1} \nu\left(\tilde{q} \in W\left(\frac{\lambda_1 - \beta_2}{\beta_1 - \lambda_2\delta_1} q + \frac{\lambda_2\delta_2}{\beta_1 - \lambda_2\delta_1} p\right)\right)\right\}
\end{aligned}$$

We work through the terms in the max. Expanding $\tilde{q}$ and rearranging, and using the fact (from the third part of Lemma 21) that

$$\nu(q \in W(p)) = \max_{\lambda \in [0,1]} \nu(\lambda q + (1 - \lambda)b \in W(p)).$$

we can rewrite the first term as $\nu(q \in W(p))$. By Lemma 19, the second term is no greater than

$\nu(q \in W(\{p, \tilde{q}\}))$, which is equal to

$$\nu(q \in W(\{p, \beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\}))$$

by Lemma 22. Similarly, the third term is no greater than $\nu(\tilde{q} \in W(\{p, q\}))$, which we already showed was less than $\nu(q \in W(\{p, \tilde{q}))$. Combining all three terms, we have (14) as desired.

By Lemma 22,

$$\nu(p \in W(\{q, \beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\})) = \nu(p \in W(\{q, \tilde{q}\})) = 0.$$

Since

$$\begin{aligned}\nu(q \in W(\{p, \beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\})) \geq \max\{&\nu(p \in W(\{q, \beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\})), \\ &\nu(\beta_1\tilde{q} + \beta_2 q \in W(\{p, q, \delta_1\tilde{q} + \delta_2 p\}))\},\end{aligned}$$

the only DMs who would choose q from $\{p, q, \beta_1, \tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\}$ are the ones who like q best. We have

$$\begin{aligned}\rho(q|\{p, q, \beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\}) &= \mu(q \succ \tilde{q}, p)\nu(q \notin W(\{p, \tilde{q}\}) \\ &= \rho(q|\{p, q, \tilde{q}\}).\end{aligned}$$

This completes the first possibility for $\tilde{q}$. The second possibility is

$$\tilde{q} = \alpha_1 q + \alpha_2 w + (1 - \alpha_1 - \alpha_2)p$$

where $w \in D_P(q)$ and $\alpha_1, \alpha_2 \geq 0$ (but $\alpha_1 + \alpha_2$ may exceed 1). We show that $\nu(p \in W(x)) = 0$ for all $x \in \text{co}(\{q, \tilde{q}\})$. Suppose not. By Independence, $\nu(p \in W(y)) > 0$ for some $y \in \text{co}(\{q, w\}) \setminus \{q\}$. Since $\nu(y \in W(q)) > 0$, $\nu(p \in W(q)) > 0$ by the first part of Lemma 17—contradiction.

For $\tilde{q}$ sufficiently close to q, there are two subcases. The first is $\tilde{q} \in D_P(y)$ for some $y \in \text{co}(\{p, q\})$. By Independence, this implies

$$\begin{aligned}&\exists y \in \text{co}(\{p, q\}) \quad \beta_1\tilde{q} + \beta_2 q \in D_P(y) \\ &\exists y \in \text{co}(\{p, q\}) \quad \delta_1\tilde{q} + \delta_2 p \in D_P(y)\end{aligned}$$

for all $\beta_1, \beta_2, \delta_1, \delta_2 > 0$ such that $\beta_1 + \beta_2 = \delta_1 + \delta_2 = 1$. We have

$$\begin{aligned}\nu(\tilde{q} \in W(\{p, q\})) &= \nu(\beta_1\tilde{q} + \beta_2 q \in W(\{p, q, \delta_1\tilde{q} + \delta_2 p\})) \\ &= \nu(\delta_1\tilde{q} + \delta_2 q \in W(\{p, q, \delta_1\tilde{q} + \delta_2 p\})).\end{aligned}$$

We also have

$$\nu(\tilde{q} \in W(\{p,q\})) \geq \nu(q \in W(\{p,\tilde{q}\})),$$

so

$$\nu(q \in W(\{p,\tilde{q}\})) = \nu(q \in W(p))$$

by Lemma 20. There are two groups of DMs that may choose q from $\{p,q,\tilde{q}\}$: those who like q best, and those who have $\tilde{q} \succ q \succ p$, but $\tilde{q} \in W(\{p,q\})$. We have

$$\rho(q|\{p,q,\tilde{q}\}) = \mu(q \succ p,\tilde{q})\nu(q \notin W(p)) + \mu(\tilde{q} \succ q \succ p)[\nu(q \notin W(p)) - \nu(\tilde{q} \notin W(\{p,q\}))].$$

Similarly,

$$\nu(\{\beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\} \subset W(\{p,q\})) \geq \nu(q \in W(\{p,\beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\})),$$

so Lemma 20 implies

$$\nu(q \in W(\{p,\beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\})) = \nu(q \in W(p)).$$

We have

$$\begin{aligned}
\rho(q|\{p,q,\beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\}) &= \mu(q \succ p,\tilde{q})\nu(q \notin W(p)) \\
&\quad + \mu(\beta_1\tilde{q} + \beta_2 q \succ q \succ p) \\
&\quad \times [\nu(q \notin W(p)) - \nu(\beta_1\tilde{q} + \beta_2 q \notin W(\{p,q\}))] \\
&= \mu(q \succ p,\tilde{q})\nu(q \notin W(p)) \\
&\quad + \mu(\tilde{q} \succ q \succ p)[\nu(q \notin W(p)) - \nu(\tilde{q} \notin W(\{p,q\}))] \\
&= \rho(q|\{p,q,\tilde{q}\}).
\end{aligned}$$

The second subcase is

$$q \in \underset{x \in \text{co}(\{p,q\})}{\arg\max}\ \nu(\tilde{q} \in W(x)).$$

We show that

$$\nu(q \in W(\{\tilde{q},p\}) = \nu(q \in W(p)).$$

By Lemma 19,

$$\begin{aligned}\nu(q \in W(\{p, \tilde{q}\})) &= \max_{\lambda_1,\lambda_2 \geq 0, \lambda_1+\lambda_2=1} \nu(q \in W(\lambda_1 \tilde{q} + \lambda_2 p)) \\ &= \max \left\{ \max_{\lambda_1\alpha_1 < 1, \lambda_1\alpha_3+\lambda_2 \geq 0} \nu\left(\frac{\lambda_1\alpha_2}{1-\lambda_1\alpha_1} w + \frac{\lambda_1\alpha_3+\lambda_2}{1-\lambda_1\alpha_1} p\right), \right. \\ &\qquad \left. \max_{\lambda_1\alpha_1 \geq 1, \lambda_1\alpha_3+\lambda_2 < 0} \nu\left(p \in W\left(\frac{\lambda_1\alpha_1 - 1}{-(\lambda_1\alpha_3+\lambda_2)} q + \frac{\lambda_1\alpha_2}{-(\lambda_1\alpha_3+\lambda_2)} w\right)\right) \right\}\end{aligned}$$

For the first term, recall that $\nu(w \in W(\{p, q\})) = \nu(w \in W(q)) \geq \nu(q \in W(\{w, p\}))$. By Lemma 20, $\nu(q \in W(\{w, p\})) = \nu(q \in W(p))$. By the third part of Lemma 21,

$$\nu(q \in W(p)) = \max_{x \in \text{co}(\{w,p\})} \nu(q \in W(x))$$

so the first term is $\nu(q \in W(p))$. The second term is 0 because $\nu(p \in W(x)) = 0$ for all $x \in \text{co}(\{q, \tilde{q}\})$.

Now we show that $\nu(q \in W(p)) \leq \nu(\tilde{q} \in W(\{p, q\}))$. Suppose not. By Lemma 20, $\nu(\tilde{q} \in W(\{p, q\})) = \nu(\tilde{q} \in W(p))$. Expanding $\tilde{q}$, we have

$$\nu(\tilde{q} \in W(p)) = \nu\left(\frac{\alpha_1}{\alpha_1+\alpha_2} q + \frac{\alpha_2}{\alpha_1+\alpha_2} w \in W(p)\right).$$

Since $\text{co}(\{q, w\}) \subset W(p)$ whenever $q \in W(p)$, we have

$$\begin{aligned}\nu(q \in W(p)) &\leq \nu\left(\frac{\alpha_1}{\alpha_1+\alpha_2} q + \frac{\alpha_2}{\alpha_1+\alpha_2} w \in W(p)\right) \\ &= \nu(\tilde{q} \in W(p)) \\ &= \nu(\tilde{q} \in W(\{p, q\})).\end{aligned}$$

This contradicts the assumption that $\nu(q \in W(p)) > \nu(\tilde{q} \in W(\{p, q\}))$.

We can now compute $\rho(q|\{p, q, \tilde{q}\})$. We have

$$\rho(q|\{p, q, \tilde{q}\}) = \mu(q \succ p, \tilde{q})\nu(q \notin W(p)) + \mu(\tilde{q} \succ q \succ p)[\nu(q \notin W(p)) - \nu(\tilde{q} \notin W(\{p, q\}))]$$

exactly as in the first subcase.

We show that

$$\nu(\beta_1 \tilde{q} + \beta_2 q \in W(\{p, q, \delta_1 \tilde{q} + \delta_2 p\})) = \nu(\tilde{q} \in W(q)). \tag{15}$$

By Lemma 22,

$$\nu(\beta_1 \tilde{q} + \beta_2 q \in W(\{p, q, \delta_1 \tilde{q} + \delta_2 p\}))) = \nu(\beta_1 \tilde{q} + \beta_2 q \in W(\{q, \delta_1 \tilde{q} + \delta_2 p\}))).$$

By Lemma 19,

$$\begin{aligned}\nu(\beta_1\tilde{q}+\beta_2 q \in W(\{q,\delta_1\tilde{q}+\delta_2 p\}))) &= \max_{\lambda_1,\lambda_2\geq 0,\lambda_1+\lambda_2=1} \nu(\beta_1\tilde{q}+\beta_2 q \in W(\lambda_1(\delta_1\tilde{q}+\delta_2 p)+\lambda_2 q)) \\ &= \max\left\{ \max_{\beta_1\geq\lambda_1\delta_1,\beta_2\geq\lambda_2} \nu\left(\frac{\beta_1-\lambda_1\delta_1}{\lambda_1\delta_2}\tilde{q}+\frac{\beta_2-\lambda_2}{\lambda_1\delta_2}q \in W(p)\right),\right. \\ &\quad \max_{\beta_1<\lambda_1\delta_1,\beta_2>\lambda_2} \nu\left(q\in W\left(\frac{\lambda_1\delta_1-\beta_1}{\beta_2-\lambda_2}\tilde{q}+\frac{\lambda_1\delta_2}{\beta_2-\lambda_2}p\right)\right), \\ &\quad \left. \max_{\beta_1>\lambda_1\delta_1,\beta_2<\lambda_2} \nu\left(\tilde{q}\in W\left(\frac{\lambda_2-\beta_2}{\beta_1-\lambda_1\delta_1}q+\frac{\lambda_1\delta_2}{\beta_1-\lambda_1\delta_1}p\right)\right)\right\}\end{aligned}$$

We work through the terms in the max. For the first term, suppose that $\nu(x \in W(p)) > \nu(\tilde{q} \in W(q))$ for some $x \in \text{co}(\{\tilde{q},q\}) \setminus \{\tilde{q}\}$. Since $\{y \in \Delta(Z) : y \in W(p)\}$ is a cone, it cannot be that $\nu(x \in W(p)) > 0$ and $\nu(\tilde{q} \in W(p)) = 0$ for ϵ small. Thus, $\nu(\tilde{q} \in W(p)) > 0$. By Independence, $\nu(\tilde{q} \in W(q)) = \nu(\tilde{q} \in W(x))$, so $\nu(x \in W(p)) > \nu(\tilde{q} \in W(x))$. By the second part of Lemma 17, $\nu(\tilde{q} \in W(p)) > \nu(\tilde{q} \in W(x)) = \nu(\tilde{q} \in W(q))$. This contradicts the assumption that $\nu(\tilde{q} \in W(q)) = \nu(\tilde{q}inW(\{p,q\}))$. Conclude that the first term cannot exceed $\nu(\tilde{q} \in W(q))$. Now consider the second term. By Lemma 19, $\nu(q \in W(x)) \leq \nu(q \in W(\{p,\tilde{q}\}))$ for all $x \in \text{co}(\{p,\tilde{q}\})$. We already showed that $\nu(q \in W(\{p,\tilde{q}\})) = \nu(q \in W(p)) \leq \nu(\tilde{q} \in W(q))$. Finally, consider the third term. By Lemma 19, $\nu(\tilde{q} \in W(x)) \leq \nu(\tilde{q} \in W(\{p,q\}))$ for all $x \in \text{co}(\{p,q\})$. Since $\nu(\tilde{q} \in W(\{p,q\})) = \nu(\tilde{q} \in W(q))$, the third term cannot exceed $\nu(\tilde{q} \in W(q))$. This maximum can be achieved by setting $\lambda_1 = 0$, so the third term equals $\nu(\tilde{q} \in W(q))$. Conclude that (15) holds.

By Lemma 19,

$$\nu(\delta_1\tilde{q}+\delta_2 p \in W(\{p,q,\beta_1\tilde{q}+\beta_2 q\})) \geq \max_{x\in\text{co}(\{p,q\})} \nu(\delta_1\tilde{q}+\delta_2 p \in W(x)).$$

By Independence,

$$\nu(\delta_1\tilde{q}+\delta_2 p \in W(\delta_1 q+\delta_2 p)) = \nu(\tilde{q} \in W(q)).$$

Conclude that

$$\nu(\delta_1\tilde{q}+\delta_2 p \in W(\{p,q,\beta_1\tilde{q}+\beta_2 q\})) \geq \nu(\tilde{q} \in W(q)).$$

As in the first subcase, we have

$$\begin{aligned}\nu(q \in W(\{p,\beta_1\tilde{q}+\beta_2 q,\delta_1\tilde{q}+\delta_2 p\})) &= \nu(q \in W(p,\tilde{q})) = \nu(q \in W(p)) \\ \nu(p \in W(\{p,\beta_1\tilde{q}+\beta_2 q,\delta_1\tilde{q}+\delta_2 p\})) &= \nu(p \in W(q,\tilde{q})) = 0.\end{aligned}$$

There are two groups of DMs who may choose q from $\{p,q,\beta_1\tilde{q}+\beta_2 q,\delta_1\tilde{q}+\delta_2 p\}$: those who like q

best, and those who have $\beta_1\tilde{q} + \beta_2 q \succ q \succ p$ but $\beta_1\tilde{q} + \beta_2 q \in W(\{p, q, \delta_1\tilde{q} + \delta_2 p\})$. We have

$$\begin{aligned}\rho(q|\{p, q, \beta_1\tilde{q} + \beta_2 q, \delta_1\tilde{q} + \delta_2 p\}) &= \mu(q \succ \beta_1\tilde{q} + \beta_2 q, p)\nu(q \notin W(p)) \\ &\quad + \mu(\beta_1\tilde{q} + \beta_2 q \succ q \succ p) \\ &\quad \times [\nu(q \notin W(p)) - \nu(\beta_1\tilde{q} + \beta_2 q \notin W(\{p, q\}))] \\ &= \mu(q \succ \tilde{q}, p)\nu(q \notin W(p)) \\ &\quad + \mu(\tilde{q} \succ q \succ p)[\nu(q \notin W(p)) - \nu(\tilde{q} \notin W(p))] \\ &= \rho(q|\{p, q, \tilde{q}\}).\end{aligned}$$

This completes the second subcase and, by extension, the second possibility for $\tilde{q}$. We have shown that (p, q) cannot be anomalous for ϵ sufficiently small, provided there exists $\bar{\epsilon} > 0$ such that $\nu(p \in W(\tilde{q})) = 0$ for all $\tilde{q} \in B_{\bar{\epsilon}}(q)$.

Now suppose there exist $\tilde{q}$ arbitrarily close to q such that $\nu(p \in W(\tilde{q})) > 0$. Since $\nu(p \in W(q)) = 0$, q must be on the boundary of $\{x \in \Delta(Z) : \nu(p \in W(x)) > 0\}$. There are q' arbitrarily close to q that are not on this boundary. For any such q', there exists $\epsilon > 0$ such that $\nu(p \in W(\tilde{q})) = 0$ for all $\tilde{q} \in B_\epsilon(q')$. We have already seen that (p, q') cannot be anomalous. Thus, there is no $\epsilon > 0$ such that (p, q') is anomalous for all $q' \in B_\epsilon(q)$.

A.14 Proof of Theorem 5

If ρ has an REU representation, the unique RJ representation has $\nu(\mathcal{U}) = 1$. μ is simply the REU representation. For the remainder of this proof, we assume that ρ has an REU representation with $\nu(\mathcal{U}) < 1$.

A.14.1 Identifying $D(p)$

Fix any interior $q, \tilde{q}$. Suppose that $q \in D(\tilde{q})$. For any $p \in \Delta(Z)$, we have

$$\nu(q \in W(\tilde{q})) \geq \max\{\nu(p \in W(\{q, \tilde{q}\})), \nu(\tilde{q} \in W(\{p, q\}))\}$$

by the second part of Lemma 21. This implies

$$\rho(q|\{p, q, \tilde{q}\}) = \mu(q \succ p, \tilde{q})\nu(q \notin W(\tilde{q})).$$

By Independence, $q \in D(\beta\tilde{q} + (1-\beta)q)$ for any $\beta \in (0,1)$, so

$$\begin{aligned}\nu(q \in W(\beta\tilde{q} + (1-\beta)q)) \geq \max\{&\nu(p \in W(\{q, \beta\tilde{q} + (1-\beta)q, \delta\tilde{q} + (1-\delta)p\})),\\ &\nu(\beta\tilde{q} + (1-\beta)q \in W(\{p, q, \delta\tilde{q} + (1-\delta)p\})),\\ &\nu(\delta\tilde{q} + (1-\delta)p \in W(\{p, q, \beta\tilde{q} + (1-\beta)q\}))\}\end{aligned}$$

for any $\delta \in (0,1)$. This implies

$$\begin{aligned}\rho(q|\{p, q, \beta\tilde{q} + (1-\beta)q, \delta\tilde{q} + (1-\delta)q\}) &= \mu(q \succ p, \beta\tilde{q} + (1-\beta)q)\nu(q \notin W(\beta\tilde{q} + (1-\beta)q))\\ &= \mu(q \succ p, \tilde{q})\nu(q \notin W(\tilde{q}))\\ &= \rho(q|\{p, q, \tilde{q}\})\end{aligned}$$

Now suppose that $\nu(q \in W(\tilde{q})) > 0$, but $q \notin D(\tilde{q})$. (We have already seen how to tell whether $\nu(q \in W(\tilde{q})) > 0$.) We showed at the end of Section A.13.1 that

$$\rho(q|\{p, q, \tilde{q}\}) > \rho(q|\{p, q, \beta\tilde{q} + (1-\beta)q, \delta\tilde{q} + (1-\delta)p\})$$

for some $p \in \Delta(Z)$ (in particular, any $p \in D(q)$ will work) and some $\beta, \delta \in (0,1)$. This allows us to identify $D(\tilde{q})$ and, by extension, $D(x)$ for any $x \in \Delta(Z)$.

A.14.2 Identifying ν

Take any $p, q \in \text{int}(\Delta(Z))$ such that $\nu(p \in W(q)) > 0$ and $p \notin D(q)$. Construct x^* and $\tilde{q}$ exactly as in Section A.13.1.

Since

$$\nu(p \in W(\{q, \tilde{q}\})) = \nu(p \in W(q)) > \nu(q \in W(\{p, \tilde{q}\})) = \nu(q \in W(\tilde{q})) > \nu(\tilde{q} \in W(\{p, q\})) = 0,$$

we have

$$\begin{aligned}\rho(p|\{p, q, \tilde{q}\}) &= \mu(p \succ q, \tilde{q})\nu(p \notin W(q))\\ &= \left[\mu\left(\frac{1}{2}q + \frac{1}{2}p \succ q, \tilde{q}\right) + \mu\left(p \succ \tilde{q} \succ \frac{1}{2}q + \frac{1}{2}p\right)\right]\nu(p \notin W(q))\end{aligned}$$

Now consider menu

$$\left\{\frac{1}{2}q + \frac{1}{2}p, q, \tilde{q}\right\}.$$

By Lemma 22,

$$\nu\left(q \in W\left(\left\{\frac{1}{2}q + \frac{1}{2}p, q, \tilde{q}\right\}\right)\right) = \nu(q \in W(\{p, \tilde{q}\})) = \nu(q \in W(\tilde{q})).$$

Since $\nu(\{p,q\} \in W(\tilde{q})) > 0$, $\nu(\tilde{q} \in W(\{p,q\})) = 0$. By Lemma 19, $\nu(\tilde{q} \in W(A)) = 0$ for all $A \subset \text{co}(\{p,q\})$. In particular,

$$\nu\left(\tilde{q} \in W\left(\left\{\frac{1}{2}q + \frac{1}{2}p, q\right\}\right)\right) = 0.$$

By Lemma 19,

$$\begin{aligned}
\nu\left(\frac{1}{2}q + \frac{1}{2}p \in W(q, \tilde{q})\right) &= \max_{\lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1} \nu\left(\frac{1}{2}q + \frac{1}{2}p \in W(\lambda_1 q + \lambda_2 \tilde{q})\right) \\
&= \max\left\{\max_{\lambda_1 \geq \frac{1}{2}} \nu\left(p \in W((2\lambda_1 - 1)q + 2\lambda_2 \tilde{q})\right),\right. \\
&\qquad \left.\max_{\lambda_1 < \frac{1}{2}} \nu\left(\frac{1 - 2\lambda_1}{2\lambda_2}q + \frac{1}{2\lambda_2}p \in W(\tilde{q})\right)\right\}
\end{aligned}$$

We work through the terms in the max. For the first term, recall that

$$\nu(p \in W(q)) = \nu(p \in W(\{q, \tilde{q}\})) = \max_{x \in \text{co}(\{q, \tilde{q}\})} \nu(p \in W(x)).$$

Thus, the first term is maximized at $\nu(p \in W(q))$ by setting $\lambda_1 = 1$. For the second term, suppose $\nu(x \in W(\tilde{q})) > \nu(p \in W(q))$ for some $x \in \text{co}(\{p,q\})$. We showed in Section A.13.1 that $\nu(p \in W(\tilde{q})) < \nu(p \in W(q))$, so $x \neq p$. We have $\nu(p \in W(x)) = \nu(p \in W(q))$ by Independence. By the first part of Lemma 17, $\nu(p \in W(\tilde{q})) \geq \nu(p \in W(q))$—contradiction. Conclude that the second term is no greater than $\nu(p \in W(q))$, so

$$\nu\left(\frac{1}{2}q + \frac{1}{2}p \in W(q, \tilde{q})\right) = \nu(p \in W(q)).$$

We have

$$\begin{aligned}
\rho\left(\frac{1}{2}q + \frac{1}{2}p \middle| \left\{\frac{1}{2}q + \frac{1}{2}p, q, \tilde{q}\right\}\right) &= \mu\left(\frac{1}{2}q + \frac{1}{2}p \succ q, \tilde{q}\right)\nu(p \in W(q)) \\
\rho(p|\{p,q,\tilde{q}\}) - \rho\left(\frac{1}{2}q + \frac{1}{2}p \middle| \left\{\frac{1}{2}q + \frac{1}{2}p, q, \tilde{q}\right\}\right) &= \mu\left(p \succ \tilde{q} \succ \frac{1}{2}q + \frac{1}{2}p\right)\nu(p \in W(q)). \qquad (16)
\end{aligned}$$

Finally, consider menu

$$\left\{\frac{1}{2}q + \frac{1}{2}p, x^*, q, \tilde{q}\right\}.$$

By Lemma 22,

$$\nu\left(q \in W\left(\left\{\frac{1}{2}q + \frac{1}{2}p, x^*, \tilde{q}\right\}\right)\right) = \nu(q \in W(\{p, \tilde{q}\})) = \nu(q \in W(\tilde{q}))$$
$$\nu\left(\tilde{q} \in W\left(\left\{\frac{1}{2}q + \frac{1}{2}p, x^*, q\right\}\right)\right) = \nu(\tilde{q} \in W(\{p, q\})) = 0.$$

Since x^* was chosen so that

$$\nu(x^* \in W(q)) \geq \nu\left(\frac{1}{2}q + \frac{1}{2}p \in W\left(\{x^*, q, \tilde{q}\}\right)\right),$$

Lemma 20 gives

$$\begin{aligned}\nu\left(\frac{1}{2}q + \frac{1}{2}p \in W\left(\{x^*, q, \tilde{q}\}\right)\right) &= \nu\left(\frac{1}{2}q + \frac{1}{2}p \in W(\{q, \tilde{q}\})\right) \\ &= \nu\left(\frac{1}{2}q + \frac{1}{2}p \in W(q)\right) \\ &= \nu(p \in W(q)).\end{aligned}$$

We have

$$\begin{aligned}
\rho\left(x^* \middle| \left\{\frac{1}{2}q+\frac{1}{2}p, x^*, q, \tilde{q}\right\}\right) &= \mu\left(x^* \succ \frac{1}{2}p+\frac{1}{2}q, \tilde{q}\right)\nu(x^* \notin W(q)) \\
\rho\left(\frac{1}{2}p+\frac{1}{2}q \middle| \left\{\frac{1}{2}q+\frac{1}{2}p, x^*, q, \tilde{q}\right\}\right) &= \mu\left(\frac{1}{2}p+\frac{1}{2}q \succ x^*, q\right)\nu(p \notin W(q)) \\
&\quad + \mu\left(x^* \succ \frac{1}{2}p+\frac{1}{2}q \succ \tilde{q}\right) \\
&\quad \times [\nu(p \notin W(q)) - \nu(x^* \notin W(q))] \\
\rho\left(\left\{x^*, \frac{1}{2}p+\frac{1}{2}q\right\} \middle| \left\{\frac{1}{2}p+\frac{1}{2}q, x^*, q, r\right\}\right) &= \left[\mu\left(\frac{1}{2}p+\frac{1}{2}q \succ d^*, q\right)\right. \\
&\quad \left. +\mu\left(d^* \succ \frac{1}{2}p+\frac{1}{2}q \succ \tilde{q}\right)\right]\nu(p \notin W(q)) \\
&\quad + \mu\left(d^* \succ \tilde{q} \succ \frac{1}{2}p+\frac{1}{2}q\right)\nu(x^* \notin W(q)) \\
&= \mu\left(\frac{1}{2}p+\frac{1}{2}q \succ q, \tilde{q}\right)\nu(p \notin W(q)) \\
&\quad + \mu\left(x^* \succ \tilde{q} \succ \frac{1}{2}p+\frac{1}{2}q\right)\nu(x^* \notin W(q)).
\end{aligned}$$

This implies

$$\begin{aligned}
&\rho\left(\left\{x^*, \frac{1}{2}p+\frac{1}{2}q\right\} \middle| \left\{\frac{1}{2}p+\frac{1}{2}q, x^*, q, r\right\}\right) - \rho\left(\frac{1}{2}q+\frac{1}{2}p \middle| \left\{\frac{1}{2}q+\frac{1}{2}p, q, \tilde{q}\right\}\right) \\
&\qquad = \mu\left(p \succ \tilde{q} \succ \frac{1}{2}p+\frac{1}{2}q\right)\nu(x^* \notin W(q)). \quad (17)
\end{aligned}$$

Combining (16) and (17),

$$\frac{\rho(p|p,q,\tilde{q}) - \rho(\frac{1}{2}p+\frac{1}{2}q|\left\{\frac{1}{2}p+\frac{1}{2}q, q, \tilde{q}\right\})}{\rho\left(\left\{x^*, \frac{1}{2}p+\frac{1}{2}q\right\} | \left\{\frac{1}{2}p+\frac{1}{2}q, x^*, q, \tilde{q}\right\}\right) - \rho\left(\frac{1}{2}p+\frac{1}{2}q| \left\{\frac{1}{2}p+\frac{1}{2}q, q, \tilde{q}\right\}\right)} = \frac{\nu(p \notin W(q))}{\nu(x^* \notin W(q))}. \quad (18)$$

Now we show how to identify $\nu(x^* \notin W(q))$. Choose any point $p^* \neq q$ on the boundary of $\{x \in \Delta(Z) : \nu(x \in W(q)) > 0\}$. We have $\nu(p^* \in W(q)) = 0$. but some sequence $\{p_i\}$ converging to p such that $\nu(p_i \in W(q)) > 0$ for all i. For each p_i, we can construct x_i^* and $\tilde{q}_i$ as above, and use them to recover

$$\frac{\nu(p_i \notin W(q))}{\nu(x_i^* \notin W(q))}.$$

Since each $x_i^* \in D(q)$, $\nu(x_i^* \notin W(q)) = \nu(\mathcal{U})$ for all i.

Fix any $t \in [\nu(\mathcal{U}), 1]$. Let $\mathcal{M}_t$ be the unique element of $\text{supp}(\nu)$ such that $\nu(\{\mathcal{M} : \mathcal{M} \supset t\}) = t$. $\mathcal{M}_t$ always exists by the second property of ν. Let $W_t(q) := W_{\mathcal{M}_t}(q)$. Let ϵ_t be the (Hausdorff)

distance between $W_t(q)$ and $W_1(q)$. Notice that $\lim_i \epsilon_{t_i} = 0$ by the first property of ν. Since no element of $B_{\epsilon_t}(p^*)$ can belong to $W_t(q)$, we must have $\nu(\tilde{p} \notin W(q)) \geq t$ for all $\tilde{p} \in B_{\epsilon_t}(p^*)$. Since $\lim_i p_i = p^*$, we have $\lim_i \nu(p_i \notin W(q)) = 1$.

This implies

$$\lim_i \frac{\nu(p_i \notin W(q))}{\nu(x_i^* \notin W(q))} = \frac{1}{\nu(\mathcal{U})}.$$

Plugging into (18) and using $\nu(x^* \notin W(q)) = \nu(\mathcal{U})$, we recover $\nu(p \notin W(q))$.

Since p was chosen arbitrarily subject to the requirement $\nu(p \in W(q)) > 0$, we can recover $\nu(x \in W(q))$ for all x such that $\nu(x \in W(q)) > 0$. For each $t \in (\nu(\mathcal{U}), 1]$, let

$$W_t(q) := \{x \in \Delta(Z) : \nu(x \notin W(q)) < t\}.$$

Let $W_{\nu(\mathcal{U})}(q) = \text{int}\left(\bigcap_{t \in (\nu(\mathcal{U}),1]} W_t(q)\right)$. (This set may be empty.) For $t \in [\nu(\mathcal{U}, 1]$, let

$$\mathcal{M}_t := \{\succsim \in \mathcal{U} : q \succ W_t(q)\}.$$

ν is pinned down by $\nu(\mathcal{U})$ and, for $t \in (\nu(\mathcal{U}), 1]$,

$$\nu(\{\mathcal{M} : \mathcal{M} \supset \mathcal{M}_t\}) = t.$$

A.14.3 Identifying μ

Now we recover μ. Fix any $A \in \mathcal{F}(\Delta(Z))$. Since we have already recovered ν, we can index the elements of A as follows:

$$\nu(a_1 \in W(A)) \geq \nu(a_2 \in W(A)) \geq \cdots \geq \nu\left(a_{|A|} \in W(A)\right).$$

To simplify notation, let

$$A^i := \{a_i, \ldots, a_{|A|}\}.$$

For each $i \in \{1, \ldots, |A|\}$, we have

$$\begin{aligned}
\rho(a_i|A) &= \mu(a_i \succsim A)\nu(a_i \notin W(A^{i+1})) \\
&\quad + \sum_{j<i} \mu\left(a_j \succ a_i \succsim A^{j+1}\right)\left[\nu(a_i \notin W(A^{i+1})) - \nu(a_j \notin W(A^{j+1}))\right] \qquad (19) \\
&= \mu(a_i \succsim A^i)\nu(a_i \notin W(A^{i+1})) - \sum_{j<i} \mu\left(a_j \succ a_i \succsim A^{j+1}\right)\nu(a_j \notin W(A^{j+1})). \qquad (20)
\end{aligned}$$

By Lemma 20, $\nu(a_i \in W(A)) = \nu(a_i \in W(A^j))$ for $j \leq i$. The ordering we used for A still works for A^j:

$$\nu(a_j \in W(A^j)) \geq \cdots \geq \nu\left(a_{|A|} \in W(A^j)\right).$$

For $j \leq i$, we have

$$\rho(a_i|A^j) = \mu(a_i \succsim A^i)\nu(a_i \notin W(A^{i+1})) \\ - \sum_{k \in \{j,\ldots,i-1\}} \mu\left(a_k \succ a_i \succsim A^{k+1}\right) \nu(a_k \notin W(A^{k+1})).$$

Combining this with (20),

$$\rho(a_i|A^{j+1}) - \rho(a_i|A^j) = \mu\left(a_j \succ a_i \succsim A^{j+1}\right) \nu(a_j \notin W(A^{j+1})).$$

Since we have already recovered $\nu(a_j \notin W(A^{j+1}))$, we can use this equation to recover $\mu(a_j \succ a_i \succsim A^{j+1})$. Plugging this back into (19) and rearranging, we can recover $\mu(a_i \succsim A)$. Since a_i was an arbitrary member of A, and A was an arbitrary menu, this is enough to fully recover μ.

A.15 Proof of Proposition 6

Suppose that $\mathcal{M}$ contains at least one preference that weakly prefers p to both q_1 and q_2, and at least one preference that weakly prefers both q_1 and q_2 to p. This implies $\mathcal{M}^{\text{avoid}}(\{p, q\}) = \{p, q\}$. Since $\mathcal{M}^{\text{avoid}} \subset \mathcal{M}$, it also implies $\mathcal{M}(\{p, q_1\}) = \{p, q_1\}$ and $\mathcal{M}(\{p, q_2\}) = \{p, q_2\}$. Since the DM does not face any constraints on any of the feasible menus, he is weakly better off acquiring information.

Now suppose that $\mathcal{M}$ contains at least one preference that weakly prefers p to both q_1 and q_2, but no preference that weakly prefers both q_1 and q_2 to p. This implies $\mathcal{M}^{\text{avoid}})(\{p, q\}) = \{p\}$. It also implies $p \in \mathcal{M}(\{p, q_1\})$ and $p \in \mathcal{M}(\{p, q_2\})$. Since the DM must choose p if he avoids information, and has the option of choosing p if he acquires information, he is weakly better off becoming informed.

Finally, suppose that $\mathcal{M}$ contains at least one preference that weakly prefers both q_1 and q_2 to p, but no preference that weakly prefers p to both q_1 and q_2. This implies This implies $\mathcal{M}^{\text{avoid}})(\{p, q\}) = \{q\}$. It also implies $q_1 \in \mathcal{M}(\{p, q_1\})$ and $q_2 \in \mathcal{M}(\{p, q_2\})$. Since $q = \alpha q_1 + (1 - \alpha)q_2$, the DM is weakly better off ex ante if he acquires information. This covers all the cases.

A.16 Proof of Proposition 7

Let $\mathcal{M}$ be the element of $\text{supp}(\nu)$ such that

$$\nu(\{\tilde{\mathcal{M}} : \tilde{\mathcal{M}} \subset \mathcal{M}) = \nu(q \in W(p)).$$

We need to find $q_1 \in \Delta(Z)$ such that $s(q_1) = s(p)$ and $m(q_1) > m(p)$ for all $m \in M := \{m \in \mathcal{M} : m(q) = m(p)\}$. Suppose there is no such q_1. First, suppose there is no x such that $m(x) > m(p)$ for all $m \in M$. Then, there is no x such that $m(p) > m(p + \lambda(p - x))$ for all $m \in M$ and all λ sufficiently small. Since $M \subset \mathcal{M}$, $W_{\mathcal{M}}(p)$ must be empty—contradiction.

Now suppose $s(x) > s(p)$ for all x such that $m(x) > m(p)$ for all $m \in M$. Since q is on the boundary of $W_{\mathcal{M}}(p)$, we can find $\tilde{q}$ arbitrarily close to q such that $m(\tilde{q}) > m(p)$ for all $m \in M$. In particular, we can choose $\tilde{q}$ close enough to q that $s(p) > s(\tilde{q})$—contradiction. Now suppose $s(x) < s(p)$ for all x such that $m(x) > m(p)$ for all $m \in M$. Take $\tilde{q}$ as above. Notice that $m(p + \lambda(p - \tilde{q})) < m(p)$ for all $m \in M$ for any λ such that $p + \lambda(p - \tilde{q}) \in \Delta(Z)$. Similarly, $s(p + \lambda(p - \tilde{q})) > s(p)$—contradiction. There must exist x, y such that $s(x) > s(p) > s(y)$ and $m(x), m(y) > m(p)$ for all $m \in \mathcal{M}$. Some combination of x and y is the desired q_1.

Notice that $m(q + \lambda(q - q_1)) < m(p)$ for all $m \in M$ and any λ such that $q + \lambda(q - q_1) \in \Delta(Z)$. By choosing λ sufficiently small, we can ensure that $m(q + \lambda(q - q_1)) < m(p)$ for all $m \in \mathcal{M}$, i.e. $q + \lambda(q - q_1) \in W_{\mathcal{M}}(p)$. Suppose not. Then we have sequences $\{m_i : m_i \in \mathcal{M} \setminus M\}$ and $\lambda_i \to 0$ such that $m_i(q + \lambda_i(q - q_1)) > m_i(p)$. We can shift and rescale the m_i so they belong to a compact subset of $\mathbb{R}^Z$, then pass to a convergent subsequence. Call the limit m^*. We have $m^*(q) \geq m^*(p)$. Since $\mathcal{M}$ is closed, $m^* \in \mathcal{M}$, so it cannot be that $m^*(q) > m^*(p)$. Conclude that $m^*(q) = m^*(p)$, so $m^* \in M$. Since each $m_i \in \mathcal{M} \setminus M$, $m_i(p) > m_i(q)$ for all i. Since $m_i(q + \lambda_i(q - q_1)) > m_i(p)$, we must have $m_i(q) > m_i(q_1)$ for all i, so $m^*(q) \geq m^*(q_1)$ for all i. This contradicts the definition of q_1, which requires $m^*(q_1) > m^*(p) = m^*(q)$.

Thus, we can find λ such that $q_2 := q + \lambda(q - q_1) \in W_{\mathcal{M}}(p)$. This implies $\nu(q_2 \in W(p)) > \nu(q \in W(p))$. Set $\alpha = \lambda/(1 + \lambda)$, so (q_1, q_2, α) is a signal for q. We show that

$$S_{(\mu,N)}\left(\alpha\delta_{\{p,q_1\}} + (1 - \alpha)\delta_{\{p,q_2\}}\right) > S_{(\mu,N)}\left(\delta_{\{p,q\}}\right).$$

for any μ such that $\text{supp}(\mu) = \{u \in \mathcal{U} : u(q) \geq u(p)\}$ and any N. (Self-knowledge N does not matter here because the DM does not get to choose whether to acquire information; he is required to be informed, or required to remain ignorant.) To simplify the notation, normalize $s(p)$ to 0. We have $0 = s(p) = s(q_1) > s(q) > s(q_2)$. Since the social planner does not care whether q_1 or p is chosen, we can break the support of μ into two groups: u such that $u(q_2), u(q) > u(p) = 0$, and u

such that $u(q_1) > u(q) > u(p) = 0 > u(q_2)$. In the first case,

$$S_{(u,N)}\left(\alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right) - S_{(u,N)}\left(\delta_{\{p,q\}}\right) \\ = (1-\alpha)s(q_2)\left[\nu(q_2 \notin W(p)) - \nu(q \notin W(p))\right] > 0.$$

In the second case,

$$S_{(u,N)}\left(\alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right) - S_{(u,N)}\left(\delta_{\{p,q\}}\right) = -(1-\alpha)s(q_2)\nu(q \notin W(p)) > 0.$$

Integrating over all u such that $u(q) \geq u(p)$, we conclude that the social planner is strictly better off imposing information than withholding it.

To show that $S_{(\mu,N)}\left(\left\{\delta_{\{p,q\}}, \alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right\}\right)$ is strictly between these two extremes, it suffices to show that one positive-measure group of DMs voluntarily acquires information, and another positive-measure group avoids it. For the first part, consider a DM with beliefs $\tilde{\nu}$ and utility u such that $u(q_1) > u(q) > u(p) = 0 > u(q_2)$. It is strictly optimal for him to acquire information if

$$\alpha u(q_1)\left[\tilde{\nu}(q_1 \notin W(p)) - \tilde{\nu}(q \notin W(p))\right] - (1-\alpha)u(q_2)\tilde{\nu}(q \notin W(p)) > 0. \tag{21}$$

Recall that $q_2 \in W_{\mathcal{M}}(p)$ and $q \notin W_{\mathcal{M}}(p)$. Since q is a convex combination of q_1 and q_2 and $W_{\mathcal{M}}(p)$ is convex, $q_1 \notin W_{\mathcal{M}}(p)$, so $\tilde{\nu}(q_1 \notin W(p)) \geq \tilde{\nu}(q \notin W(p))$. Thus, (21) will hold if $\tilde{\nu}(q \notin W(p)) > 0$, i.e. if $\tilde{\nu}(\{\tilde{\mathcal{M}} : \tilde{\mathcal{M}} \supseteq \mathcal{M}) > 0$. Let E be the set of $\tilde{\nu} \in \text{supp}(N)$ that satisfy this condition. By definition of self-knowledge,

$$\begin{aligned}\int_{\tilde{\nu}} \tilde{\nu}(\{\tilde{\mathcal{M}} : \tilde{\mathcal{M}} \supseteq \mathcal{M})dN(\tilde{\nu}) &= N(E)\int_{\tilde{\nu}\in E} \tilde{\nu}(\{\tilde{\mathcal{M}} : \tilde{\mathcal{M}} \supseteq \mathcal{M})dN(\tilde{\nu}|E) \\ &= \nu(\{\tilde{\mathcal{M}} : \tilde{\mathcal{M}} \supseteq \mathcal{M}) \\ &> 0.\end{aligned}$$

This implies $N(E) > 0$, so (21) holds for a positive-measure group of DMs.

Now consider a DM with beliefs $\tilde{\nu}$ and utility u such that $u(q_2) > u(q) > 0 > u(q_1)$. A DM in this set strictly prefers to avoid information if

$$(1-\alpha)u(q_2)\left[\tilde{\nu}(q_2 \notin W(p)) - \tilde{\nu}(q \notin W(p))\right] - \alpha u(q_1)\tilde{\nu}(q \notin W(p)) < 0. \tag{22}$$

For any $\tilde{\nu}$ such that $\tilde{\nu}(q_2 \notin W(p)) < \tilde{\nu}(q \notin W(p))$, this condition will hold if $u(q_2)$ is sufficiently large relative to $-u(q_1)$. We can use the same arguments as the previous step to show that N puts positive probability on $\tilde{\nu}(q_2 \notin W(p)) < \tilde{\nu}(q \notin W(p))$. Thus, (22) holds for a positive-measure group of DMs.

A.17 Proof of Proposition 8

Suppose that, for some $\mathcal{M} \in \text{supp}(\nu)$, there exist $m_1, m_2 \in \mathcal{M}$ such that $m_1(q_1) \geq m_1(p)$ and $m_2(q_2) \geq m_2(p)$. Suppose further that $m(p) > \min\{m(q_1), m(q_2)\}$ for all $m \in \mathcal{M}$. This implies

$$m(p) < \max\{m(p + \lambda(p - q_1)), m(p + \lambda(p - q_2))\} \tag{23}$$

for all $m \in \mathcal{M}$ and all $\lambda > 0$ such that $p + \lambda(p - q_1), p + \lambda(p - q_2) \in \Delta(Z)$. (Since p is interior, some such λ must exist.)

We showed in the proof of Theorem 3 (specifically, in the necessity proof of Convexity) that (23) implies the existence of $\alpha \in [0, 1]$ such that

$$m(p + \lambda(p - (\alpha q_1 + (1 - \alpha)q_2))) > m(p)$$

for all $m \in \mathcal{M}$. Rearranging, we have

$$m(p) > m(\alpha q_1 + (1 - \alpha)q_2)$$

for all $M \in \mathcal{M}$. That is, $q \in W_{\mathcal{M}}(p)$. Let $q := \alpha q_1 + (1 - \alpha)q_2$. Since $m_1(q_1) \geq m_1(p)$ and $m_2(q_2) \geq m_2(p)$, $q \neq q_1$ and $q \neq q_2$.

Fix any utility u such that $\min\{u(q_1), u(q_2)\} > u(p) = 0$ and any belief $\tilde{\nu}$. We have

$$\begin{aligned} &U_{(u,\tilde{\nu})}\left(\alpha\delta_{\{p,q_1\}} + (1 - \alpha)\delta_{\{p,q_2\}}\right) - U_{(u,\tilde{\nu})}\left(\delta_{\{p,q\}}\right) \\ &\quad = \alpha u(q_1)[\tilde{\nu}(q_1 \notin W(p)) - \tilde{\nu}(q \notin W(p))] + (1 - \alpha)u(q_2)[\tilde{\nu}(q_2 \notin W(p)) - \tilde{\nu}(q \notin W(p))]. \end{aligned} \tag{24}$$

Let $\underline{\mathcal{M}}$ be the member of $\text{supp}(\nu)$ such that

$$\nu(\{\tilde{\mathcal{M}} : \tilde{\mathcal{M}} \supseteq \underline{\mathcal{M}}\}) = \max\{\nu(q_1 \notin W(p)), \nu(q_2 \notin W(p))\},$$

and let $\bar{\mathcal{M}}$ be the member of $\text{supp}(\nu)$ such that

$$\nu(\{\tilde{\mathcal{M}} : \tilde{\mathcal{M}} \supseteq \bar{\mathcal{M}}\}) = \nu(q \notin W(p)).$$

Since we have $\mathcal{M}$ such that $q_1, q_2 \notin W_{\mathcal{M}}(p)$ but $q \in W_{\mathcal{M}}(p)$, we must have $\underline{\mathcal{M}} \subset \bar{\mathcal{M}}$. For any $\tilde{\nu}$ such that

$$\tilde{\nu}\left(\left\{\tilde{\mathcal{M}} : \bar{\mathcal{M}} \supset \tilde{\mathcal{M}} \supset \underline{\mathcal{M}}\right\}\right) > 0, \tag{25}$$

we have $\max\{\tilde{\nu}(q_1 \notin W(p)), \tilde{\nu}(q_2 \notin W(p))\} > \tilde{\nu}(q \notin W(p))$, so (24) is strictly positive. The DM strictly prefers to acquire information. Otherwise, the DM is indifferent to information. By assumption, he acquires information in this case as well.

Now consider a social planner with $s(p) = 0 > \max\{s(q_1), s(q_2)\}$ and a utility u such that $\min\{u(q_1), u(q_2)\} > u(p) = 0$. We have shown that a DM with utility u and beliefs $\tilde{\nu}$ strictly prefers to acquire information if and only if (25) holds. Fix any self-knowledge N, and let Let E be the set of $\tilde{\nu} \in \text{supp}(N)$ that satisfy (25). An equivalent definition is

$$E := \{\tilde{\nu} \in \text{supp}(N) : \max\{\tilde{\nu}(q_1 \notin W(p)), \tilde{\nu}(q_2 \notin W(p))\} > \tilde{\nu}(q \notin W(p))\}.$$

By definition of self-knowledge,

$$\begin{aligned}\int_{\tilde{\nu}} \tilde{\nu}\left(\left\{\tilde{\mathcal{M}} : \bar{\mathcal{M}} \supset \tilde{\mathcal{M}} \supset \underline{\mathcal{M}}\right\}\right) dN(\tilde{\nu}) &= N(E) \int_{\tilde{\nu} \in E} \tilde{\nu}\left(\left\{\tilde{\mathcal{M}} : \bar{\mathcal{M}} \supset \tilde{\mathcal{M}} \supset \underline{\mathcal{M}}\right\}\right) dN(\tilde{\nu}|E) \\ &= \nu\left(\left\{\tilde{\mathcal{M}} : \bar{\mathcal{M}} \supset \tilde{\mathcal{M}} \supset \underline{\mathcal{M}}\right\}\right) \\ &> 0.\end{aligned}$$

Thus, $N(E) > 0$. If all the DMs with beliefs in E acquire information, the cost to the social planner is

$$N(E)\left[\alpha[-s(q_1)]\int_E \tilde{\nu}(q_1 \notin W(p)) - \tilde{\nu}(q \notin W(p))N(d\tilde{\nu}) \right. \\ \left. +(1-\alpha)[-s(q_2)]\int_E \tilde{\nu}(q_2 \notin W(p)) - \tilde{\nu}(q \notin W(p))N(d\tilde{\nu})\right] > 0.$$

Since information does not change the behavior of any DM with beliefs outside E, we conclude that

$$S_{(u,N)}\left(\delta_{\{p,q\}}\right) > S_{(u,N)}\left(\alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right).$$

Fix any μ such that $\mu(\{u : \min\{u(q_1), u(q_2)\} > u(p)\}) = 1$. Integrating over $\text{supp}(\mu)$, we can replace subscript u with subscript μ.

Since the DMs in this society acquire information with probability 1, we also have

$$S_{(u,N)}\left(\alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right) = S_{(u,N)}\left(\left\{\delta_{\{p,q\}}, \alpha\delta_{\{p,q_1\}} + (1-\alpha)\delta_{\{p,q_2\}}\right\}\right).$$

B Model variants

B.1 Continuity without recoverability

We slightly strengthen local non-satiation.

Definition 41 (Locally non-satiated*)**.** *$\mathcal{M}$ is locally non-satiated* if, for any $a \in \mathcal{A}$, there exists $Z \in \mathcal{F}(B_\epsilon(a) \cap \mathcal{Z})$ such that Z is strictly preferred to a by all $m \in \mathcal{M}$, and a is strictly preferred to*

Z by u.

To get a version of Theorem 2 with local non-satiation* in place of non-satiation, we slightly strengthen Improvability.

Axiom 14 (Improvability*)**.** *For any $a \in \mathcal{A}$ and any $\epsilon > 0$, there is some $Z \in \mathcal{F}(B_\epsilon(a) \cap \mathcal{Z})$ such that $a \succ Z$ and $a \notin c(\{a\} \cup Z)$.*

Proposition 9. *Fix $(\succsim, c)$, and let*

$$\hat{c}(A) := \{a \in A : a \sim c(A)\}.$$

If $(\succsim, c)$ has a representation $(u, \mathcal{M})$ with $\mathcal{M}$ closed and locally non-satiated, then $(\succsim, \hat{c})$ has a recoverable representation $(u, \hat{\mathcal{M}})$ with $\hat{\mathcal{M}}$ closed and locally non-satiated*.*

Proof. Take any representation $(u, \mathcal{M})$ for $(\succsim, c)$, where $\mathcal{M}$ closed and locally non-satiated*. Since $\mathcal{M}$ is closed,

$$\bigcap_{m \in \mathcal{M}} \left\{ a \in \mathcal{A} : v(a) < \max_{b \in B} m(b) \right\}$$

is open for any $B \in \mathcal{F}(\mathcal{A})$. Enumerate the indifference classes of B from best to worst according to u: $u(B_1) > \cdots > u(B_n)$. The set

$$\bigcup_{i=1}^{n} \left\{ a \in \mathcal{A} : u(a) \geq u(B_i) \text{ and } \forall m \in \mathcal{M}\ v(a) < \max_{b \in B_i \cup \cdots \cup B_n} m(b) \right\}.$$

need not be open. Specifically, there may be a such that $u(a) = u(B_i)$ and $m(a) < \max_{b \in B_i \cup \cdots \cup B_n} m(b)$ for all $m \in \mathcal{M}$. In choice terms, there may be $a \sim B_i$ such that $a \notin c(\{a\} \cup B_i \cup \cdots \cup B_n)$. $\hat{c}$ eliminates this problem; $a \in \hat{c}(\{a\} \cup B_i \cup \cdots \cup B_n)$. $a \notin \hat{W}(B_i \cup \cdots \cup B_n)$, so $a \notin \hat{W}(B)$. We conclude that $\hat{W}(B)$ is open, so $\hat{c}$ satisfies Continuity.

It remains to show that $\hat{c}$ satisfies C-IUA and Improvability. Consider C-IUA first. Suppose that $d \in W(B)$ and $B \subset \bar{W}(A)$. We can assume $d \succsim B$ and $d \notin c(\{d\} \cup B)$. We will have $d \notin \hat{c}(\{d\} \cup B)$ only if $d \succ c(\{d\} \cup B)$; assume this is the case. By IUA, we can eliminate any $b \in B$ such that $b \sim d$ without changing choice. Thus, we can assume $d \succ B$. We can also assume that there are sequences $b_i \to b$ and $A_i \to A$ such that $b_i \in \hat{W}(A_i)$ for all i, and that $b \succsim A$ for some $b \in B$. We conclude that $d \succ A$. By C-IUA, we have $d \notin c(\{d\} \cup A)$. Since d is not indifferent to any item in A, $d \notin \hat{c}(\{d\} \cup A)$ as well. A parallel argument goes through if $d \in \bar{W}(B)$ and $B \subset W(A)$. Since $\hat{c}$ does not add any new item to any W or $\bar{W}$ this is all we need to check.

Now consider Improvability. Since $\mathcal{M}$ is locally non-satiated*, we can find $Z \in \mathcal{F}(\mathcal{Z})$ arbitrarily close to a such that $a \succ Z$ and $a \notin c(\{a\} \cup Z)$. This is still true for $\hat{c}$, so $(\succsim, \hat{c})$ satisfies Improvability*.

By Theorem 2, $\hat{c}$ has a recoverable representation $(u, \hat{\mathcal{M}})$ with $\hat{\mathcal{M}}$ closed and locally non-satiated*. □

B.2 Preferences fixed, constraints random, $|\mathcal{A}| = 3$

Let ρ be a stochastic choice function on $\mathcal{A}$. Let Π be the set of strict preferences on $\mathcal{A}$.

Definition 42 (Fixed-preference representation)**.** *A fixed-preference representation for ρ is $(\succ, \nu) \in \Pi \times \Delta(\mathcal{F}(\Pi))$ such that*

$$\rho(a|A) = \nu\left(\{\mathcal{M} \in \mathcal{F}(\Pi) : a = \arg\max(\mathcal{M}(A), \succ)\}\right)$$

where

$$\mathcal{M}(A) := \bigcup_{\succ_m \in \mathcal{M}} \arg\max(A, \succ_m).$$

Proposition 10. *Suppose $|\mathcal{A}| = 3$. The following are equivalent:*

1. *ρ has a fixed-preference representation.*
2. *There is at most one pair $(x, y) \in \mathcal{A}^2$ such that*

$$\rho(x|\{x, y\}) < \rho(x|\mathcal{A}).$$

Proof. Write $\mathcal{A} = \{a, b, d\}$. Suppose that

$$\rho(b|\{a, b\}) < \rho(b|\mathcal{A}).$$

(If there is no pair x, y such that $\rho(x|\{x, y\}) < \rho(x|\mathcal{A})$, then we have a standard Random Utility representation. We can define $\succ$ as usual, and let $\nu = \delta_{\succ}$.)

We take $\succ$ to be $a \succ b \succ d$. Now we show how to construct an appropriate ν. It is helpful to divide $\mathcal{F}(\Pi)$ into subsets ("states"), where each state is defined by the restrictions that prevent the DM from maximizing $\succ$. To formalize this, fix any $\mathcal{M} \in \mathcal{F}(\Pi)$, any $X \in \mathcal{F}(\mathcal{A})$ and $y \notin X$. Write $X \triangleright_{\mathcal{M}} y$ if the following three conditions are met: (1) $y \succ X$, (2) for each proper subset X' of X, $y \succ_m X'$ for some $\succ_m \in \mathcal{M}$, and (3) $\neg(y \succ_m X)$ for all $\succ_m \in \mathcal{M}$. Finally, define the state $X \triangleright y$ to be the set of $\mathcal{M}$ such that $X \triangleright_{\mathcal{M}} y$, but $\neg(Z \triangleright_{\mathcal{M}} w)$ for all $(Z, w) \neq (X, x)$.

When $|\mathcal{A}| = 3$ and $a \succ b \succ d$, there are eight states:

$$
\begin{aligned}
&(1) : \{b, d\} \triangleright a \\
&(2) : d \triangleright a \\
&(3) : d \triangleright a \text{ and } d \triangleright b \\
&(4) : b \triangleright a \\
&(5) : d \triangleright a \text{ and } b \triangleright a \\
&(6) : d \triangleright a \text{ and } d \triangleright b \text{ and } b \triangleright a \\
&(7) : d \triangleright b \\
&(8) : \text{ no constraints}
\end{aligned}
$$

The probabilities of states (7) and (8) are uniquely determined:

$$
\begin{aligned}
\Pr(7) &= \rho(d|b, d) - \rho(d|\mathcal{A}) \\
\Pr(8) &= \rho(b|b, d) - \rho(b|\mathcal{A}).
\end{aligned}
$$

The probabilities of the remaining states are not uniquely determined, but certain sums are:

$$
\begin{aligned}
\Pr(1) + \Pr(2) + \Pr(3) &= \rho(a|a, b) - \rho(a|\mathcal{A}) \\
\Pr(4) + \Pr(5) + \Pr(6) &= \rho(b|a, b) \\
\Pr(1) + \Pr(4) &= \rho(a|a, d) - \rho(a|\mathcal{A}) \\
\Pr(2) + \Pr(5) &= \rho(d|a, d) - \rho(d|\mathcal{A}) \\
\Pr(3) + \Pr(6) &= \rho(d|\mathcal{A}).
\end{aligned}
$$

Notice that states (1), (2), (3) can be transformed into (4), (5), (6) respectively by imposing $b \triangleright a$. Similarly, states (1), (4) can be transformed into (2), (5) by adding $d \triangleright a$, and then states (2), (5) can be transformed into (3), (6) by adding $d \triangleright b$. We will assume that, conditional on states (1)-(6) (equivalently, conditional on a being ruled out by a subset of $\{b, d\}$), $b \triangleright a$ is independent of

$d \triangleright a$ and $d \triangleright a, d \triangleright b$. This gives

$$\Pr(1) = \frac{(\rho(a|a,b) - \rho(a|\mathcal{A}))\,(\rho(a|a,d) - \rho(a|\mathcal{A}))}{1 - \rho(a|\mathcal{A})}$$
$$\Pr(2) = \frac{(\rho(a|a,b) - \rho(a|\mathcal{A}))\,(\rho(d|a,d) - \rho(d|\mathcal{A}))}{1 - \rho(a|\mathcal{A})}$$
$$\Pr(3) = \frac{(\rho(a|a,b) - \rho(a|\mathcal{A}))\,\rho(d|\mathcal{A})}{1 - \rho(a|\mathcal{A})}$$
$$\Pr(4) = \frac{\rho(b|a,b)\,(\rho(a|a,d) - \rho(a|\mathcal{A}))}{1 - \rho(a|\mathcal{A})}$$
$$\Pr(5) = \frac{\rho(b|a,b)\,(\rho(d|a,d) - \rho(d|\mathcal{A}))}{1 - \rho(a|\mathcal{A})}$$
$$\Pr(6) = \frac{\rho(b|a,b)\rho(d|\mathcal{A})}{1 - \rho(a|\mathcal{A})}.$$

Given our assumptions on ρ, all these probabilities will be positive. Moreover, the probabilities of all the states will sum to 1. It is easy to check that this assignment delivers the right predictions on all menus.

To complete the proof, we just need to pick a $\mathcal{M}$ that corresponds to each "state." For the state with no restrictions, we could pick $\{\succ\}$. For (7), we could pick $\{\succ_1\}$ such that $a \succ_1 d \succ_1 b$. For (1), we could pick $\{\succ_2, \succ_3\}$ such that $b \succ_2 a \succ_2 d$ and $d \succ_3 a \succ_3 b$. Proceeding in this way, we get an appropriate $\nu \in \Delta(2^{\Pi} \setminus \emptyset)$. □

As suggested in the proof, the distribution over exclusion relationships is not unique, even for $|\mathcal{A}| = 3$. One distribution can be picked out by imposing a particular independence assumption. It says: given that a is excluded by $\{b, d\}$ or a subset, the probability that b excludes a does not depend on whether d excludes a and b, d excludes a alone, or d excludes neither a nor b. Similarly, the probability that d excludes a (or that d excludes both a and b) does not depend on whether b excludes a. It may be possible to extend this idea to $|\mathcal{A}| > 3$, but the number of states is large even for $|\mathcal{A}| = 4$, and it is not obvious how to extend the independence assumption.

B.3 Preferences random, constraints fixed

Take $|\mathcal{A}| < \infty$. Take a stochastic choice function ρ on $\mathcal{F}(\mathcal{A})$. Let Π be the set of strict preferences on $\mathcal{A}$.

Definition 43 (Fixed-constraint representation)**.** *A fixed-constraint representation is* $\mathcal{M} \subseteq \Pi$ *and* $\mu \in int(\Delta(\Pi))$ *such that*

$$\rho(a|A) = \mu(\{\succ \in \Pi : a \succsim \mathcal{M}(A)\})$$

where

$$\mathcal{M}(A) = \bigcup_{\succ_m \in \mathcal{M}} \arg\max(\succ_m, A).$$

For any $A \subseteq \mathcal{A}$, let

$$S(A) = \text{supp}(\rho(\cdot|A)).$$

Axiom 15 (Plott)**.** *For any $A, B \in \mathcal{F}(\mathcal{A})$,*

$$S(A \cup B) = S(S(A) \cup B).$$

Axiom 16 (Support Dependence)**.** *For any $A, B \in \mathcal{F}(\mathcal{A})$, $S(A) = S(B)$ implies $\rho(\cdot|A) = \rho(\cdot|B)$.*

Let

$$S := \{(a, A) \in \mathcal{A} \times \mathcal{F}(\mathcal{A}) : a \in A \text{ and } A = S(A)\}.$$

Axiom 17 (No Arbitrage)**.** *Fix $\lambda \in \mathbb{R}^{|S|}$. If*

$$\sum_{(a,A)\in S} \lambda_i \mathbb{1}\{a \succsim A\} \geq 0, \tag{26}$$

for all $\succ \in \Pi$, then

$$\sum_{(a,A)\in S} \lambda_i \rho(a|A) \geq 0 \tag{27}$$

If in addition (26) holds with strict inequality for some $\succ$, then (27) holds with strict inequality.

Proposition 11. *ρ has a fixed-constraint representation if and only if it satisfies Plott, Support Dependence and No Arbitrage.*

Proof. Plott (1973) showed that the Plott axiom delivers $\mathcal{M} \subseteq \Pi$ such that

$$S(A) = \bigcup_{\succ_m \in \mathcal{M}} \arg\max(\succ_m, A).$$

By Support Dependence, it suffices to show that some $\mu \in \text{int}(\Delta(\Pi))$ explains choice on $\{A \in \mathcal{F}(\mathcal{A}) : A = S(B) \text{ for some } B\}$. Take any A, B such that $A = S(B)$. For each $a \in A$, there exists $\succ_m \in \mathcal{M}$ such that $a \succsim_m B$. Since $A \subseteq B$, $a \succsim_m B$ implies $a \succsim_m A$, so $A = S(A)$. Thus, it suffices to show that some μ explains choice on $\{A \in \mathcal{F}(\mathcal{A}) : A = S(A)\}$.

This follows from No Arbitrage. To see why, index the preferences in Π from $\succ_1$ to $\succ_{|\Pi|}$, and index the item-menu pairs in S from (a_1, A_1) to $(a_{|S|}, A_{|S|})$. Let X be the $|S|$-by-$|\Pi|$ matrix such that

$$X_{ij} = \mathbb{1}\{a_i \succsim_j A_i\}$$

for each i, j. Let y be the $|S|$-length vector in which

$$y_i = \rho(a_i|A_i).$$

We need a strictly positive $|\Pi|$-length vector μ such that $X\mu = y$. It is well known that such a μ exists if and only if there is no vector λ such that $X'\lambda \geq 0$ and $y'\lambda \leq 0$ with at least one inequality strict. This is precisely the No Arbitrage condition.[11] Now we show that μ has a unit sum. Fix any $a \in \mathcal{A}$. Since $a \succsim_i \{a\}$ for all i and since $\{a\} = S(\{a\})$, there is at least one row of X that consists entirely of ones. Since $\rho(a|\{a\}) = 1$, we have $\sum_j \mu_j = 1$. □

B.4 Constrained information choice

We extend the notion of a justification distribution to allow first-stage information choice to be less constrained than second-stage choice (without making it fully unconstrained). This is done by having each DM draw two different sets of justifications, one of which is weakly larger than the other. The larger set of justifications constrains information choice, while the smaller set constrains second-stage choice as usual.

Definition 44 (Augmented justification distribution)**.** *An augmented justification distribution is* $\nu \in \Delta(\mathfrak{U} \times \mathfrak{U})$ *that satisfies the following conditions:*

1. *The marginal of* ν *in its first dimension,* ν_1*, is a justification distribution.*
2. *For each* $\mathcal{M}_1$ *in supp*(ν_1)*, the conditional of* ν *on* $\mathcal{M}_1$*,* $\nu_2(\cdot; \mathcal{M}_1)$*, satisfies*

$$\nu_2(\{\mathcal{M}_2 : \mathcal{M}_2 \subseteq \mathcal{M}_1\}; \mathcal{M}_1) = 1.$$

The notion of self-knowledge can easily be extended to this setup by allowing private information about the second-stage justifications to depend on $\mathcal{M}_1$. Formally, a self-knowledge N must now satisfy

$$\int_{\tilde{\nu}} \tilde{\nu}(\cdot) N(d\tilde{\nu}; \mathcal{M}_1) = \nu_2(\cdot; \mathcal{M}_1).$$

Consider the following choice procedure, working backward from second-stage choice. Suppose the DM faces menu $\{p, q\}$ at $t = 2$ because he chose ignorance at $t = 1$. In that case, he maximizes his primary utility u over $\mathcal{M}_2(\mathcal{M}_1^{\text{avoid}}(\{p, q\}))$. Now suppose the DM faces menu $\{p, q\}$ at $t = 2$ because he was not offered information at $t = 1$. In that case, he maximizes u over $\mathcal{M}_2(\{p, q\})$ as usual. Finally, suppose the DM faces menu $\{p, q_i\}$ at $t = 2$ because he chose information, or

[11] The application of No Arbitrage to random utility is not new; it was done by Clark (1996). Clark used a slightly different version of the axiom, which does not require $\mu \in \text{int}(\Delta(\Pi))$.

was compelled to receive information, at $t = 1$, and q_i was realized. Again, he maximizes u over $\mathcal{M}_2(\{p, q_i\})$ as usual.

Now consider a DM who faces a choice between information and ignorance at $t = 1$. He knows $\mathcal{M}_1$ and has self-knowledge $N(\cdot; \mathcal{M}_1)$ about $\mathcal{M}_2$. If $\mathcal{M}_1^{\text{avoid}}$ is empty, he acquires information. Otherwise, he maximizes his expected utility given N and the $t = 2$ behavior spelled out above.

This version of the model connects the two extremes considered in Section 5.2. We recover the unconstrained-information case by setting $\nu_1 := \delta_{\mathcal{U}}$. We recover the fully-constrained information case by setting $\nu_2(\cdot; \mathcal{M}_1) = \delta_{\mathcal{M}_1}$.

Unlike the fully-constrained case, intermediate cases can predict information avoidance. To illustrate, consider a DM who knows both $\mathcal{M}_1$ and $\mathcal{M}_2$ at $t = 1$. Consider p, q_1, q_2 such that $q_1, q_2 \succ p$ and

$$\begin{aligned}
\mathcal{M}_1^{\text{avoid}}(\{p, q\}) &= \{p, q\} \\
\mathcal{M}_2(\{p, q\}) &= \{p, q\} \\
\mathcal{M}_2(\{p, q_1\}) &= \{p, q_1\} \\
\mathcal{M}_2(\{p, q_2\}) &= \{p\}.
\end{aligned}$$

(This will happen if $\mathcal{M}_1$, but not $\mathcal{M}_2$, contains a preference such that $q_1, q_2 \succsim_m p$, while both $\mathcal{M}_1$ and $\mathcal{M}_2$ contain a preference such that $q_1 \succ_m q \succ_m p \succ_m q_2$.) If the DM avoids information at $t = 1$, he will get q for sure, since

$$\mathcal{M}_2(\mathcal{M}_1^{\text{avoid}}(\{p, q\})) = \{p, q\}.$$

If the DM acquires information at $t = 1$, he will have to choose p when q_2 is realized. Thus, the DM will avoid information.

This is not to say that the DM will avoid information just as often as he would in the fully unconstrained case. Consider r_1, r_2 such that $r_1, r_2 \succ p$ and

$$\begin{aligned}
\mathcal{M}_1^{\text{avoid}}(\{p, r\}) &= \{p\} \\
\mathcal{M}_2(\{p, r\}) &= \{p, r\} \\
\mathcal{M}_2(\{p, r_1\}) &= \{p, r_1\} \\
\mathcal{M}_2(\{p, r_2\}) &= \{p\}.
\end{aligned}$$

(This will happen if neither $\mathcal{M}_1$ nor $\mathcal{M}_2$ contains any preference such that $r_1, r_2 \succsim_m p$, but both contain some preference such that $r_1 \succ_m r \succ_m p \succ_m r_2$.) If the DM avoids information at $t = 1$, he will have to choose p. If he acquires information, he can choose r_1 when it is realized. Thus, the DM will not avoid information. By contrast, he would avoid information if information choice

were fully unconstrained, i.e. if $\mathcal{M}_1$ were replaced with $\mathcal{U}$. He would get r for sure by remaining ignorant, but would have to choose p if he observed r_2.

Proposition 8 holds without modification in the augmented model. This is because no DM has an incentive to avoid information, so the constraints on information choice are not relevant.

Now consider Proposition 7. It is clear that the social planner will still prefer forcibly informing everyone to withholding information, since information choice does not play a role in that result. Moreover, the planner will still prefer providing information freely to withholding it, since some DMs will voluntarily become informed even if information choice is unconstrained (and even more DMs will make that choice if information choice is constrained). The only question is whether the planner will prefer forcibly informing everyone to withholding information—equivalently, whether some DMs will avoid information. This will certainly not be the case if information choice is fully constrained, so we will need to place additional restrictions on the augmented justification distribution ν. A natural restriction is, for each $\mathcal{M}_1^* \in \text{supp}(\nu_1)$,

$$\nu_2(\cdot; \mathcal{M}_1) = \nu_1(\cdot | \mathcal{M}_1 \subseteq \mathcal{M}_1^*). \tag{28}$$

This restriction can be interpreted as follows: the DM draws $\mathcal{M}_1$ from a standard justification distribution, and then draws $\mathcal{M}_2$ from the same justification distribution, conditional on $\mathcal{M}_2 \subseteq \mathcal{M}_1$. Notice that the subset of DMs who draw $\mathcal{M}_1 = \mathcal{U}$ will behave exactly like the population of DMs in Proposition 7. Since $\nu_1(\mathcal{U}) > 0$, there is a positive mass of such DMs. We showed in the proof of Proposition 7 that a positive mass of these DMs will avoid information. Thus, Proposition 7 holds given (28).