Expectations-Based Loss Aversion May Help Explain Seemingly Dominated Choices in Strategy-Proof Mechanisms

Bnaya Dreyfuss Ori Heffetz Matthew Rabin*

First draft: May 1, 2019
This version: June 6, 2021

Abstract

Deferred Acceptance (DA), a widely implemented algorithm, is meant to improve allocations: under classical preferences, it induces preference-concordant rankings. However, recent evidence shows that—in both real, large-stakes applications and experiments—participants frequently play seemingly dominated, significantly costly, strategies that avoid small chances of good outcomes. We show theoretically why, with expectations-based loss aversion, this behavior may be partly intentional. Reanalyzing existing experimental data on random serial dictatorship (a restriction of DA), we show that such reference-dependent preferences, with a degree and distribution of loss aversion that explain common levels of risk aversion elsewhere, fit the data better than no-loss-aversion preferences.

Keywords: dominated strategies, market design, obvious misrepresentation, school choice, strategy-proof, expectations, beliefs, reference-dependent preferences, endowment effect, prospect theory, first-order-stochastic-dominance violations

JEL Classification: B49, D47, D82, D84, D91

*Dreyfuss: Department of Economics, Harvard University (e-mail: bdreyfuss@g.harvard.edu); Heffetz: S.C. Johnson Graduate School of Management, Cornell University, Bogen Family Department of Economics and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, and NBER (e-mail: oh33@cornell.edu); Rabin: Department of Economics and Business School, Harvard University (email: matthewrabin@fas.harvard.edu). A version of this paper constituted Dreyfuss's MA thesis at Hebrew University. We thank Ned Augenblick, Alon Eizenberg, Zoë Hitzig, Shengwu Li, Alex Rees-Jones, Assaf Romm, Charlie Sprenger, the JESC-lab group and especially Matan Gibson, Ofer Glicksohn, Guy Ishai, Lev Maresca and Kobi Mizrahi, and seminar/conference participants at the Hebrew University, UCSB, University of Utah, UC Berkeley, the AEA Annual Meeting and BEAM for invaluable comments. This research project was supported by the I-CORE program of the Planning and Budgeting Committee and the Israel Science Foundation (grant no. 1821/12).

Blessed is he who expects nothing, for he shall never be disappointed.

Alexander Pope

Introduction

Centralized clearinghouses are used across the world to match applicants to positions. A common example is assigning students to schools. The assignment is determined as a function of students' rankings of schools and some combination of schools' rankings of students and predetermined priorities. These clearinghouses are increasingly adopting strategy-proof (SP) mechanisms, in which, with standard utility, a ranking of alternatives that does not coincide with an applicant's true preferences is a (weakly) dominated strategy. One such mechanism that is particularly popular is the deferred-acceptance algorithm (Gale and Shapley, 1962, henceforth DA).[1]

However, seemingly dominated rankings in DA-based allocation contexts are evident in a number of recent field and lab studies, reviewed in detail below. From the point of view of classical economic theory, the evidence is puzzling: in the lab, a substantial fraction of participants in simple allocation games submit lists that rank smaller amounts of money above larger ones; in real-world clearinghouses, supposedly well-informed school applicants do not rank a program that comes with a fellowship above the same program without the fellowship and, when such a ranking prompts a pop-up notification questioning their choice, they click OK to "keep the present preference ranking anyway."

This paper explores the idea that at least some apparently suboptimal behavior is in fact *intentional* behavior by individuals with expectations-based reference-dependent (EBRD) preferences. Intuitively, submission of a list creates expectations, and loss-averse individuals may want to avoid generating high expectations in order to avoid experiencing future disappointment. Specifically, an individual who is loss-averse around her endogenous expectation-based reference point may submit a list with a lower-value position (or a smaller sum of money) that she expects to get, above a higher-value position (or a larger sum) that she has little chance of getting. In the first part of the paper, using the framework of Kőszegi and Rabin (2009), we show this theoretically, and derive specific empirical predictions. In the second part, reanalyzing experimental data from Li (2017a), we show that the EBRD model fits the data substantially better than the no-loss-aversion, standard model.[2]

[1]For example, DA-based mechanisms are used in school choice in the US (e.g. Abdulkadiroğlu et al., 2005), Israel (Hassidim et al., 2018), Hungary (Shorrer and Sóvágó, 2018), and Australia (Artemov et al., 2017). For a recent review of the use of DA-based mechanisms, see Fack et al. (2019).

[2]In independent work, Meisner and von Wangenheim (2019) also study loss aversion and school choice,

Within the literature, a ranking that coincides with (true) preferences is referred to as *truthful*, and one that does not coincide with preferences is referred to as a *misrepresentation*. We reluctantly keep using these terms, even though they may be misnomers both from the EBRD-theory's perspective and from the point of view of those participating in the mechanisms, especially since participants are not typically told that their ranking is a matter of honesty. We keep this terminology mainly for consistency with the literature, use it purely descriptively, and imply no normative interpretation.

Briefly, the DA mechanism works—and is typically communicated to applicants—as follows: "The process begins with an attempt to match an applicant to the program most preferred on that applicant's rank-order list (ROL). If the applicant cannot be matched to that first-choice program, an attempt then is made to place the applicant into the second-choice program, and so on, until the applicant obtains a **tentative** match or all the applicant's choices on the ROL have been exhausted. When all applicant rank-order lists have been considered, all tentative matches become final."[3] The tentativeness of matches, which is emphasized to participants, means that in each round, all yet-unmatched and all already-(tentatively)-matched applicants to each program compete with each other anew. It is this feature of DA that guarantees that applicants cannot gain by misrepresenting their preferences. In contrast, the "Boston mechanism," explained in, e.g., Hakimov and Kübler (2019), is not SP. In that mechanism, all matches are final—hence it is sometimes referred to as the immediate-acceptance, rather than deferred-acceptance, algorithm. That "first come first served" algorithm is easy to manipulate: ranking a school higher than its value in order to increase its admission probability is in fact, in many cases, a best response.

Evidence of misrepresentation in DA settings is found in several lab experiments. In a typical experiment, participants submit ROLs of hypothetical positions that are assigned different amounts of real money (see Hakimov and Kübler, 2019 for a recent review). Chen and Sönmez (2006), Calsamiglia et al. (2010), Klijn et al. (2013), Ding and Schotter (2017), and Li (2017a), for example, find that between 19 and 45 percent of submitted ROLs are not ranked monotonically by their monetary value. Notably, Rees-Jones and Skowronek (2018) conduct an online experiment with 1,714 medical students immediately after their partici-

focusing on strategic interactions.

[3]This text is copied (unedited, except for the addition of four hyphens) from the FAQ page on the website of the National Resident Matching Program (NRMP), which we briefly discuss below. The original text ends with a weblink to a "How the Matching Algorithm Works" page (`http://www.nrmp.org/matching-algorithm/`), which contains a short animated video that explains the mechanism and demonstrates it with a simple example. The introduction to the video reads: "The NRMP uses a mathematical algorithm to place applicants into residency and fellowship positions. Research on the algorithm was the basis for awarding the 2012 Nobel Prize in Economic Sciences. To make the matching algorithm work best for you, create your rank order list in order of your *true* preferences, not how you think you will match."

pation in the National Resident Matching Program (NRMP), a much-studied example of a carefully designed strategy-proof market, and find that when placed in an analogous, incentivized matching task, 23 percent of participants misrepresent their preferences.[4] In these experiments, applicants often seem to factor into their ranking the probability of admission on top of monetary value.

In contrast with these lab settings, where (because respondents are essentially asked to rank amounts of money) preferences are assumed to be known, in field settings identifying misrepresentation is harder, because applicants' preferences over programs are not generally known. In recent studies, however, researchers identified an important exception: in some DA-based allocation mechanisms, applicants submit ROLs that may include, as two separately ranked options, a funded position—e.g., a study program with a fellowship—and an identical non-funded position—i.e., the *same* program without the fellowship. This allows researchers to detect what they term *obvious misrepresentations*: ROLs where a funded position is not ranked above an identical non-funded one (Hassidim et al., 2018; Shorrer and Sóvágó, 2018; Artemov et al., 2017). Obvious misrepresentations are further categorized by researchers into *obvious flippings*, i.e., ROLs with a funded position ranked below an identical non-funded one, and *obvious droppings*, i.e., ROLs that include a non-funded position, but not an identical funded one.[5]

Like the misrepresentations found in the lab, the obvious misrepresentations identified in the field are prevalent. The studies above find that of the ROLs that contain at least one program that offers funding, between 19 and 35 percent obviously misrepresent. Generally, applicants who are less likely to receive funding (as determined by measures of ability and socioeconomic status) are found more likely to obviously misrepresent. While this finding may suggest that misrepresentation is common mainly when it is likely to be irrelevant, these misrepresentations are sometimes costly. For example, Shorrer and Sóvágó (2018) show that in the Hungarian college-admissions market 12–19 percent of the obvious misrepresentations are ex-post costly, with an average cost of \$347–\$735 per misrepresentation (\$3,000–\$3,500 per costly misrepresentation). Similar evidence has been found in Israel (2–8 percent of obvious misrepresentations are ex-post costly; Hassidim et al., 2018) and Australia (1–20 percent; Artemov et al., 2017). Of course, obvious misrepresentations are a lower bound; their prevalence in these markets raises the possibility of less-than-obvious misrepresentation

[4]Although the NRMP itself uses a modified version of DA that complicates the strategic incentives, simulations in Roth and Peranson (1999) suggest that effectively all students remain incentivized to truthfully reveal their preferences.

[5]Note, however, that an obvious flipping and an obvious dropping result in the same outcome. If the applicant is admitted to the non-funded program, then all the programs that she ranked below the non-funded program become irrelevant. If she is rejected from the non-funded program, then she will also be rejected from an identical funded one.

as a potentially broader phenomenon in various DA-based markets. We return to this point below.

In all the above studies, the authors offer explanations for costly preference misrepresentation. These fall into two main categories: incorrect beliefs and non-classical preferences. We start with the first. Extreme pessimism could lead applicants to incorrectly assign essentially zero probability to the event of getting funding, and so they simply do not bother ranking funded positions. Such over-pessimism seems less likely to explain results from the lab, where it is not clear what prevalent bias would generate that particular belief. Alternatively, applicants could fail to understand the mechanism's strategy-proofness. The error most plausible to explain the observed misrepresentation is a mistaken belief that higher ranking will increase the probability of admission—as is true for mechanisms such as the Boston mechanism discussed earlier. However, such a mistaken belief cannot explain obvious droppings. Moreover, in markets with admission-score cutoffs, where applicants are matched with their highest-ranked program whose (ex-post common-knowledge) cutoff does not exceed their (privately known) admission score (e.g., Artemov et al., 2017; Shorrer and Sóvágó, 2018), such a mistake would amount to believing that ranking a program higher could *decrease* its admission-score cutoff. In addition, this explanation appears implausible in settings where the information regarding strategy-proofness is made salient to applicants.[6]

The second category of explanations is closer to the focus of our paper, as it concerns utility-maximizing behavior with correct beliefs but non-classical *preferences*. Ego utility (Kőszegi, 2006), for example, can predict that applicants may choose to rank higher those schools at which they have higher chances of admission, in order to avoid learning that they would be rejected by more attractive schools. However, with cutoff scores, this explanation too seems implausible, because applicants generally cannot avoid the information regarding what would happen if their ROL were truthful. Altruistic motives can also explain why some applicants may omit funded positions. Finally, by not applying for funding, applicants may attempt to signal (including to themselves) pro-sociality, generosity, or wealth. While we believe that wrong beliefs, simple mistakes, and the above non-classical preferences may all play a role in explaining some of the misrepresentation, in the rest of this paper we explore the role that EBRD may play.

In section 1 we outline the EBRD model. Developed by Kőszegi and Rabin (2006; 2007; 2009), the model builds on Prospect Theory (Kahneman and Tversky, 1979), and is meant to generalize and formalize its value function. (The ideas of framing, editing, and probability

[6]The FAQ section in Hassidim et al. (2018) explicitly advises applicants to rank according to preferences. For experimental evidence on the effect of advice on behavior in a different strategy-proof mechanism (top trading cycles), see Guillen and Hing (2014) and Guillen and Hakimov (2018).

weighting are not included; in section 3.6 we discuss how including probability weighting may affect our predictions.) The model's two main purposes are, first, to explain how reference points affect behavior, i.e., how people react to various departures from a posited reference point, and, second, to define how said reference points are endogenously determined. Because different ROLs induce different expectations and hence different reference points, the utility function in the EBRD model is ROL-dependent: when the agent optimizes she takes into account the effect of the ROL on her expectations and hence on her reference point. Other models of disappointment aversion, such as Bell (1985), Loomes and Sugden (1986) and Gul (1991) have similar predictions. However, the last two do not predict obvious misrepresentations, as they preclude violations of first-order stochastic dominance (FOSD).

In section 2 we derive predictions. First, we analyze a simple case where there are only two schools. We show that for loss-averse agents, there are plausible conditions where the utility-maximizing ROL differs from ranking by value, with either the order of schools flipped or a school omitted. Intuitively, an applicant who is likely to get matched with a school will feel a loss when matched with any other school (even a better one); this can create attachment to the high-probability school—an *endowment effect* for schools. By moving the (possibly less-preferred) school up on her ROL, and updating her reference point accordingly, she can reduce the chance of loss. In addition, avoiding high expectations of matching with other, possibly better schools reduces the chances of the painful loss of not matching with them. For agents with a high degree of loss aversion, when this latter effect is strong enough, it can even result in the omission of a school altogether (rather than just flipping the order of schools), violating FOSD. We provide the most general result in subsection 2.5, where we show that the main two-schools result can be generalized to any number of schools. Specifically, we show that so long as the likelihood of a school's admission does not increase with the school's value, for plausible combinations of parameters misrepresentation always occurs.

Subsection 2.6 discusses field evidence consistent with this general result. That evidence, of misrepresentations that are not necessarily FOSD violating, is inherently weaker (and hence is not discussed above): it relies on assumptions regarding preferences that may be less obvious than that people like money. At the same time, such less-than-obvious misrepresentations are arguably more important for three main reasons. First, they may occur under *any* degree of loss aversion, and are hence predicted to be much more prevalent. Second, generic DA-governed markets typically do not contain a direct monetary aspect. Third, even when they do, the lifetime material and non-material loss from not matching with the most preferred schools can be much higher than the funding loss due to a payoff-relevant obvious misrepresentation.

Since the most conclusive evidence for misrepresentation concerns FOSD violations, we

return to focusing on them in the remainder of the paper. Prefacing our main empirical analysis in section 3, we close section 2 with a simple example that demonstrates the model's explanation of existing field evidence of obvious misrepresentation.

All else equal, the EBRD model predicts that some applicants with low chances of matching with funded positions—generally, weaker students—will prefer not to rank them above identical non-funded ones. As discussed above, this prediction seems to be consistent with existing empirical evidence. Of course, in the field, low probability of getting into a funded program may be highly correlated with low ability, and therefore may also be correlated with ability-related mistakes, such as misunderstanding of the strategy-proofness of the DA mechanism. However, Shorrer and Sóvágó (2018) exploit a reform that substantially increased the selectivity of admission with funding in some fields of study in order to show that lowering the probability of receiving funding *causally* increases the number of observed misrepresentations.

Naturally, with this field evidence alone, a theory in which applicants with classical preferences are more likely to make mistakes when they are less costly may be enough to explain the data. In section 3 we reanalyze the data from one of the experimental treatments in Li (2017a), and test our EBRD explanation against such a benchmark.[7] In that treatment, over ten rounds, four different sums of money, between \$0 and \$1.25, are allocated to four participants using Random Serial Dictatorship (RSD)—a restriction of DA to one-sided markets with priorities. In RSD, players' ROLs are processed by a randomly-determined priority, and each player receives the highest-ranked prize on her ROL that is still left to choose from by the time her list is processed. Participants first learn their randomly drawn "priority score"—an integer between 1 and 10 that strongly signals their actual priority—and then rank the four monetary prizes in any order they choose. After describing the experimental setup, we present detailed theoretical predictions: for each possible choice situation, we show the expected EBRD utility from the submission of each possible ROL as a function of the degree of loss aversion. A loss-averse participant may rank high amounts of money lower when her priority score is lower.

Li (2017a) finds that 29 percent of submitted ROLs are not ordered by value. Further analyzing Li's data, Hassidim et al. (2018) find that misrepresentation significantly increases as priority score decreases. We take a further step and explore the extent to which the empirical misrepresentation patterns match the specific patterns predicted by our model. The distributions of ROLs conditional on different priority scores seem to fit the EBRD

[7]We do not analyze data from Li's other DA-related treatment since there, the EBRD model's predictions coincide with those of the classical, reference-independent-preferences model. (These coinciding predictions are largely confirmed; see footnotes 24 and 32 below.) Li has additional treatments that involve auctions, which we do not explore in this paper.

model's predictions: some subjects indeed seem to try to maximize, given their priority score, their chances of winning their first *ranked* choice, even at the expense of reducing their expected monetary payoff. For example, as we show in table 3, as the ROL-submitter's priority score decreases, the percent of ROLs with prizes ranked by their value decreases steeply: it is at most 91, 79, and 61 percent, respectively, for subjects with the three highest, four middle, and three lowest priority scores. At the same time, the percent of ROLs with the *lowest* prize ranked *first* increases as priority decreases: it is up to 3, 11, and 21 percent for high-, middle-, and low-priority-score subjects, respectively. In between, a hump-shaped upper-bound of 13, 29, and 25 percent of ROLs, respectively, rank one of the two middle prizes on top of their list—consistent with the theory.

Our main empirical results are in subsection 3.4. We formally test the fit of the EBRD model to the data, comparing it to the fit of the no-loss-aversion alternative. We first use a logit, random-utility model (where choosers make randomly drawn mistakes, whose probability decreases with their cost) to estimate a single (population-wide) coefficient of loss aversion, λ. We estimate $\lambda = 2.0$, with a standard error of 0.2, so the data clearly favors $\lambda > 1$ over the nested classical non-EBRD preferences, corresponding to $\lambda = 1$. A likelihood-ratio test statistic $= 32$ (p-value < 0.00001) suggests that given the model, the observed choices are more than 10^7 times likelier with $\lambda = 2$ than with $\lambda = 1$. This result passes several placebo/permutation tests, as the same equality-of-fit hypothesis is never rejected at such significance levels on synthetic, randomly generated datasets that we construct to match the real data on several dimensions. The result holds in both earlier and later rounds of the experiment, and is also replicated ($\lambda = 2.2$, SE $= 0.4$) on a second dataset: a one-shot (rather than a ten-round) version of the same experiment, with stakes that are 12 times higher, conducted by Li (2017a) after the original experiment. Finally, we estimate both individual-level and (parametric) population-wide models of heterogeneity in the coefficient of loss aversion, and find mean $\lambda = 1.6$–3.0, with substantial heterogeneity. We close section 3 with further evidence that Li's subjects did not misperceive the mechanism in the experiment to be the Boston mechanism, followed by a brief discussion of past experimental evidence of FOSD violations that can be explained by the EBRD model.

In section 4 we conclude, and discuss implications of our findings. Following our analysis, a seemingly dominated choice may not be a calculation mistake. But, if EBRD preferences themselves are mistaken with respect to some notion of agents' true preferences, then a seemingly dominated choice may still suggest a mistake of a different kind.

1 A Model of News Utility

We use a simplified version of Kőszegi and Rabin's (2009) EBRD framework. Lori is a decision maker (DM) with EBRD preferences, about to submit her ROL of schools to a DA-based centralized clearinghouse.

In the general model, Lori lives $T+1$ periods. In each period, a K-dimensional consumption vector $\boldsymbol{c_t} = (c_t^1, ...c_t^K)$ is realized. In our setting, we parse Lori's life into 3 periods, $t = 1, 2, 3$. Lori consumes only in period 3, and the only consumption dimensions are school-determined lifetime consumption—a separate dimension for each school—and, in the last example below, money. Since in many cases, matching with a specific school determines much of the rest of an applicant's lifetime experience—including one's social circle, economic status, friends, spouse, city of residence, etc.—the "consumption" of a specific school in our setting represents attending the school as well as enjoying the resulting flow of lifetime consumption. In period 1, Lori learns about the submission process, forms beliefs about her chances, and submits her ROL. In period 2, the uncertainty resolves and she learns about her match. In period 3, which represents the rest of Lori's life, she consumes. (We discuss this periodization in section 2.2.)

In the general model, Lori starts every period t with beliefs she inherited from the previous period, denoted $F_{t-1} = \{F_{t-1,\tau}\}|_{\tau=t}^{T}$, where $F_{t,\tau}$ are the beliefs Lori holds in period t regarding her consumption in period $\tau \geq t$ in each of the K dimensions. Within each period, up to five things may happen: some of the uncertainty may resolve, Lori may take an action, further uncertainty may resolve, $\boldsymbol{c_t}$ is consumed, and Lori updates her beliefs $F_t = \{F_{t,\tau}\}|_{\tau=t}^{T}$, where $F_{t,t}$ assign probability 1 to $\boldsymbol{c_t}$.

In our setting, Lori takes one action: the submission of a ROL in period 1, which determines $F_{1,3}$, Lori's beliefs at the end of period 1 regarding her consumption in period 3. These beliefs are the lottery generated by the ROL she submitted. After resolution in period 2, Lori's beliefs are $F_{2,3} = F_{3,3}$, which assign probability 1 to the consumption of the matched school.

In the general model, Lori's utility is the sum of two components: classical consumption utility, and news utility:

$$\boldsymbol{u_t} = m\left(\boldsymbol{c_t}\right) + \sum_{\tau=t}^{T} N\left(F_{t,\tau}|F_{t-1,\tau}\right).$$

The term $m\left(\boldsymbol{c_t}\right)$ corresponds to classical (reference-independent) consumption utility. Consumption utility is additively separable across dimensions, and the consumption-utility function in each dimension k, $m^k\left(\cdot\right)$, is assumed to be differentiable and strictly increasing.

$N\left(F_{t,\tau}|F_{t-1,\tau}\right)$, defined below, is the "gain-loss" utility function. For $\tau = t$, it is Lori's "contemporaneous gain-loss" utility, which measures how she experiences consumption relative to what she believed entering period t; for $\tau > t$, $N\left(F_{t,\tau}|F_{t-1,\tau}\right)$ is her "prospective gain-loss" utility, which measures how Lori reacts to news about future consumption.[8]

The model defines $N\left(F|F\right)$ as follows. For any distribution F over $\mathbb{R}$ and any $p \in (0,1)$ let $c_F(p)$ denote the consumption level at percentile p. $c_F\left(p\right)$ is essentially the inverse of $F(\cdot)$, and is defined by satisfying the following conditions:

$$F\left(c_F\left(p\right)\right) \geq p,$$

and

$$F\left(c\right) < p \text{ for all } c < c_F\left(p\right).$$

The gain-loss utility from changes in beliefs regarding the consumption in dimension k is defined as

$$N^k\left(F^k_{t,\tau}|F^k_{t-1,\tau}\right) = \int_0^1 \mu\left(m^k\left(c_{F^k_{t,\tau}}\left(p\right)\right) - m^k\left(c_{F^k_{t-1,\tau}}\left(p\right)\right)\right) dp,$$

where $\mu\left(\cdot\right)$ is a gain-loss utility function defined in the following way:

$$\mu\left(x\right) = \begin{cases} \eta x & x \geq 0 \\ \eta\lambda x & x < 0. \end{cases}$$

$\eta > 0$ is the weight Lori assigns to changes in her beliefs, and $\lambda > 1$ is the coefficient of loss aversion, which measures how losses are weighted relative to gains.[9]

While the definition of $N(F|F)$ is notationally cumbersome, the intuition behind it is quite simple: Lori compares her old beliefs to her new beliefs percentile by percentile: in each period t, Lori compares the worst outcomes under $F^k_{t,\tau}\left(\cdot\right)$ to the worst outcomes under $F^k_{t-1,\tau}\left(\cdot\right)$, the second-worst outcomes under $F^k_{t,\tau}\left(\cdot\right)$ to the second-worst outcomes under $F^k_{t-1,\tau}\left(\cdot\right)$, and so on.

Table 1 summarizes Lori's utility and beliefs from a period-t surprise lottery L that will be resolved in period $t+1$. The lottery will give Lori consumption, in period $t+2$, of x units in dimension k with probability p and 0 units otherwise. We normalize $m^k\left(0\right) = 0$.

[8] Kőszegi and Rabin (2009) have an additional parameter, $\gamma_{t,\tau}$, the weights Lori assigns to hearing news in period t regarding her consumption in period $\tau \geq t$. In our analysis, we assume news about period-3 consumption (the only consumption period) to have the same weight in periods 1 and 2, and therefore further normalize all γ's $= 1$.

[9] In all the settings examined in this paper, $\lambda \in [0,1]$ yields predictions identical to those of classical preferences.

Table 1: News Utility, Consumption Utility, and Beliefs in the Three Time Periods

	t Learning	$t+1$ Resolution		$t+2$ Consumption	
Probability	1	p	$1-p$	p	$1-p$
Consumption in Current Period	0	0	0	x	0
Expectations About Consumption in $t+2$	$L = \{x, p; 0, 1-p\}$	$\{x, 1\}$	$\{0, 1\}$	$\{x, 1\}$	$\{0, 1\}$
News Utility	$N^k(L\|0) = \eta p m^k(x)$	$N^k(x\|L) = \eta(1-p)m^k(x)$	$N^k(0\|L) = -\eta\lambda p m^k(x)$	$N^k(x\|x) = 0$	$N^k(0\|0) = 0$
Consump. Utility	0	0	0	$m^k(x)$	0

Summing over the bottom two lines of the table, weighted by the corresponding probabilities in the top line, Lori's expected lifetime utility in dimension k from this lottery is given by

$$u^k = p(1+\eta)m^k(x) - p(1-p)\eta(\lambda-1)m^k(x), \tag{1}$$

where the first term represents expected consumption utility and news utility (or gain-loss utility) from learning about the lottery, and the second term, which is always negative (because $\lambda > 1$), represents expected news utility from the resolution of the lottery.

Although the model as written above has two parameters, equation (1) can be rearranged to have a single parameter, $\Lambda \equiv \frac{1+\eta\lambda}{1+\eta}$, sometimes referred to as the *behavioral*, or *de facto*, coefficient of loss aversion. However, it is common in the literature to discuss results in terms of λ (rather than Λ) under the identifying assumption that $\eta = 1$, and we follow this practice below. With this convention, equation (1) simplifies to

$$u^k = 2pm^k(x) - p(1-p)(\lambda-1)m^k(x). \tag{2}$$

2 Predictions

We start with a setting where Lori ranks n different schools, and we analyze school flipping and omission. We postpone our analysis of funded positions to subsection 2.7.

2.1 Defining the Problem

There are N schools $s \in S$, and a unit continuum of students with a representative element $\theta \in [0,1]$. Students have strict preferences over S. Each school has capacity $\psi_s \in (0,1]$, and a priority score for each student denoted by $\epsilon_{\theta s}$. As shown in Azevedo and Leshno (2016), under some regularity conditions, the unique stable matching in this economy can be described by a vector of cutoffs, with κ_s the cutoff at school s. Schools are said to accept students whose priority scores are above the cutoff. The market is governed by (the continuum-economy equivalent of) DA, and therefore each student receives her top-ranked school in which she is acceptable, that is, the highest-ranked school such that $\epsilon_{\theta s} > \kappa_s$.

Lori is a student who has preference over a set of schools $\{s_1, \ldots, s_N\}$, ordered from most to least preferred. To get a match using DA, Lori submits her ranking over schools in an l-long list (ROL) with $0 \leq l \leq N$. Lori does not know her $\epsilon_{\theta s}$ at each school, and is uncertain about which school (if any) she will be matched with given a submitted ROL. Therefore, choosing a ROL is, effectively, choosing a lottery over schools.

One important modeling assumption regards consumption dimensions; results depend on whether different schools belong to the same consumption dimension or to separate ones. Since, as mentioned above, we view a match with a program as potentially determining many aspects of Lori's life, we view different schools as different lifetime sequences, and model them as at least partially separate consumption dimensions. This important assumption is simplifying and restrictive, equivalent to assuming that each school has one dominating dimension that in all other schools is at the same lower level.[10]

To analyze ROLs as lotteries, we distinguish between the probability of admission to a school *conditional on proposing to it* (that is, conditional on being rejected from all schools ranked above it), and the *unconditional* probability of admission to a school given a submitted ROL. As discussed above, conditional on proposing to school s, Lori is admitted if $\epsilon_{\theta s} > \kappa_s$.

Lori knows κ_s, and has prior beliefs F_s about her $\epsilon_{\theta s}$ at each school s.[11] Since we analyze the decision of a single DM, for ease of presentation we drop the subscript θ. The probability of admission to a school conditional on proposing to it is therefore given by

$$Pr(\epsilon_s > \kappa_s | \text{rejections from all schools ranked above } s).$$

[10] Another de facto modeling assumption regards background risk (see Kőszegi and Rabin, 2009, web appendix). Since a match determines a significant part of Lori's lifetime consumption, we view the match uncertainty as itself representing much of Lori's background risk; we effectively assume that additional background risk is negligible.

[11] Notice that this setup can easily incorporate uncertainty about the cutoff κ_s by redefining $\epsilon_{\theta s}$ as the sum of two terms, one representing uncertain applicant-school characteristics, and the other uncertain school cutoff.

We analyze Lori's decision under two opposite extreme assumptions regarding the correlations between priorities at different schools. First, we assume that they are independent: $\epsilon_s \perp \epsilon_{s'}$ for all $s \neq s'$. This assumption simplifies the presentation and analysis in the rest of this section, but may not be realistic in some real-world applications. Second, we assume that they are perfectly correlated: $\epsilon_s = \epsilon$ for all $s \in S$. We focus on this case as arguably the most empirically relevant alternative: it applies to all situations where schools rank applicants according to some common standardized test (as is the case in, e.g., Shorrer and Sóvágó, 2018; see Fack et al., 2019 for a review).[12] The analytical details of this case are in appendix A, but throughout this section we present (in figure 1 below) and discuss our results under both cases. We show that they are remarkably similar (our main results hold under even weaker conditions under the perfect-correlation assumption).

Assuming priorities at different schools are independent, we can simply write the conditional probability of admission to a school as $Pr(\epsilon_{s_i} > \kappa_{s_i}) \equiv q_i$. The set of possible ROLs, $\mathcal{R}$, contains all possible permutations of all sets contained in the power set of S. Given market conditions, submission of $r \in \mathcal{R}$ corresponds to a lottery over schools. We will later refer to the probability distribution that defines this lottery as the vector $\boldsymbol{p}(r)$. Under independence of priorities, $p_i(r)$ is the product of the probabilities of rejections from s_j for all j ranked above i, times the probability of admission to s_i. For example, if r^* is the (consumption-utility) truthful list, $p_i(r^*) = q_i \cdot \prod_{j=1}^{i-1} (1 - q_j)$.[13]

We start by assuming $S = \{s_1, s_2\}$, so Lori's consumption vector is two-dimensional: (s_1, s_2). (In subsection 2.5 we extend our main result to n schools.) Denote $s_i = 1$ if Lori is matched with school s_i. Note that $(1, 1)$ is not possible, but $(0, 0)$ is. Denote $m^i(1) = m_i > 0$ the utility from being matched with school s_i, and normalize $m^i(0) = 0$. Finally, WLOG assume $m_1 > m_2$, i.e., Lori's consumption utility is higher for s_1.

The lottery that corresponds to each ROL r is denoted by $L(r)$, and the ranking implied by r is denoted by $\widehat{\succ}$. If Lori ranks only s_1, she gets the lottery

$$L(s_1) = ((1,0), q_1; (0,0), 1 - q_1),$$

and similarly, if she ranks only s_2, her lottery is

$$L(s_2) = ((0,1), q_2; (0,0), 1 - q_2).$$

[12] Rees-Jones et al. (2019) induce such perfect correlation in the lab. Incidentally, they find that participants fail to fully take it into account.

[13] Other lists take a similar form, with a different order of indices. Whether Lori knows the *correct* probabilities associated with each list is of course irrelevant.

Finally, if she ranks truthfully, the corresponding lottery is

$$L\left(s_1 \widehat{\succ} s_2\right) = \left((1,0), q_1; (0,1), (1-q_1)q_2; (0,0), (1-q_1)(1-q_2)\right),$$

with an analogous expression if she flips.

2.2 Timing and Initial Beliefs

As discussed in section 1, there are three relevant periods in this setting. In period 1, Lori learns that she can participate in the match by submitting a ROL; she plans which ROL to submit; and she submits (i.e., actually chooses) her planned ROL. By including learning, planning, and choosing/submitting in the same period (the only action period), we assume away the possibility that Lori plans a choice in one period, but implements her planned choice only in a later period. Our periodization therefore rules out, by assumption, the possibility of deviating from a planned choice, and hence the possibility of dynamic inconsistencies.

In the main text, we assume as above that Lori's inherited probabilistic beliefs entering period 1 regarding period-3 consumption are of zero consumption, and therefore her updated period-1 beliefs, represented by the lottery implied by her submitted ROL, yield positive news utility. In appendix B, we allow Lori to hold different inherited beliefs entering period 1, and show that misrepresentation is still predicted in the cases we consider, although in some cases only under a narrower range of parameters. (We provide more detail in section 2.4 below.)

In period 2 uncertainty is resolved and Lori learns her match. Consumption occurs in period 3, which contains the rest of Lori's life. Using equation (2), Lori's utility from submitting a truthful ROL is therefore

$$\begin{aligned} U\left(L\left(s_1 \widehat{\succ} s_2\right)\right) = {} & 2q_1 m_1 - q_1(1-q_1)(\lambda-1)m_1 \\ & + 2(1-q_1)q_2 m_2 - (1-q_1)q_2(1-(1-q_1)q_2)(\lambda-1)m_2, \end{aligned}$$

and her utility from submitting a flipped ROL is a similar expression with the subscripts 1 and 2 flipped.

2.3 Analysis of Flipping

First, consider the case where Lori is required to include both schools on her list. Lori will choose to rank $s_2 \widehat{\succ} s_1$ iff $U\left(L\left(s_2 \widehat{\succ} s_1\right)\right) > U\left(L\left(s_1 \widehat{\succ} s_2\right)\right)$, i.e., if

$$\frac{m_1}{m_2} \equiv m < \frac{3-\lambda+(\lambda-1)q_2(2-q_1)}{3-\lambda+(\lambda-1)q_1(2-q_2)}, \tag{3}$$

(and for all $m > 1$ if only the denominator is negative).

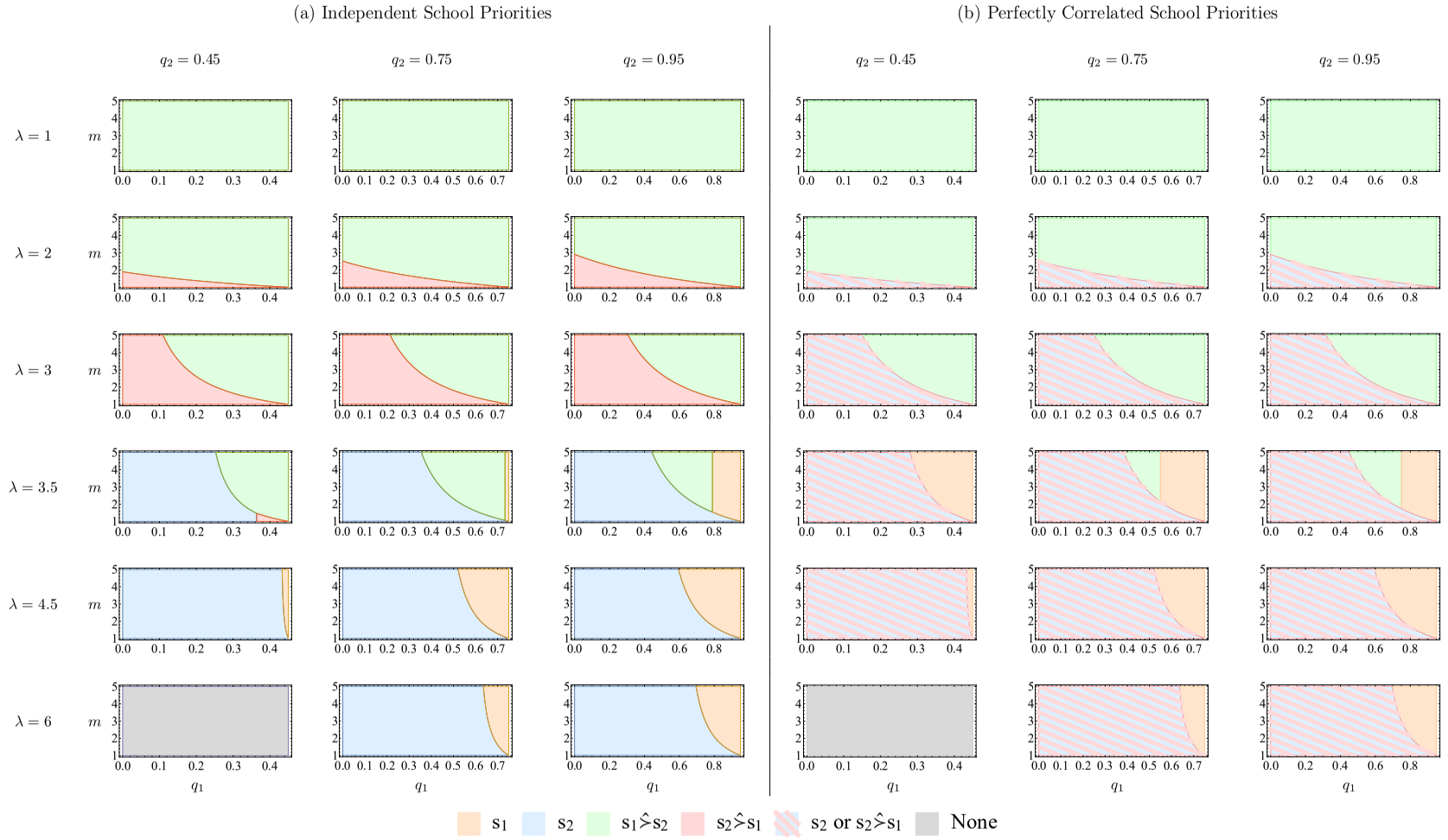
For example, if $q_2 = 1$, (3) reduces to $m < \frac{2+(\lambda-1)(1-q_1)}{2-(\lambda-1)(1-q_1)}$, which, with $\lambda = 3$ (or $\Lambda = 2$) and $q_1 = \frac{1}{2}$, predicts that Lori will misrepresent her preference as long as $m < 3$, that is, as long as the consumption utility from her preferred school is less than three times that from her less-preferred school. Of course, with no loss aversion, DA is strategy-proof: with $\lambda = 1$, the condition in (3) reduces to $m < 1$, and, since $m > 1$ by assumption, flipping never occurs.

Fixing λ, the RHS of (3) (which determines the upper bound of m such that flipping is predicted) increases in q_2 and decreases in q_1. Note that given some q_2, as $q_1 \to q_2$, the RHS of (3) approaches 1 (the no-flipping value). As q_1 decreases, the RHS of (3) increases, i.e., there is a wider range of values of m where flipping occurs. Intuitively, a decrease in q_1 creates attachment to s_2: expecting s_2 with high probability makes *not* matching with s_2 more costly, and Lori may choose to reduce this cost by increasing the probability of matching with s_2. This increase is achieved by flipping, which results in lower expected consumption utility but more certainty. As the value of q_1 approaches q_2, disappointment occurs with the same probability and the incentive to flip disappears, so Lori is predicted to rank schools by their consumption value.[14]

The top three rows of subfigures in figure 1 present graphically the above analysis (in panel a) and its perfect-correlation counterpart from the appendix (in panel b). Each subfigure presents the range of combinations of m and q_1 for which flipping is predicted for different values of q_2 and λ. As the top row shows, without loss aversion ($\lambda = 1$), no misrepresentation occurs. As λ increases to 2 (second row) and then 3 (third row) and, for a given λ, as q_2 increases and as q_1 decreases, flipping (panel a), and flipping or omitting s_1 (panel b), are predicted for increasingly wider ranges of m. (Under perfectly correlated priorities, flipping is theoretically identical to omitting s_1.)

[14] If we assume that q_1 decreases sufficiently faster than q_2, this analysis implies that, fixing preferences and loss aversion, flipping is more prevalent among those with weaker applications. For example, if s_1 is extremely selective relative to s_2 (because it is smaller or more prestigious), it is reasonable to assume that a flaw in the application (e.g. a somewhat weak letter of recommendation) hurts a candidate's chances of admission to s_1 much more than it hurts her chances of admission to s_2.

Figure 1: Theoretical Predictions



Notes: Model-predicted submitted list as a function of the model's parameters: each subplot shows the model's predictions for combinations of q_1 and m, fixing q_2 and λ. "s_1" and "s_2" denote lists containing only one school, with the other school omitted (s_1 is the preferred school). "None" denotes an empty list.

2.4 Analysis of Omissions

We can relax the complete-lists assumption and consider the case where Lori can choose to omit one (or both) of the schools. The exact conditions for each optimum are cumbersome and are relegated to appendix C. Here we outlay the general reasoning behind it. By plugging in the ROL-dependent probability, the utility u^k from submission in each consumption dimension is given by (2) above. For $\lambda \leq 3$, u^k is always positive, implying that adding a school to Lori's ROL always increases her utility. Therefore, with $\lambda \leq 3$, Lori will always rank all available schools, and allowing for omissions does not change the analysis in 2.3 above. However, if $\lambda > 3$, u^k can be negative for a sufficiently small probability and a sufficiently high λ. As we discuss in 2.7 below, this violates FOSD. Note, however, that the sign of u^k is independent of m_k. In general, whenever

$$p_k < \frac{\lambda - 3}{\lambda - 1} \equiv \text{þ}, \tag{4}$$

we have $u^k < 0$ and so on this dimension, Lori is better off when she omits the school altogether, violating FOSD. One can think of the threshold þ, which will crop up repeatedly, as the severity of loss aversion (þ$\to -\infty$ as $\lambda \to 1$; þ$\to 1$ as $\lambda \to \infty$). In appendix C we show how the optimum is found in a few simple steps.

The bottom three rows in both panels of figure 1 present the model's predictions for three values of $\lambda > 3$. They show that comparisons between a full list and a partial list depend only on the probabilities (since the sign of u^k is independent of m_k), while comparisons between full lists with s_1 versus s_2 on top depend both on the probabilities (which determine expected loss) and on m (which captures consumption preferences).

To summarize, with $\lambda \leq 3$, Lori is predicted to submit complete lists. She may show an endowment effect and flip the order of schools. This effect negatively depends on m. Intuitively, when the difference in utility from attending the two schools is small, flipping is less costly in terms of payoff, and is therefore more prevalent. Ranking two schools as opposed to one school adds another consumption dimension to utility, and therefore more potential disappointment. When $\lambda > 3$, the effect of additional expected disappointment can dominate the effect of added expected consumption utility in this dimension. Lori may find it optimal to omit a school from her list, in order to avoid future disappointment from not getting it. This effect becomes stronger for larger values of λ and smaller values of q_1 and q_2. The same intuition holds for both the uncorrelated (panel a) and the perfectly correlated (panel b) cases, with the technical caveat that in the latter, when Lori ranks s_2 on top, she is never matched with s_1 and so the second place on the list becomes irrelevant.

This entire analysis assumes that Lori enters the submission period with prior belief of

0. Appendix B shows the model's predictions assuming different prior beliefs. Specifically, we consider three possible sets of beliefs that Lori holds when entering period 1: (a) expecting to attend s_1 with probability 1; (b) expecting to attend s_2 with probability 1; and (c) expecting the lottery generated by ranking truthfully (that is, attending s_1 with probability q_1 and s_2 with probability $(1-q_1)q_2$). Relative to our zero-prior-expectations benchmark, if Lori expects, entering period 1, to attend s_2, she is more likely to misrepresent (by flipping or by omitting s_1). If she expects to attend s_1, she is less likely to misrepresent, but misrepresentation is still predicted (especially for high values of λ), mostly by dropping s_2. Last, if Lori expects to rank truthfully, she is much less likely to misrepresent, but misrepresentation (by flipping) is still predicted for high q_2, as long as q_1 and m are not too high.

2.5 Misrepresentation with More Than Two Schools

We prove that unless the conditional admission probabilities, $q_1, \ldots, q_n$, happen to weakly increase with preference, for a sufficiently small m, preference misrepresentation is strictly preferred.

Proposition. *Let $S = \{s_1, s_2, ..., s_n\}$ be a set of schools, ordered from the most to least preferred, let $\boldsymbol{q} = (q_1, ..., q_n)$ be the corresponding market conditions where $q_j \in (0,1)\ \forall j = 1, ..., n$, such that there exists at least one pair of schools s_i, s_{i+1} with $q_{i+1} > q_i$. If $\lambda > 1$ then there exist m_i, m_{i+1}, such that the optimal ranking submission is non-truthful.*

We prove the proposition in appendix D, under the two alternative assumptions regarding the correlation between priorities at different schools. The intuition behind this result is simple: if preference ordering does not coincide with probability ordering, then we can always find a pair of schools that satisfy the conditions in 2.3, and with the appropriate m find a ranking that strictly improves Lori's utility as compared to the truthful ranking.[15]

2.6 Field Evidence and Discussion

The analysis above suggests that all else equal, students with a lower belief q_1 will tend to avoid ranking s_1 on top, by flipping or—if their $\lambda > 3$—possibly also by omitting it. Using data from the DA-based Mexico City high school match, Chen and Pereyra (2019) find that among students whose survey responses strongly indicate a preference for a locally famous set of selective schools, one fifth did not top-rank any school from that set, consistent

[15] As shown in appendix B, if we consider different prior expectations, we may have to impose stronger assumptions on the market conditions. For example, if prior expectations are either s_1 for sure, or the lottery generated by ranking truthfully, then we have to assume that q_1 is sufficiently small and q_2 is sufficiently high.

with either flipping or the omission of s_1. Of these apparent misrepresentations, 23 percent are consequential, i.e., the student would have been admitted to a selective school had they top-ranked it. Notably, the strongest predictors for misrepresentation in the data are average grade from secondary school and socioeconomic background. To the extent that these are associated with a lower q_1, these findings are consistent with the model's predictions regarding omission and, assuming a sufficiently high q_2, regarding flipping as well.

As we mention in the introduction, such empirical examples, that appear consistent with the prediction that lower q_1 leads to misrepresentations, can also be explained by, e.g., extreme pessimism (in line with the explanation in Chen and Pereyra, 2019). In contrast, the model's counterpart prediction (see equation (3)), that all else equal, higher q_2 *also* leads to misrepresentation—an "attachment effect" created by high expectations—*cannot* result from pessimism. Consistent with this prediction, Grenet et al. (2019) show that in Germany's DA-based university admissions, receiving an "early offer"—in our setting, the equivalent of a signal that guarantees $q_i = 1$—strongly predicts top-ranking it, despite it not being more desirable. Specifically, receiving such a signal from a school increases the probability that a student top-ranks it by 8–11 percentage points (25–36 percent increase). (The authors explain their findings with reference to search costs and regret.)[16]

Empirically, our paper focuses on obvious misrepresentations, i.e., FOSD violations: we discuss field evidence in the introduction, draw theoretical predictions below, and analyze Li's (2017a) data in section 3. Obvious misrepresentations are easier to identify: one has to assume only that people like money. However, we opened section 2 with the more interesting case of horizontally differentiated schools because we believe that it is more important and relevant. First, it is relevant to a higher share of the population—the conditions for misrepresentation in multidimensional settings are weaker. As we show below, $\lambda > 3$ is a necessary condition for obvious misrepresentation and for misrepresentation in (money-based) lab experiments. In contrast, as we show in 2.5 above, flipping between a desirable low-probability school and a less-desirable high-probability school may occur at any $\lambda > 1$. Second, while providing a way to cleanly identify obvious misrepresentations, the unidimensional (mone-

[16]In addition to the above choice-based evidence of misrepresentation in non-monetary contexts, there is also some survey-based evidence. Rees-Jones (2018) reports that out of 558 respondents surveyed after their NRMP submission, 17 percent self-assessed their submitted ROLs as non-truthful. However, this evidence is hard to interpret because in many cases respondents' verbal explanations suggest that their ROLs *did* truly represent some other notion of preferences, e.g., their *family* preferences—and should therefore not be considered here as evidence of misrepresentation. Hassidim et al. (2018) report that in two surveys conducted in 2014 and 2015, 18 percent of 367 respondents reported submitting a ROL that was only "partially truthful" (2014), and 20 percent of 292 respondents reported that their ROL ranked some program higher or lower relative to their true preferences (2015). In the 2015 survey, where respondents provide verbal explanations for this behavior, they often mention considerations of chances of admission (which may also suggest that in this case respondents do interpret the preferences question as intended).

tary) aspect (i.e., funding) is hardly at the center of DA—monetary lab experiments and the ranking of funded vs. non-funded positions are both very special cases of DA. In general, DA is put to use in settings of multidimensional consumption, where agents' preference ordering is far from obvious. Finally, analyzing multidimensional settings is arguably more important in cases where the monetary and non-monetary lifetime consequences of not matriculating at the most preferred school can be much higher than the funding lost as a result of costly misrepresentation.

2.7 Funded Positions

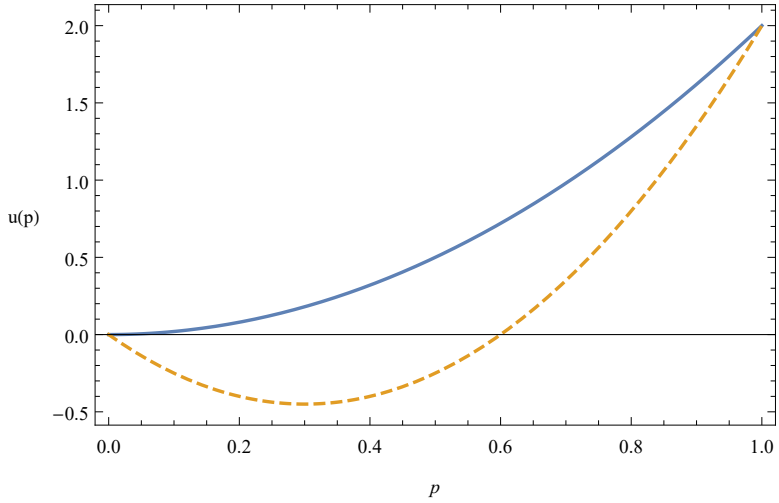
While we view the less-than-obvious misrepresentations analyzed in this section as more important in the field, they are difficult to directly identify (because the m_i's are not generally observable). We now analyze easier-to-identify, *obvious* misrepresentations—the omission or flipping of funded positions—which have been discussed in the introduction.

We again assume two schools, only one of which, s_1, offers funding. We keep all assumptions and notation as above, and denote the funded position by $\widetilde{s_1}$. The probability of getting into the funded position *conditional on getting into* s_1 is π. The amount of funding is $\$x$. Conditional on rejection from (the unfunded) s_1, the probability of admission to $\widetilde{s_1}$ (i.e., the probability of getting funding at s_1) is, naturally, 0.[17] Note that there are now three different options to rank, which we assume to represent three different consumption bundles (there are three consumption dimensions: one for each school and one for money; money can only be consumed when bundled with s_1). While this results in many more ranking options to analyze, we can show a simple sufficient condition for obvious misrepresentation. For simplicity, assume that the probabilities are such that s_1 is always ranked (i.e., we focus on the binary decision whether to rank the funded version above the non-funded version or not). Note that Lori is admitted to the funded program and gets a discount of $\$x$ with maximum probability of $q_1\pi$. Plugging in these values into (2), her utility in the money dimension in this case is $u^{\text{money}} = 2q_1\pi m^{\text{money}}(x) - q_1\pi(1-q_1\pi)(\lambda-1)m^{\text{money}}(x)$, where $m^{\text{money}}(\cdot)$ is the consumption utility from money. As noted in 2.4, with $\lambda \leq 3$ this expression is always non-negative. However, if $q_1\pi < \text{Þ}$ (which is possible only for $\lambda > 3$), Lori never ranks $\widetilde{s_1}$ above s_1, i.e., she violates FOSD: she gives up on the possibility to win a prize with positive probability.[18]

[17] Formally (using the notation from 2.1), the admission cutoff to s_1 with funding is $\hat{\kappa}_{s_1} > \kappa_{s_1}$. We denote $\pi \equiv Pr(\epsilon_{s_1} \geq \hat{\kappa}_{s_1} | \epsilon_{s_1} \geq \kappa_{s_1})$. Rejection from the unfunded position implies rejection from the funded position since $\epsilon_{s_1} < \kappa_{s_1} < \hat{\kappa}_{s_1}$.

[18] In appendix B we allow for non-zero prior expectations entering the submission period. We show that the model predicts obvious misrepresentations only when Lori is sufficiently loss averse and sufficiently pessimistic (relative to $q_1\pi$), i.e., only if her λ is high enough and her prior expectation to receive funding is low enough.

Figure 2: Upward and downward slopping utilities at $p = 0$



Note: Utility from a surprise binary lottery with winning probability p. Parameters: $m_i = 1$, $\lambda = 3$ (solid), $\lambda = 6$ (dashed).

Figure 2 presents the RHS of (2) as a function of the probability in two possible cases (upward- and downward-sloping at $p = 0$). It is easy to see that everything else equal, if u^{money} is positive, a decrease in $q_1\pi$ decreases u^{money}, possibly changing its sign. As a result, ranking the funded position is suboptimal (if the expression is negative, it will remain negative with the decrease). Put differently, a decrease in $q_1\pi$ weakly increases the number of applicants who forgo the funding opportunity in order to avoid future disappointment.

As discussed in the introduction, this comparative static is consistent with an empirical finding common to all relevant papers we know of: applicants with lower standardized test scores and therefore lower chances of being admitted to a funded position are more likely to obviously misrepresent (Hassidim et al., 2018; Shorrer and Sóvágó, 2018; Artemov et al., 2017). As shown by Shorrer and Sóvágó (2018), this correlation does not seem to be driven solely by misunderstanding or mistakes related to low ability. For example, they show that when the selectivity of some funded positions increased, more applicants flipped or omitted those positions. Interpreted as an exogenous shock to π, this suggests a causal relationship, as implied by our model.[19]

On the other hand, if her priors are overoptimistic, the model predicts that Lori will never violate FOSD by omitting the funded position (independent of her λ). Interestingly, Rees-Jones and Skowronek (2018) find that all else equal, overconfident participants are more likely to submit truthful ROLs compared with non-overconfident participants.

[19]As mentioned in the introduction, with this field evidence alone, we cannot rule out a theory in which applicants are more likely to make mistakes when those mistakes are less costly. In order to distinguish between such a theory and ours, we would need to lower $q_1\pi$ and increase x, keeping the expected utility from payoff, $q_1\pi m^{\text{money}}(x)$, constant; only our model would predict an increase in obvious misrepresentation. (Unlike this field evidence, the lab evidence we analyze in the next section does allow us to distinguish between

If there are more than two schools, some of which offer funded positions with the same amount of funding $\$x$, it is easy to see that at the optimum Lori's list will either rank *each* funded position above its non-funded counterpart, or *none* of the funded positions above its non-funded counterpart. This makes another empirical prediction that could be tested on a dataset that had pairs of funded and non-funded positions, where all funded positions offered the same amount of funding.

3 Empirical Analysis

In this section we reanalyze data on subjects' submitted ROLs from an experimental setting in which alternatives are allocated through a strategy-proof mechanism similar to DA. Of the experiments mentioned in the introduction, we chose to analyze Li (2017a), a clean and carefully conducted recent experiment. Li's experiment has a simple setup (which minimizes potential subject misunderstanding and mistakes), and its data is publicly available. It is the only data we analyzed, which also limits the generalizability of our findings. As in other experiments, the alternatives that subjects rank are different sums of money, so that intrinsic preferences are obvious. Since all alternatives are on the same consumption dimension (money), the results from the previous section do not directly apply in this setup. However, we show below that the main prediction remains essentially unchanged: sufficiently loss-averse subjects ($\lambda > 3$) with lower chances of getting the highest prizes are predicted to rank them lower (violating FOSD), in order to reduce expected future disappointment.

3.1 Experimental Design

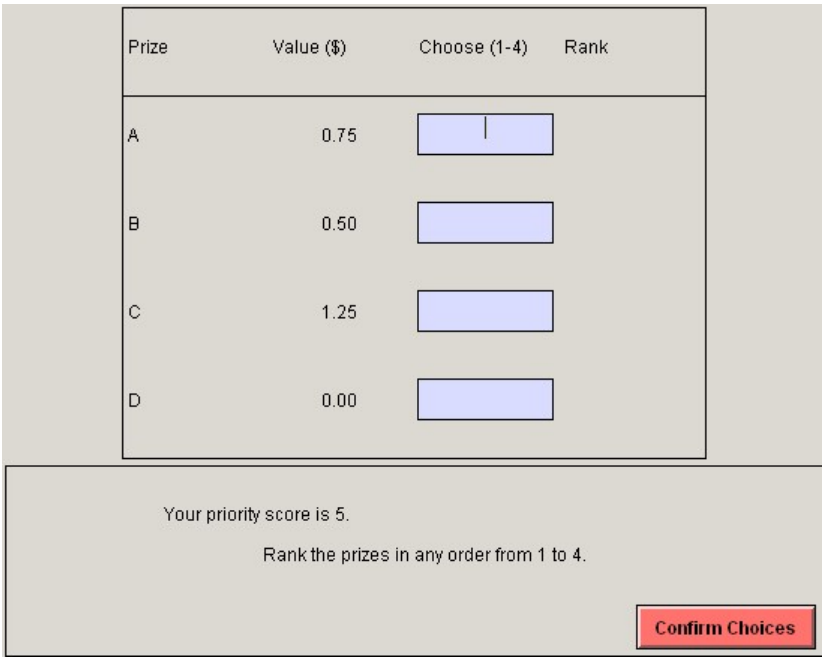
In Li's (2017a) Strategy-Proof Random Serial Dictatorship (SP-RSD) treatment, four participants play ten rounds. In each round, four prizes are allocated between the participants. The prizes are drawn uniformly and without replacement from the set {\$0.00, \$0.25, \$0.50, \$0.75, \$1.00, \$1.25}.

At the start of each round, subjects observe the four prizes, and learn their priority score: an integer drawn uniformly and with replacement from 1 to 10. Subjects' priority scores are private information. After learning their priority scores, subjects have 90 seconds to rank-order each of the four prizes. They do this by typing a different number, from 1 (top) to 4 (bottom), next to each prize. Figure 3, reproduced from Li's (2017a) instructions, shows the user interface. The mechanism is explained to participants in the instructions as follows:[20]

the two theories.)

[20]Subjects have printed copies of the instructions, and the experimenter reads them aloud just before this part begins. Li (2017a) reports that instructions are approximately at a fifth-grade reading level according

Figure 3: User Interface in Li (2017a), SP-RSD Treatment



Note: Screenshot reproduced from Li (2017a), web appendix.

> "After all the lists have been submitted, we will assign prizes using the following rule:
>
> 1. The player with the highest priority score will be assigned the top prize on his list.
> 2. The player with the second-highest priority score will be assigned the top prize on his list, among the prizes that remain.
> 3. The player with the third-highest priority score will be assigned the top prize on his list, among the prizes that remain.
> 4. The player with the lowest priority score will be assigned whatever prize remains.
>
> If two players have the same priority score, we will break the tie randomly."

(We report the full instructions page from Li (2017a) in appendix E.) This allocation mechanism, serial dictatorship, is a restriction of DA to one-sided markets with priorities (that is, markets where only one side has heterogeneous preference rankings; the other side shares a universal priority ranking).

At the end of each round, subjects learn their realized priority—i.e., whether they are first, second, etc.—and the prize they receive. The instructions neither provide recommendations on how to play nor mention the existence of a dominant strategy. Eighteen groups of four players participated in this experimental condition, amounting to a total of 72 subjects and 720 subject-round observations. See Li (2017a) for further details.

---

to the Flesch-Kincaid readability test, a standard measure for how difficult a piece of text is to read.

3.2 Theoretical Predictions

In our model, the utility from a (unidimensional) lottery $\{x_1, ..., x_4; p_1, ..., p_4\}$, with four monetary outcomes ordered from largest to smallest, is given by

$$u(\boldsymbol{x}, \boldsymbol{p}) = 2 \cdot \sum_{i=1}^{4} p_i x_i - (\lambda - 1) \sum_{i=1}^{4} p_i \sum_{j>i} p_j (x_i - x_j), \tag{5}$$

under the following assumptions. First, utility is linear in money. Second, the timing within each round is: learning and submission in period 1, resolution in period 2, and consumption in period 3.[21] Third, we assume that subjects perceive prizes earned beyond the \$20 show-up fee as a positive surprise. In the multiround data, this assumption implies that subjects do not form new expectations about subsequent rounds; in the robustness analysis below we show that our empirical findings replicate separately for earlier and later rounds, as well as in a one-shot experiment. We derive equation (5) in appendix F.

The first term in (5) is the expected consumption utility plus period-1 news utility. It equals twice the lottery's expected value (EV). The second term is the expected period-2 news utility from the resolution of the lottery. It compares, in each contingency, the realized outcome x_i against all other possible outcomes x_j, weighted by their probability. It equals $-(\lambda - 1)$ times half the lottery's *mean absolute difference* (MD), a measure of statistical dispersion also known as the *Gini mean distance* (GMD). For a loss-averse player ($\lambda > 1$), this comparison term is negative (in expectation, the resolution of uncertainty yields negative news utility). Unlike in the multidimensional-lottery case analyzed in the previous section, in this unidimensional case where all outcomes are monetary amounts, there is no endowment effect: a subject who expects to receive a small amount but wins a larger one experiences only a gain—she does not also experience a loss from not getting the smaller amount. However, for a sufficiently high λ, the negative expected news utility from the resolution of the lottery may dominate the positive expected consumption utility, and a subject may thus prefer, ex ante, to give up a positive-expected-value lottery, violating FOSD.

To calculate her vector $\boldsymbol{p}$, a subject needs to combine two pieces of information. First, she needs to calculate the probability distribution of her realized priority within her group given her (privately known) priority score. For example, if her priority score is 10 (the highest), then the probability that she gets first priority is the probability that each of the three other players either got a score lower than 10, or got a 10 and lost the random tie breaker. Table 2 presents the probability distribution of realized priorities as a function of a subject's

[21] See the discussion on timing and initial beliefs in 2.2. We analyze the discrete-background-risk case as in Kőszegi and Rabin (2009).

priority score.[22] While we do not believe that subjects actually calculate in their heads the exact probability in each cell of the table—a calculation complicated by tie breaking—we do believe that subjects have a reasonably good intuition regarding the approximate shape of the distribution. For example, with the highest priority score, it is very likely (86 percent chance) to place first, not very likely (13 percent) to place second, and extremely unlikely (1 percent) to place third of lower. With priority score of 5 or 6, it is pretty likely (74 percent) to place second or third, and less likely (26 percent) to place first or fourth. And so on.

Table 2: Probability Distribution of Realized Priorities as a Function of Priority Score

Priority Score	Probability of Realized Priority			
	First	Second	Third	Fourth
10	0.86	0.13	0.01	0.00
9	0.62	0.32	0.06	0.00
8	0.42	0.42	0.14	0.02
7	0.28	0.44	0.24	0.04
6	0.17	0.41	0.33	0.09
5	0.09	0.33	0.41	0.17
4	0.04	0.24	0.44	0.28
3	0.02	0.14	0.42	0.42
2	0.00	0.06	0.32	0.62
1	0.00	0.01	0.13	0.86

Notes: Probability distribution of realized priorities in the experiment as a function of a subject's priority score. The probabilities were estimated through simulation using 1,000,000 draws.

The second input into the vector $\boldsymbol{p}$ is a subject's beliefs regarding the ROLs submitted by the other subjects in her group. We start by assuming "face-value" beliefs: when a subject makes her ranking decision, she believes that the other subjects in her group always rank the prizes by face value. These beliefs are out of equilibrium because, as we show below, given these beliefs, the model predicts that some subjects will misrepresent (relative to face value). But they may be a natural point of departure for subjects and, importantly, an instructive benchmark that simplifies our presentation. We later replace this assumption by the assumption that subjects accurately predict the actual distribution of misrepresentation we observe. The two alternative assumptions yield qualitatively similar predictions.

We denote the truthful list by 1234, an abbreviation for 1st-2nd-3rd-4th. Under the beliefs implied by our face-value assumption—namely, that other subjects submit 1234—a subject who ranks a prize lower than its position by value, wins it with zero probability. For

[22]The probabilities were estimated through simulation using 1,000,000 draws.

example, flipping the first two prizes, which we denote by 2134, means that if the subject is placed first or second, she wins the second-highest prize. If she is placed third (fourth), then the highest two (three) prizes are already taken by the time her list is processed, and hence she receives the third-highest (lowest) prize.

Subjects are required to rank all four prizes, so there are $4! = 24$ possible rankings. However, when a subject believes that all other players play 1234, these 24 rankings are reduced to 8 sets of rankings that yield 8 different lotteries. For example, if a subject ranks the lowest prize first, her lottery is degenerate: she wins the lowest prize with probability 1. Independent of her realized priority, the lowest prize will still be available by the time her list is processed. We denote by 4XXX the set of six lists that rank the lowest prize at the top. We use the same notation for the rest of the sets, with two exceptions. First, all lists with the second prize on top and the third prize ranked above the fourth yield the same lottery, regardless of the position of the first prize. We denote this set, {2134, 2314, 2341}, by 2*3*4*. Similarly, we denote by 2*4*3* the set of lists with the second prize on top and the fourth prize ranked above the third, {2143, 2413, 2431}.

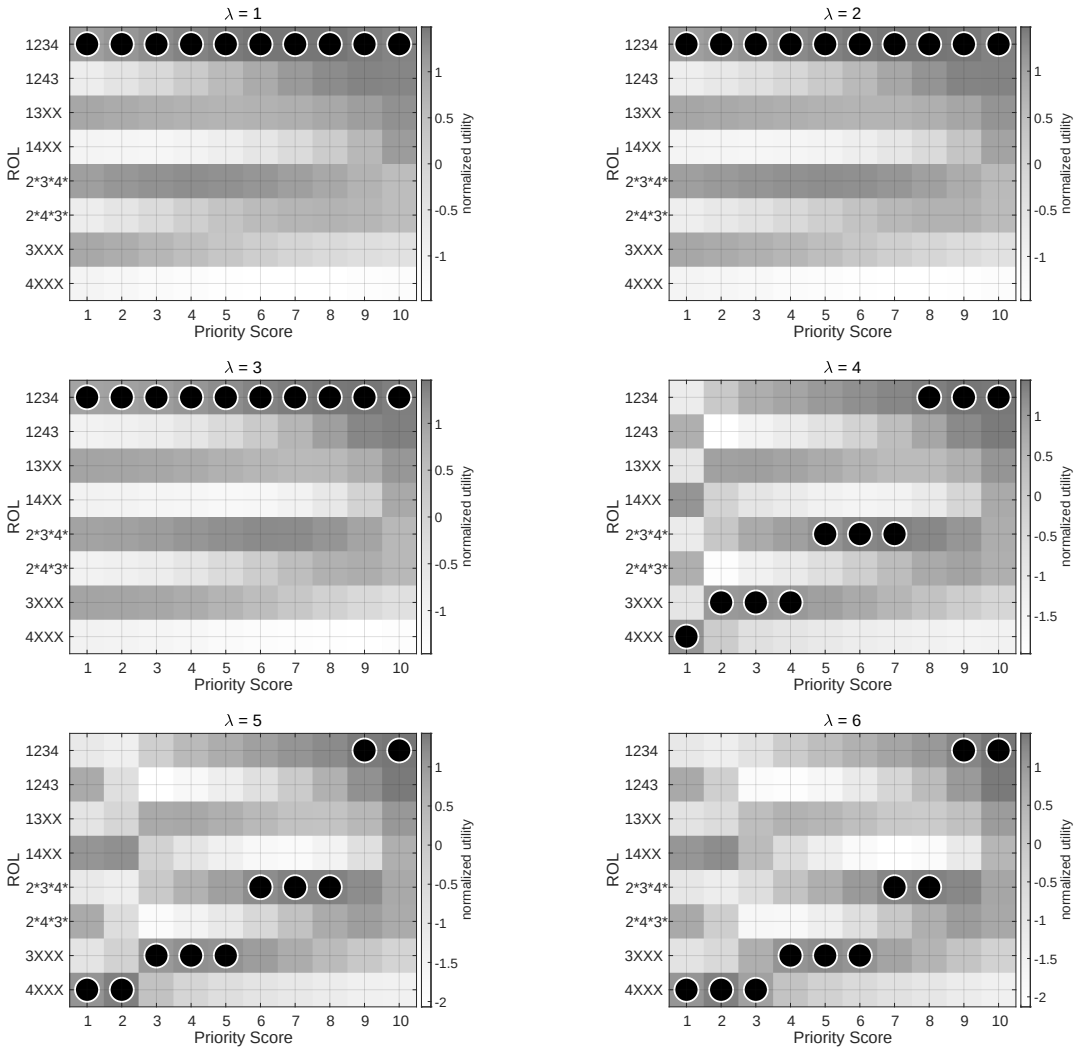
Figure 4a plots the utility from the eight ranking sets, conditional on a subject's priority score, for different values of λ. The figure is drawn for the case of prizes equally distanced from each other—the most common case—but figures based on other distance realizations are qualitatively similar (and, importantly, the econometric analysis in section 3.4 is based on the actual realization of prize distances in each round).[23] The utility in each column is normalized such that the value of each cell is in terms of within-column standard deviations from the mean. The cells marked with a black circle are the optimal (utility-maximizing) lists, given a priority score. For $\lambda = 1$, the model is reduced to the standard model with linear utility from money. Naturally, in this case 1234 maximizes expected payoff for all priority scores, reflecting strategy-proofness. However, when the priority score is the lowest (1), all lists with 3 ranked higher than 4 give approximately the same expected payoff, because the probability of being placed first or second is extremely low, so all rankings that have 3 above 4 can be easily rationalized by the standard model with a small error term. In contrast, listing 4 above 3 is expected to be costly, because it makes the probability of winning the lowest prize very close to 1 instead of around 0.86 (with around 0.14 probability of winning a higher prize). The same pattern holds for $\lambda = 2, 3$.

With $\lambda > 3$, however, the model predicts that subjects will intentionally misrepresent. Specifically, as the priority score goes down from 10 to 1, subjects will submit at the top of their ROLs values that are increasingly smaller than the highest prize. For example, a

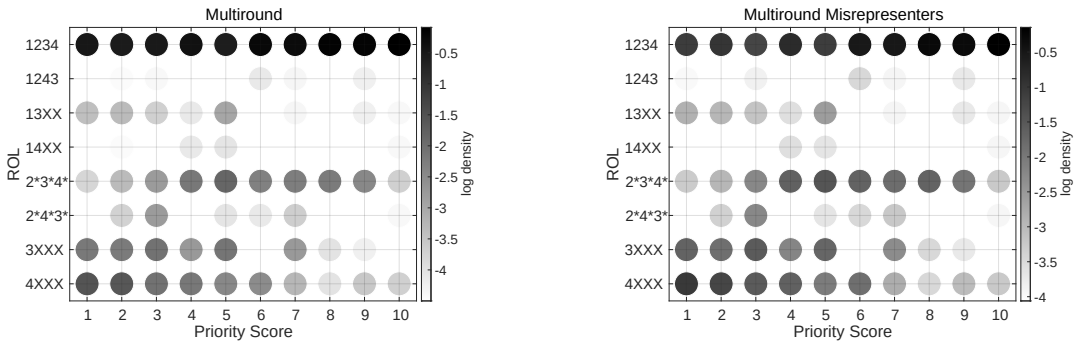
[23] Equally distanced prizes are the most common case both theoretically (20%, or 3 of the $6!/(2 \cdot 4!) = 15$ possible realizations of prize values) and empirically (17.8 percent of the observations in Li's data).

Figure 4: Preferences over Lists and Empirical Distribution of Lists (Face-Value Beliefs)

(a) Theoretical Predictions



(b) Empirical Distribution (Multiround)



Notes: 13XX = {1324, 1342}, 14XX = {1423, 1432}, 2*3*4* = {2134, 2314, 2341}, 2*4*3* = {2143, 2413, 2431}, 3XXX = {3124, 3142, 3214, 3241, 3412, 3421}, 4XXX = {4123, 4132, 4213, 4231, 4312, 4321}. Panel (a): Utility is normalized for each priority score. Distance between prizes is equal. Black circles: optimum. Panel (b): $N = 720$ (Multiround), $N = 440$ (Misrepresenters). Log density: share of observations per priority score.

subject with $\lambda = 4$ is predicted to rank truthfully when she gets a priority score 8–10, to flip the first two prizes when her score is 5–7, to rank the third-highest prize on top of her list when her score is 2–4, and to rank the lowest prize on top when she gets the lowest priority score. For higher values of λ, this effect becomes weakly stronger, and subjects are predicted to rank lower amounts on the top of their list for even higher priority scores.

It is instructive to look at the extreme case of priority score = 1. As noted above, a realized priority above third place is very unlikely in this case, less than 1 percent before rounding. If we approximate this probability to be exactly zero, and normalize the lowest prize to be zero, we get that the expected utility from ranking 4XXX is zero, and the expected utility from ranking 1234 is given by

$$2x_3p_3 - x_3p_3(1-p_3)(\lambda - 1),$$

with p_3 the probability of raking third in terms of realized priority. When $p_3 < \text{þ}$ (þ is defined in equation (4) on page 17), the second term dominates and the expression is negative. That is, the subject wants to completely eliminate the variance (and thus avoid the expected loss) and prefers 4XXX, or any list with 4 above 3, over 1234. The same intuition can be applied to all other predicted violations of FOSD: subjects may wish to increase their chances of getting their first *ranked* choice, at the expense of their expected payoff. This effect becomes stronger for subjects with higher λ.

These predictions imply that, unless all players have sufficiently low λ, it is not a Nash equilibrium (NE) for everyone to play 1234: with the belief that all other players play 1234, players with $\lambda > 3$ may find it optimal to rank lower prizes at the top of their lists. (Moreover, since such $\lambda > 3$ players are the only ones without a dominant strategy, they are the only ones whose beliefs about other players' behavior matter; it seems unlikely that these $\lambda > 3$ players would believe that all players other than themselves have $\lambda \leq 3$.) In appendix G we analyze the model's predictions under the alternative assumption that subjects' beliefs are empirically correct: subjects can predict the empirical distribution of decisions for each priority score. When making their decision, subjects therefore take into account that other subjects (especially those who get a low priority score) may submit a list that is different from 1234. To the extent that the same empirical distribution of decisions that constitutes subjects' beliefs is also *generated* by these beliefs—i.e., the same distribution of decisions is also *predicted* by the model—the data could be consistent with a NE (for a certain distribution of λ's in the population).

The appendix provides details on how, for each score-list combination, we use simulation to estimate the probability of winning each of the four prizes given the empirical distribution

of decisions. Figure G.1a there presents the theoretical predictions. They are qualitatively similar to the predictions in figure 4a above: subjects with a high-enough λ submit lists with lower prizes on top, the higher is their λ and the lower is their priority score. However, for a given λ, less misrepresentation is predicted. Intuitively, when other subjects submit non-1234 lists, the expected loss in payoff from misrepresentation is greater. In addition, while under face-value beliefs ranking a prize lower than its relative value completely eliminates the probability of winning it, under empirically correct beliefs the probability of winning the downranked prize is always positive, so flipping is less effective in reducing the dispersion of the distribution and hence in avoiding loss.

3.3 Empirical Patterns

In this descriptive subsection we analyze the empirical patterns and compare them with the theoretical predictions above. In the next subsection we move from eyeballing to econometric analysis: we formally test the fit of the model relative to the standard ($\lambda = 1$) model, and estimate the distribution of λ.

In Li's SP-RSD experiment, 36 percent of games do not end in the dominant-strategy outcome, meaning at least one player submitted an *ex-post-payoff-relevant* non-truthful list. At the subject-list level, in 29 percent of submitted ROLs, the four prizes are not listed in order of their value. Out of 72 subjects, 44 submitted a non-1234 list in at least one round. We later refer to this subsample as the *misrepresenters* (while continuing to refer to non-1234 lists as misrepresented submissions). The average loss from a misrepresented submission is \$0.21, around one-third of the average per-round earning (\$0.64). Li (2017a) is focused on testing the performance of SP mechanisms against what he defines as Obviously Strategy-Proof (OSP) mechanisms, and is therefore mostly interested in the difference in the share of games that do not end in a dominant-strategy outcome. He offers a behavioral interpretation based on cognitively limited agents who fail to perform (a specific form of) contingent reasoning and therefore fail to recognize strategies as dominant. He does not focus on analyzing the data by either priority score or patterns of misrepresentation.[24]

Hassidim et al. (2018) reanalyze Li's SP-RSD results and find that, in line with their finding from the field, priority score strongly predicts misrepresentation: a subject is more

[24]In Li's other treatment, OSP-RSD, subjects sequentially choose, in order of their realized priority, a prize from among the remaining prizes—instead of submitting a list of preferences in advance. Only 7 percent of games in that treatment do not end in the dominant-strategy outcome, statistically significantly less than the shares of games that do in SP-RSD. In that treatment, our model has the same prediction as the classical, reference-independent-preferences model: a subject is predicted to pick the highest available prize when she is called to make the decision, regardless of λ. We note that OSP and EBRD are not theoretically equivalent in general; the mechanisms under which they yield similar predictions are yet to be characterized.

than 3 times as likely not to order the prizes by their values when she receives the lowest priority score relative to when she receives the highest score.

We can take the analysis one step further. The predictions above, summarized in figure 4a (and G.1a), predict players with specific levels of λ and specific priority scores to submit specific (types of) lists. We now compare these predictions with the lists subjects actually submitted. As above, for ease of presentation, we focus on the 8 sets of lists that result in the same lottery under the face-value-beliefs assumption, and compare their empirical distribution (conditional on priority score) to the theoretical predictions. Table 3 presents the empirical distribution; the numbers are also presented graphically in 4b. The full distribution of each of the 24 ROLs is reported in appendix H.

First, in line with the result reported by Hassidim et al. (2018), the density of 1234 lists drops sharply from 91 percent for subjects with the highest priority score to 61 percent for those with the lowest score. While not strictly monotonic—possibly due to the relatively small number of observations per score—the trend is quite clear. Indeed, Hassidim et al. (2018) show that on average, a 1-point decrease in priority score increases the probability of misrepresentation by 4 percentage points (p-value < 0.0001).[25]

Moving beyond Hassidim et al.'s (2018) analysis, the table further shows that other than 1234, the most common list sets are the three predicted by the model for $\lambda > 3$ (see figures 4a and G.1a), namely 2*3*4*, 3XXX, and 4XXX. While including $15/23 = 65$ percent of non-1234 ROLs, these sets account for 82 percent of the observed misrepresentations. In comparison, the remaining $8/23 = 35$ percent of non-1234 ROLs that are not predicted by the model account for 18 percent of observed misrepresentations. Specifically, consistent with the model with these higher λ's, 2*3*4* is most common (9–16 percent) among subjects with a medium-to-high (4–9) priority score, while 3XXX, and 4XXX are most common (7–21 percent) among those with a low-to-medium (1–5) score. Importantly, comparing figure 4a to 4b and to table 3, the empirical distribution does not seem to be consistent with $\lambda = 1$ subjects making random mistakes (independent of their cost) as can be seen from the clear patterns of misrepresentations. The empirical distribution also does not seem to be consistent with $\lambda = 1$ subjects making mistakes with the probability of a mistake proportional to its cost (as we verify with a formal test in the next subsection). For example, at high priority scores (7–10), 1243 is the least costly mistake, and yet it is almost never submitted. In contrast, 4XXX is always the costliest mistake, and yet, for subjects with low priority scores (1–4) it is the second or third most common list. Similarly, 3XXX and 2*3*4* are most

[25] See their Table 5. Results are from a linear regression of a dummy for misrepresentation on priority score, with and without round and subject fixed effects, and with standard errors clustered at the subject level ($N = 720$).

Table 3: Empirical Distribution of Eight List Sets (Multiround)

ROLs	#ROLs	Priority Score				
		1	2	3	4	5
1234	1	61.1%	57.1%	58.8%	67.7%	55.2%
1243	1	1.1%	1.2%	1.3%	0.0%	0.0%
13XX	2	3.3%	3.6%	2.5%	1.6%	5.2%
14XX	2	0.0%	1.2%	0.0%	1.6%	1.7%
2*3*4*	3	2.2%	3.6%	6.3%	11.3%	15.5%
2*4*3*	3	0.0%	2.4%	6.3%	0.0%	1.7%
3XXX	6	11.1%	10.7%	12.5%	6.5%	12.1%
4XXX	6	21.1%	20.2%	12.5%	11.3%	8.6%
Total	24	100.0%	100.0%	100.0%	100.0%	100.0%
N		90	84	80	62	58
ROLs	#ROLs	Priority Score				
		6	7	8	9	10
1234	1	79.0%	74.4%	85.7%	84.3%	91.3%
1243	1	1.6%	1.3%	0.0%	1.4%	0.0%
13XX	2	0.0%	1.3%	0.0%	1.4%	1.3%
14XX	2	0.0%	0.0%	0.0%	0.0%	1.3%
2*3*4*	3	9.7%	10.3%	10.7%	8.6%	2.5%
2*4*3*	3	1.6%	2.6%	0.0%	0.0%	1.3%
3XXX	6	0.0%	6.4%	1.8%	1.4%	0.0%
4XXX	6	8.1%	3.8%	1.8%	2.9%	2.5%
Total	24	100.0%	100.0%	100.0%	100.0%	100.0%
N		62	78	56	70	80

Notes: Share of decisions as a percentage of choice situations with the same priority score. ROLs are grouped into sets as explained in figure 4a.

common for priority scores at which they are *not* the cheapest mistakes for $\lambda = 1$ subjects. (Comparing G.1b to the predictions in G.1a, the empirical patterns seem to be consistent with the theoretical predictions (for higher λ's) also with empirically correct beliefs.)

In summary, the observed submission patterns appear to be consistent with the model's predictions, which themselves appear qualitatively similar under either of our two alternative assumptions regarding subjects' beliefs about other subjects' play. Moreover, the empirical patterns seem to imply a substantial heterogeneity in λ, with a sizable fraction of subjects that are rather loss averse, i.e., have rather high λ's.

3.4 Estimation

We first present the results from a (multinomial) logit model with a single λ for all subjects, and test our model against the standard $\lambda = 1$ model. We then allow for heterogeneity in λ by estimating the logit model for individual subjects separately, and by estimating a parametric distribution of λ with a mixed-logit model.

3.4.1 Empirical Specification

Our empirical specification is based on equation (5) (on page 24). In its most general form, it assumes the following utility function for individual i with a priority score s who submits a list l at round r:

$$U_{islr} = \alpha_1 EV_{sl} + \alpha_{2i} MD_{sl} + \varepsilon_{ilr}, \tag{6}$$

where $\alpha_1 = 2$, EV_{sl} = the expected value of the (monetary) payoff from the lottery generated by a score s and a list l (corresponding to the first term in the RHS of (5)), $\alpha_{2i} = \frac{1}{2}(1 - \lambda_i)$, and MD_{sl} = mean absolute difference, which equals twice the expected resolution gain-loss utility from this lottery (the second term in (5)). Since U is defined only up to an affine transformation, normalizing $\alpha_1 = 2$ uniquely identifies λ_i.[26] We keep the assumption that subjects' beliefs regarding other subjects' play coincide with the empirical distributions (conditional on score) of actual plays; results under the alternative, face-value-beliefs assumption are similar (see appendix I). ε_{ilr} is an i.i.d. error term, distributed extreme value type I (with variance set by the $\alpha_1 = 2$ normalization above). We interpret ε_{ilr} as subjects' mistakes; they are less likely as they become larger.[27] Note that with $\lambda_i = 1\ \forall i$, we get $\alpha_{2i} = 0$, and the model reduces to the standard model where subjects maximize EV, but are allowed to

[26] Specifically, we estimate the parameters and their standard errors, then multiply them by a free parameter to make α_1 equal 2 (with the standard errors adjusted accordingly).

[27] As we explain below, this assumption, combined with empirically correct beliefs, make our model a Logit Quantal Response Equilibrium (LRQE) with a non-standard utility function.

make mistakes whose probability decreases as they become larger (and, therefore, costlier in utils).

With $\alpha_{2i} = \alpha_2 \; \forall i$, all subjects have the same coefficient of loss aversion λ, and the model is reduced to the standard logit model. In that case, the probability that a decision maker i with a priority score s chooses the list l at round r is

$$P_{islr} = Prob\left(\alpha_1 EV_{sl} + \alpha_2 MD_{sl} + \varepsilon_{ilr} > \alpha_1 EV_{sk} + \alpha_2 MD_{sk} + \varepsilon_{ikr} \; \forall k \neq l\right),$$

and, under the assumptions made on ε_{ilr}, it simplifies to the usual expression

$$P_{islr}(\alpha) = \frac{e^{\alpha' x_{sl}}}{\sum_k e^{\alpha' x_{sk}}}, \tag{7}$$

where we collectively refer to the parameters α_1 and α_2 as α, and to the lottery attributes EV and MD as x. We estimate this model using Maximum Likelihood (ML).

3.4.2 Logit Quantal Response Equilibrium

Quantal response equilibrium (QRE; McKelvey and Palfrey, 1995) is an equilibrium notion with boundedly rational agents who might make errors when choosing which pure strategy to play, but whose probability of choosing any particular strategy is positively related to the payoff from that strategy. In a logit QRE (LQRE) specification, the strategies are chosen according to the logit probability (given in (7)) with correct beliefs about other players' behavior.[28] Our estimated model can therefore be viewed as a LQRE with a non-standard utility function. We test its fit against a standard LQRE (with $\alpha_2 = \frac{1}{2}(1-\lambda) = 0$).[29]

Table 4 column 1 reports our estimates: mean $\lambda = 1.98$ (SE $= 0.18$), highly statistically significantly different from 1, with a likelihood-ratio-test (LRT) statistic suggesting that the estimated model is more than 10^7 times likelier than the $\lambda = 1$ model (p-value $<$ 0.0005). Column 3 reports expectedly higher estimates when the sample is limited to the 44 misrepresenters: mean $\lambda = 2.74$ (SE $= 0.31$), with an even higher LRT statistic. Additional LRT tests (not reported) with the alternative null $\lambda = 3$ easily reject the null for the full sample in column 1 ($p < 0.0005$) but not for the subsample of misrepresenters in column 3 ($p = 0.28$). Column 4 further limits the sample to the 36 misrepresenters who misrepresent at least once *during the last five rounds*, as decisions in earlier rounds may be noisier. As

[28] The standard LQRE model has an additional rationality parameter, which is identified only if all the parameters in the utility function are known.

[29] Our approach to estimating beliefs closely follows the *empirical-payoff approach* described in Goeree et al. (2016), which provides a consistent estimate of players' beliefs.

expected, results further strengthen: mean $\lambda = 3.14$ (SE = 0.34), with a similar LRT statistic.[30]

Table 4: Mixed and Standard Logit Specifications

	(1) Multiround All (Logit)	(2) One Shot All (Logit)	(3) Multiround Misrepresenters (Logit)	(4) Multiround Late Misrep. (Logit)	(5) Multiround Misrepresenters (Mixed Logit)
α_1(normalized)	2	2	2	2	2
	(0.09)	(0.18)	(0.14)	(0.16)	(0.13)
λ	1.98	2.18	2.74	3.14	
	(0.18)	(0.37)	(0.31)	(0.39)	
μ_λ					2.71
					(0.48)
σ_λ					2.51
					(0.41)
Log-likelihood	−1657.81	−453.95	−1165.98	−988.90	−1138.50
Likelihood ratio test	32.24	11.47	36.54	35.73	
p-value	0.000	0.001	0.000	0.000	
N	720	192	440	360	440

Notes: The parameters were estimated through Maximum Likelihood (columns 1–4) and Maximum Simulated Likelihood with 1,000 draws (column 5). Standard errors in parentheses. The probability of winning each prize is the simulated probability as described in 3.2. Likelihood ratio test: df = 1, $H_0 : \lambda = 1$.

3.5 Robustness and Heterogeneity

3.5.1 Simulation and Permutation Tests

As discussed above, the empirical distribution of submitted ROLs (table 3 and figures 4b and G.1b) shows a distinctive pattern. When describing it in section 3.3, we emphasized three independent features. First, misrepresentations account for a certain share (29 percent) of ROLs. Second, misrepresentations tend to concentrate at specific ROLs (e.g., 2134) rather

[30] We thank an anonymous referee for suggesting the specification in column 4. Recall that misrepresenters above are defined as subjects who misrepresent in at least one round. As expected, using stricter definitions results in still higher estimates. For example, using subjects who misrepresent in at least two rounds ($N = 40$) or in at least three rounds ($N = 32$), yields estimated mean $\lambda = 2.88$ (SE = 0.34) and 3.62 (SE = 0.45), respectively. Of course, these higher means provide only indirect evidence as to how many individual subjects have $\lambda > 3$ and are hence predicted to intentionally misrepresent at lower priority scores. We provide more direct evidence in section 3.5.4, where we estimate heterogenous, individual-level λ's.

than others (e.g., 3142). Third, misrepresentations tend to concentrate at lower priority scores. In this subsection we use three simulation/permutation tests, corresponding to three alternative data-generating models. The three maintain, respectively, only the first, the first and second, and the first and third above features of the data. We find that none of the three models could yield the LRT result above by chance (simulated exact p-value < 0.0001), and conclude that our strong rejection of the $\lambda = 1$ null in the previous subsection resulted from more than one of these three subsets of features.

In all three tests, we take the real dataset, and randomly assign a synthetic decision to each subject-score observation. In each test, the randomization of decisions follows a different rule that reflects an alternative data-generating behavioral model, as follows. First, we examine a naive trembling-hand model, in which subjects make randomly chosen errors, independent of their priority score. The probability of a tremble (i.e., of making an error) is taken to be the same as the empirical density of misrepresentations ($= 209/720$). To generate the data according to this model, we randomly choose 209 subject-score observations, and assign to them synthetic ROLs, chosen uniformly out of all 23 possible non-1234 ROLs; the rest of the observations are assigned the 1234 ROL. Second, we look at a more nuanced trembling-hand model in which subjects make the *same* errors they make in the real dataset. The probability of a specific tremble (e.g., 3142) is the same as its empirical density in the actual data, but, as in the previous model, trembles are independent of priority scores. To generate data according to this model, we randomly reshuffle the actually submitted ROLs across observations, breaking the pattern that relates specific priority scores to specific misrepresentations. Last, we consider a trembling-hand model in which subjects make random errors, and the probability of making each of the 23 possible errors *given a priority score* is $1/23$ of the empirical probability of submitting a non-1234 ROL at that priority score in the actual data. Datasets in this test contain the same number of (uniformly drawn) misrepresented lists per priority score as the real dataset.[31]

For each of the three alternative data-generating models, we synthesize 10,000 datasets, and apply to each dataset a LRT as in subsection 3.4.2 (we keep assuming that beliefs about others' play coincide with the empirical distribution of plays in the actual data). In each of the three tests, none of the 10,000 simulations yields a LRT statistic higher than the actual-data LRT statistic ($=32.24$) from the previous subsection, implying a simulated exact p-value < 0.0001 in each of the three tests.

[31]This model may be interpreted as if ordering prizes by value entails some cognitive cost, and therefore at low priority scores (low expected payoff), subjects are more likely to submit a randomly (and uniformly) chosen ROL. If subjects are avoiding cognitive costs by, e.g., typing 1 2 3 4 in order next to each prize (see the instructional screenshot in figure 3), it would be equivalent to drawing a ROL from a uniform distribution.

3.5.2 Earlier vs Later Rounds

Subjects in multi-round experiments may make early mistakes that they learn to avoid with experience. In order to test whether our results are driven by subjects' behavior in early rounds, we analyze the first and last five rounds separately (appendix J). Figure J.1 replicates figures 4b and G.1b for the two subsamples. While the share of misrepresentations drops slightly and insignificantly from 31 to 27 percent from the earlier to the later rounds, the figure shows that to the extent that any differences are visible, our theoretical predictions seem to fit the empirical patterns *better* in the later than in the earlier rounds. Applying the above logit model to the two subsamples (table J.1), we estimate $\lambda = 1.86$ (SE = 0.28) and $\lambda = 2.10$ (SE = 0.24), respectively, for the earlier- and later-rounds subsamples. In summary, subjects' inexperience in the earlier rounds does not drive our results.

3.5.3 One-Shot Experimental Sessions

After running the main experiment, Li (2017a) ran additional experimental sessions comparing OSP- and SP-RSD. They consist of only a one-shot serial-dictatorship game, with stakes that are 12 times higher than in the main experiment: prizes are now drawn uniformly from {\$0, \$3, \$6, \$9, \$12, \$15}. 48 groups of 4 subjects participated in this one-shot SP-RSD experimental condition, generating a total of 192 observations. As a robustness check, we estimate our model on this additional dataset. (A one-shot game also eliminates the possibility that subjects deviate from 1234 simply because playing 1234 ten times may seem odd, feel boring, or otherwise not make sense to subjects.)

Li (2017a) reports that in this treatment 40 percent of games do not end in the dominant strategy outcome—statistically identical (p-value = 0.80) to 44 percent in the first of the ten rounds in the original experiment.[32] 33 percent of the 192 subjects did not rank prizes by their values, and as a result misrepresenters lost an average of \$2.30, a little less than one-third of the average earning (\$7.30). In their re-analysis, Hassidim et al. (2018) show that, as in Li's original experiment, priority score strongly predicts misrepresentation also in this experiment.

Figure H.1 in appendix H shows the empirical distribution of ROLs in this experiment, both grouped as in subsection 3.2 and list by list; the appendix also reports a version of table 3. The results appear qualitatively similar to those from the main (ten-round) dataset. At the same time, with only around one-fourth of the observations, many more cells are empty. Interestingly, subjects with the lowest priority score, who according to the theory should

[32] In his one-shot OSP-RSD treatment, this number is 8 percent—again statistically identical (p-value = 0.18) to 17 percent in the first round in the original experiment.

misrepresent most frequently, submit 1234 in 12 out of 13 cases (92 percent). While based on few observations, this finding appears to contrast with the ten-round results, where 1234 submissions account for only 55 out of 90 lowest-score cases (61 percent).

Table 4 column 2 presents the results from estimating, on the one-shot data, the same logit model as in column 1. Mean estimated $\lambda = 2.18$ (SE = 0.37), and LRT p-value < 0.001 (against the $\lambda = 1$ null).[33] Overall, the results from this dataset are similar to those in column 1.

3.5.4 Heterogeneity

As discussed above, the empirical submission patterns summarized in figures 4b and G.1b are consistent with substantial heterogeneity in λ. To investigate this heterogeneity, we start by estimating (6) at the individual level. Specifically, we use the 10 decisions made by each of the 44 misrepresenters to estimate 44 λ_i's. For the remaining 28 subjects who submit 1234 in all 10 rounds, λ_i is not identified.[34] Towards the end of this subsection, we report full-sample statistics under different assumptions about the 28 unidentified λ_i's.

Figure 5 shows, first, a dashed-outline histogram of the 44 estimated values of λ_i (mean = 2.98, SD = 4.63, median = 3.40, 84 percent of the 44 estimates > 1, and 57 percent > 3). It shows substantial heterogeneity, with estimates ranging from -20.86 to 12.53.[35] However, the 44 individual-level estimates are inherently noisy. Moreover, they vary widely in *how* precise they are, with standard errors ranging from 1.02 to 45.46. A second, solid-outline histogram therefore weights each point estimate by the inverse of its standard error. The weighted histogram effectively eliminates the most extreme outliers, is more concentrated, and is slightly shifted to the left (mean = 2.88, SD = 2.43, median = 3.01, 80 percent of the weighted distribution > 1, and 51 percent > 3). An additional solid curve plots a kernel estimate of the weighted distribution.

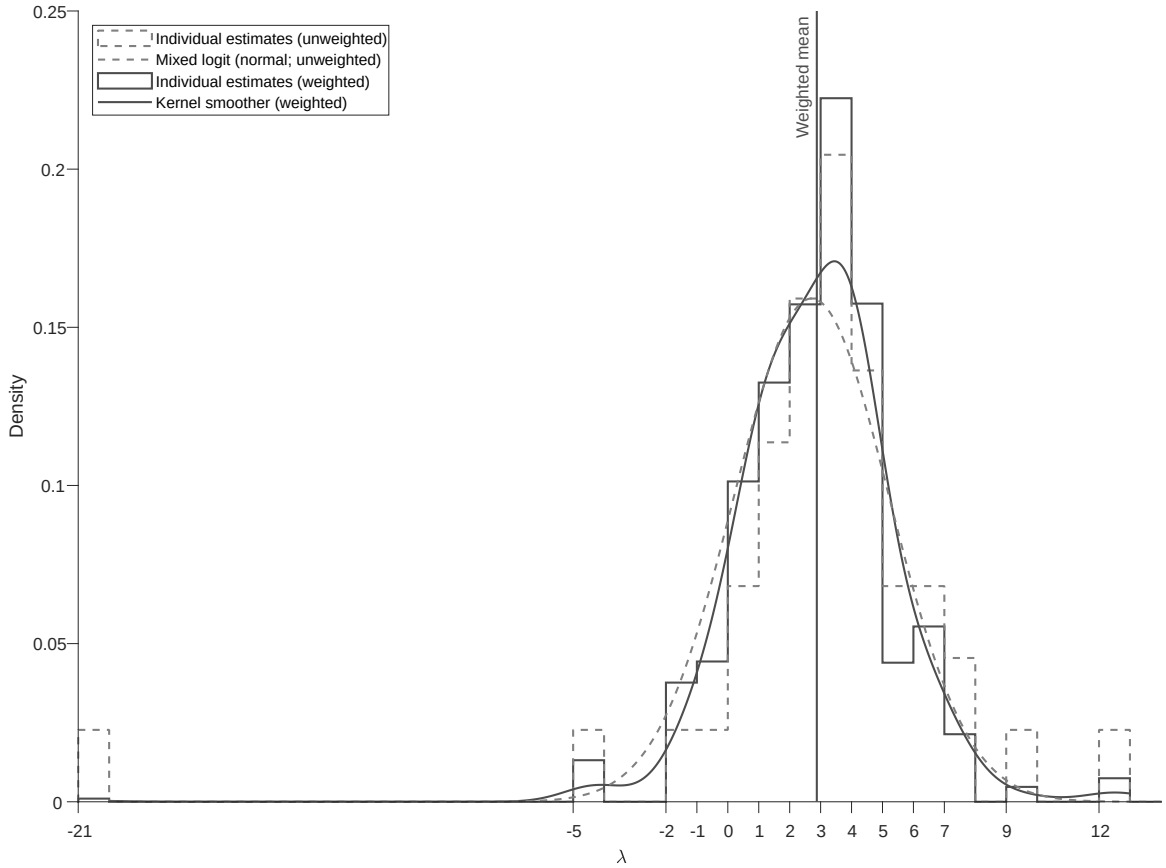
Finally, eliminating all but two degrees of freedom, the figure shows a dashed-line normal

[33] For probability beliefs, these estimates use the same simulation results as in column 1, namely those based on the main dataset (with 720 choices). However, since the distribution of choices in the one-shot dataset (with 192 choices) is rather similar, estimating beliefs from that dataset instead yields essentially the same estimates (mean $\lambda = 2.26$, SE = 0.37). Similarly, as reported in appendix I, estimating the model under the face-value-beliefs assumption yields a one-shot point estimate that is very close to the main-dataset one, albeit with larger SEs and no statistically significant LRT result.

[34] It is easy to show that for these subjects, the likelihood function is unbounded. While their λ is not point identified, it is set identified ($\lambda \in [0, 3]$).

[35] Two of the 44 subjects make choices that imply aversion to money: they rank the lowest prize first in their highest-priority-score rounds. Their individual logit estimates therefore yield negative α_1, the coefficient on EV. While such behavior cannot be explained by any model we are aware of (besides, perhaps, an extreme type of inequity aversion), for completeness we treat these estimates like all others: we mechanically normalize their $\alpha_1 = 2$ and estimate their λ_i's (= 1.99 and 7.45). Dropping them from the sample affects our results in this section only trivially (for example, the mean above changes to 2.90, and the median remains unchanged).

Figure 5: Distribution of λ (Misrepresenters Only)



Notes: Unweighted (dashed outline) and weighted (solid outline) histograms: individual-level estimates of λ_i for the 44 misrepresenters. Each observation is a point-estimate obtained by fitting equation (6) for each subject in the misrepresenters subsample, using her 10 decisions. Weights are the inverse of the SE of each point estimate. Solid curve: kernel fit of the weighted distribution (bandwidth = 1.02). Dashed curve: mixed-logit estimate with normal mixing distribution (unweighted).

distribution estimated on the raw data—the 440 choices of the 44 misrepresenters—using a mixed-logit (or random-coefficients) specification. Going back to (6), we specify a density of the distribution of α in the population of misrepresenters—a *mixing distribution*—denoted $f(\alpha|\theta)$, where θ denotes distribution parameters. The *mixed*-logit probability now takes the following form:

$$P_{islr}(\theta) = \int \left(P_{islr}(\alpha)\right) f\left(\alpha|\theta\right) d\alpha = \int \left(\frac{e^{\alpha' x_{sl}}}{\sum_k e^{\alpha' x_{sk}}}\right) f\left(\alpha|\theta\right) d\alpha,$$

where the second equality follows from equation (7). After specifying the functional form of the mixing distribution $f(\cdot)$, we estimate the parameters θ using simulation. The Maximum *Simulated* Likelihood Estimator (MSLE) is the value of θ that maximizes the simulated

likelihood function. As in the standard logit model in equation (6), we normalize the fixed coefficient α_1 to 2. We estimate (6) with MSL following the simulation and estimation methods described in Train (2009), using 1,000 draws for the simulated probabilities. Because the bell-shaped curve of the kernel estimate in figure 5 resembles a normal distribution, we choose a normal specification for $f(\cdot)$.

The (unweighted, two-parameter) normal distribution estimated with the mixed-logit model lies essentially on top of the (weighted, nonparametric) kernel distribution, and its implied summary statistics are similar to those of the weighted distribution of λ's. Table 4 column 5 reports our estimates: mean = 2.71 (SE = 0.48), SD = 2.51 (SE = 0.41). The mean λ estimate is essentially identical to the point estimate in column 3; the high statistical significance of the SD of λ implies that the null of no heterogeneity in λ is easily rejected—even among this (potentially more homogeneous) subsample of 44 misrepresenters. These mixed-logit estimates imply that 75 percent of misrepresenters have $\lambda > 1$, and 45 have $\lambda > 3$.

We can now translate these misrepresenters' statistics to full-sample statistics, under different assumptions regarding the 28 subjects who never misrepresent (and whose individual λ_i's are therefore not identified). We focus on means and medians, and consider three natural benchmark assumptions: the 28 unidentified λ_i's = 0 (the lower bound to pain from losses without loss loving), 1 (loss neutrality, i.e., no reference dependence), and 3 (the upper bound to loss aversion without FOSD violations). For the unweighted distribution, under these three assumptions, mean λ for the full 72-subjects population = 1.82, 2.21, and 2.99, respectively; median = 1.29 (assuming unidentified λ_i's ≤ 1) and 3. For the weighted distribution, assuming a mass of 28/72 at 0, 1, and 3, weighted mean = 1.76, 2.15, and 2.93; weighted median = 0.98, 1, and 3, respectively. For the estimated normal distribution of misrepresenters, assuming the same mass of 28/72 at 0, 1, and 3, mean = 1.65, 2.04, 2.82, and median = 0.43, 1, and 3, respectively.

This range of mean estimates of λ, from 1.65 to 2.99 (or, equivalently, of Λ from 1.33 to 1.99), is similar to the range of past estimates in comparable subject populations, albeit using different versions of the model (see, e.g., Gill and Prowse, 2012; Sprenger, 2015 and, more recently, Goette et al., 2018; Chapman et al., 2018; assuming immediate resolution of a surprise lottery, Kahneman and Tversky's 1979 original $\lambda^{\text{KT}} = 2$ finding would translate to $\lambda = 3$ in our formulation). However, such comparisons should be interpreted with caution. These past estimates are based on data from different settings with different strengths of the expectations manipulation. They also use different estimation methods.

3.6 Discussion

3.6.1 Misunderstanding

Throughout this section, we interpret the evidence of FOSD violations as at least partly intentional. Could this evidence instead result from experimental subjects' misunderstanding of the setting, or inattention to its implications regarding optimal behavior?

We argued above that subjects are unlikely to misunderstand Li's repeated RSD experiment. Still, subjects may be used to mechanisms that resemble the Boston mechanism ("first come first served"), and may inattentively follow common rules of thumb. They may therefore intuitively behave as they optimally would in a Boston version of RSD even if they do not misunderstand the setting. This possibility may explain why subjects aim for the prize they believe they are most likely to win, as we observe in the data. But this possibility is not supported by other features of the data. First, if subjects understand merely that each of the four players is allocated one of the four prizes, aiming for the *worst* possible outcome—i.e., choosing any ROL from the set 4XXX—is a weakly dominated choice under *any* mechanism that uses a monotone allocation rule. Yet ROLs from this set account for 10 percent of the 720 subject-round observations, and 34 percent of the 209 misrepresentations.

Second, even if subjects do not understand that each player gets a prize, and fear that unless they aim for the smallest prize they may end up with no prize at all, submitting 4XXX is still weakly dominated when the smallest prize is \$0. Yet in 520 observations the smallest prize is \$0, and in 11 percent of them (35 percent of misrepresentations in this subsample), subjects submit 4XXX. In other words, a third of the misrepresentations cannot be explained by Boston-like intuitions or rules of thumb. Third, *any* ROL in which the lowest prize is not ranked at the bottom is dominated in Boston-like mechanisms. Other than the 34 percent 4XXX submissions mentioned above, there are additional 18 percent of the misrepresentations in which the lowest prize is not ranked at the bottom. So overall, more than half of the misrepresentations are weakly dominated in any Boston-like mechanism, under *any* belief on other players' behavior.[36]

[36]More subtly and qualitatively, the most common single-ROL misrepresentation, 2134 (18 percent of misrepresentations), also seems to be a mistake under Boston-like mechanisms, given either of the two assumptions on beliefs we make in this section. In a Boston-like mechanism, after the first round, the highest prize is either never available (given face-value beliefs) or rarely available (given empirically correct beliefs). Given these beliefs and, in the case of empirically correct beliefs, given non-convex utility, 2134 is weakly dominated by 2314 (which accounts for only 6 percent of misrepresentations).

3.6.2 FOSD Violations in Other Experiments

But if the FOSD violations analyzed in this section are driven by expectations-based loss aversion rather than by mistakes or inattention of the type discussed above, should we not expect to find similar violations in other experiments?

The rejection of a small chance of winning a prize appears opposite to a central feature of Prospect Theory that the EBRD model leaves out: the idea of probability weighting. With probability weighting, a small chance is subjectively overweighted as a larger chance, and therefore should be chosen *more* than implied by its actual expected value, rather than less. What might explain why so many of the subjects have (in our interpretation) shown a dislike of small chances of gains, so much as to violate dominance? The explanation that we believe is most important is simple: people are different. As discussed in the previous subsection on heterogeneity, under arguably plausible assumptions, our findings suggest that only a *minority* of Li's subjects have the high estimated λ's required for exhibiting FOSD violations. More generally, the estimates we report are fully consistent with a majority of Li's subjects finding small-chance gains appealing. Similarly, it is possible that in experiments that focus on investigating sample averages of probability-weighting parameters, the distribution of λ is similar to what we estimate in Li's data.

Heterogeneity of experimental settings could also play a role, and could also explain a wide range of findings in the probability-weighting literature itself (Fehr-Duda and Epper, 2012; Barberis, 2013). However, the settings under which each of these apparently opposite behaviors dominates are yet to be fully understood (and modeled). At present, we can only speculate that there might be something about some settings that causes subjects to focus on, and overweight, the chance of a positive surprise, while in other settings something causes (the same or other) subjects to focus on other things, including the prospect of a potential disappointment, to the extreme of exhibiting FOSD violations.

While most economics experiments do not look for—and frequently use designs that do not allow for—FOSD violations, evidence of FOSD violations that seem to reflect intentions rather than slips has been found in past experiments. Most notably, Gneezy et al. (2006) find evidence of what they call the *uncertainty effect*: a lottery over different non-monetary prizes or deferred payments is valued less than the certainty of the lottery's worst outcome.[37] While the original finding has been challenged (Rydval et al., 2009; Keren and Willemsen, 2009) and may be sensitive to small implementation details, it is replicated by Simonsohn (2009), who also argues that it is unlikely to be caused by misunderstanding or confusion.

Most evidence for the uncertainty effect relies exclusively on between-subjects design,

[37] As we show in 3.2, the EBRD model is consistent with such effect.

in which the subjects who evaluate the certain outcomes are not those who evaluate the lottery that mixes them—as opposed to our *within*-subject evidence of FOSD violations. While Gneezy et al. (2006) report that they do not find the uncertainty effect in a within-subject design, Sonsino (2008) shows evidence of within-subject violations. Specifically, 27 percent of his subjects value a binary lottery between two gift certificates below their *own* valuation of the lowest-valued (certain) prize paid by the lottery in at least one of their fifteen choices—although they are first reminded of their valuations of the certain outcomes. This results in an overall FOSD-violation rate of 12 percent. The FOSD-violation rate is about 20 percent when the probability of winning the better outcome is 0.1–0.3, and about 4 percent when it is 0.8–0.9—consistent with the EBRD interpretation.[38] Finally, while the uncertainty-effect literature focused on FOSD violations with non-monetary prizes and deferred payments, Andreoni and Sprenger (2011) find violations with present monetary prizes. For example, using multiple price lists, 38 percent of their subjects reveal a preference for a certain monetary prize over a lottery with 95 percent of winning the same prize and 5 percent of winning a *higher* prize in at least one of three choices, with an overall violation rate of 17.5 percent.[39] Sprenger (2015) finds a violation rate of 13.5 percent in a treatment identical to that of Andreoni and Sprenger (2011), but no violations in another treatment. We are not aware of other similar experiments.

There are, however, substantial and growing literatures on both preferences over compound lotteries (e.g., Spears, 2013) and preferences over the timing of non-instrumental information about uncertainty resolution (e.g., Bellemare et al., 2005; Masatlioglu et al., 2017). Such preferences necessarily violate dominance whenever they imply a strict willingness to pay for a preferred lottery structure or resolution timing. While we do not explore these complicated topics here, we note that settings with multi-stage resolution of uncertainty—where choices sometimes resemble "complexity-induced" violations (i.e., calculation mistakes)—are also exactly the types of settings that may induce a fear of disappointment of the type that drives our predictions here.

[38] Gneezy et al. (2006) elicit hypothetical WTP for \$50 and \$100 gift certificates, and a 50/50 lottery between the two, and find that 29/30 of subjects do not violate FOSD. Sonsino (2008) elicits incentivized WTP for gift certificates for *different* objects (a weekend vacation, dinner at a gourmet restaurant, and fine chocolate or wine), and for different lotteries that mix two of the three objects, with probabilities that range from 0.1 to 0.9. The EBRD model predicts both changes to increase observed FOSD violations. When asked later to explain their violations, 34 of Sonsino's subjects (54 percent) admit having exhibited them. Of the 34 subjects who admit violations, 22 (65 percent) choose "aversion to lotteries" as their preferred explanation, while 6 choose "noise or distraction" and another 6 choose "other explanations."

[39] In this experiment, a subject reports the value of q that makes her indifferent between receiving X (with certainty), and a lottery that pays $Y > X$ with probability q and 0 otherwise, and the value of q' that makes her indifferent between the a lottery that pays X with 95 percent, and Y otherwise, and a lottery that pays Y with probability q'. Violations are identified when a subject reports $q > q'$, indicating that she prefers receiving X for sure over a lottery that pays X with 95 percent, and $Y > X$ otherwise.

4 Conclusion

Our analysis in this paper suggests that at least some of the evidence of seemingly dominated choices in DA-like mechanisms, in both the field and the lab, may in fact reflect intentional behavior by expectations-based-loss-averse individuals. This conclusion invites a reinterpretation of said evidence. It also invites a shift in how we talk about behavior in DA-like mechanisms. Rather than expressing frustration over applicants' apparent failures to understand the mechanism, we might instead ask whether we have misspecified our model of preferences.

This conclusion also invites us to rethink the advice we routinely give to applicants in DA-based matches. As mentioned in our introduction, the NRMP's website, for example, gives subjects the following advice: "To make the matching algorithm work best for you, create your rank-order list [hyphen added] in order of your *true* preferences, not how you think you will match." Such advice, however, may be intentionally ignored by loss-averse applicants. Given their concerns about disappointment, their perceived probability of matching with a program is, from these applicants' point of view, anything but irrelevant. Indeed, the evidence suggests that applicants in such mechanisms intentionally do not ignore this probability.

The above conclusion also has policy-relevant implications. For example, our analysis suggests that candidates with lower beliefs about their chances of admission into "reach" programs may avoid applying to these programs even when the mechanism is strategy-proof. As a result, the provision of more accurate matching probabilities may cause some students to apply to more selective, high-performing schools, but at the same time it may also cause other students to *avoid* these schools.

To further complicate matters, we believe that this behavior, even if intentional, may normatively be a mistake. From a welfare point of view, even if individuals intentionally make choices whose immediate consequences they understand, and even though EBRD preferences are a real component of their well-being, there is reason to believe such preferences may reflect a mistaken overweighting of the gain-loss term in the utility function. In multi-round lab experiments, narrow bracketing (Benartzi and Thaler, 1999; Rabin and Weizsäcker, 2009) may make subjects fail to realize that gains and losses may offset each other across rounds. And even when subjects do think of multiple rounds jointly, non-belief in the law of large numbers (Benjamin et al., 2016) may still mean that they under-appreciate the extent of such offsetting. We speculate that in school choice, projection bias (Loewenstein et al., 2003) in particular may cause applicants to underestimate the speed with which they will recover from disappointment. Probably most importantly, in this context of DA, it seems very likely that

people over-attend to gains and losses because they wrongly treat the immediate sensation as if it will last a long time. We emphasize, however, that even if the EBRD preferences we analyze in this paper are viewed as mistaken, the evidence suggests that their pursuit is often intentional, and not the result of people misunderstanding or miscalculating the immediate consequences of their choices.

We hope that future research will make progress on questions of stability, efficiency, and welfare analysis of different mechanisms under EBRD preferences. Ultimately, we hope that some of the relevant mechanisms may be redesigned in light of continuing analysis of the motives and preferences of people participating in those mechanisms.

References

Abdulkadiroğlu, A., Pathak, P. A., and Roth, A. E. (2005). The New York City High School Match. *American Economic Review*, 95(2):364–367.

Andreoni, J. and Sprenger, C. (2011). Uncertainty Equivalents: Testing the Limits of the Independence Axiom. Working Paper 17342, National Bureau of Economic Research.

Artemov, G., Che, Y.-K., and He, Y. (2017). Strategic 'Mistakes': Implications for Market Design Research. Working Paper.

Azevedo, E. M. and Leshno, J. D. (2016). A supply and demand framework for two-sided matching markets. *Journal of Political Economy*, 124(5):1235–1268. `https://doi.org/10.1086/687476`.

Barberis, N. (2013). The Psychology of Tail Events: Progress and Challenges. *American Economic Review*, 103(3):611–16.

Bell, D. E. (1985). Disappointment in Decision Making under Uncertainty. *Operations research*, 33(1):1–27.

Bellemare, C., Krause, M., Kröger, S., and Zhang, C. (2005). Myopic loss aversion: Information feedback vs. investment flexibility. *Economics Letters*, 87(3):319–324.

Benartzi, S. and Thaler, R. H. (1999). Risk Aversion or Myopia? Choices in Repeated Gambles and Retirement Investments. *Management Science*, 45(3):364–381.

Benjamin, D. J., Rabin, M., and Raymond, C. (2016). A model of nonbelief in the law of large numbers. *Journal of the European Economic Association*, 14(2):515–544.

Calsamiglia, C., Haeringer, G., and Klijn, F. (2010). Constrained School Choice: An Experimental Study. *American Economic Review*, 100(4):1860–74.

Chapman, J., Snowberg, E., Wang, S., and Camerer, C. (2018). Loss Attitudes in the U.S. Population: Evidence from Dynamically Optimized Sequential Experimentation (DOSE). Working Paper 25072, National Bureau of Economic Research.

Chen, L. and Pereyra, J. S. (2019). Self-selection in school choice. *Games and Economic Behavior*, 117:59 – 81.

Chen, Y. and Sönmez, T. (2006). School choice: an experimental study. *Journal of Economic theory*, 127(1):202–231.

Ding, T. and Schotter, A. (2017). Matching and chatting: An experimental study of the impact of network communication on school-matching mechanisms. *Games and Economic Behavior*, 103:94–115.

Dreyfuss, B., Heffetz, O., and Rabin, M. (2021). Data and Code for: Expectations-Based Loss Aversion May Help Explain Seemingly Dominated Choices in Strategy-Proof Mechanisms.

Fack, G., Grenet, J., and He, Y. (2019). Beyond Truth-Telling: Preference Estimation with Centralized School Choice and College Admissions. *American Economic Review*, 109(4):1486–1529.

Fehr-Duda, H. and Epper, T. (2012). Probability and Risk: Foundations and Economic Implications of Probability-Dependent Risk Preferences. *Annual Review of Economics*, 4(1):567–593.

Gale, D. and Shapley, L. S. (1962). College Admissions and the Stability of Marriage. *The American Mathematical Monthly*, 69(1):9–15.

Gill, D. and Prowse, V. (2012). A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review*, 102(1):469–503.

Gneezy, U., List, J. A., and Wu, G. (2006). The uncertainty effect: When a risky prospect is valued less than its worst possible outcome. *The Quarterly Journal of Economics*, 121(4):1283–1309.

Goeree, J., Holt, C. A., and Palfrey, T. R. (2016). *Quantal Response Equilibrium: A Stochastic Theory of Games*. Princeton University Press.

Goette, L., Graeber, T., Kellogg, A., and Sprenger, C. (2018). Heterogeneity of Loss Aversion and Expectations-Based Reference Points. Available at SSRN: https://ssrn.com/abstract=3170670.

Grenet, J., He, Y., and Kübler, D. (2019). Decentralizing Centralized Matching Markets: Implications from Early Offers in University Admissions. Working Paper.

Guillen, P. and Hakimov, R. (2018). The effectiveness of top-down advice in strategy-proof mechanisms: A field experiment. *European Economic Review*, 101:505–511.

Guillen, P. and Hing, A. (2014). Lying through their teeth: Third party advice and truth telling in a strategy proof mechanism. *European Economic Review*, 70:178–185.

Gul, F. (1991). A Theory of Disappointment Aversion. *Econometrica*, 59(3):667–686.

Hakimov, R. and Kübler, D. (2019). Experiments On Matching Markets: A Survey. Working Paper.

Hassidim, A., Romm, A., and Shorrer, R. I. (2018). 'Strategic' Behavior in a Strategy-Proof Environment. Available at SSRN: https://ssrn.com/abstract=2784659.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291.

Keren, G. and Willemsen, M. C. (2009). Decision anomalies, experimenter assumptions, and participants' comprehension: Revaluating the uncertainty effect. *Journal of Behavioral Decision Making*, 22(3):301–317.

Klijn, F., Pais, J., and Vorsatz, M. (2013). Preference intensities and risk aversion in school choice: A laboratory experiment. *Experimental Economics*, 16(1):1–22.

Kőszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Kőszegi, B. and Rabin, M. (2006). A Model of Reference-Dependent Preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Kőszegi, B. and Rabin, M. (2007). Reference-Dependent Risk Attitudes. *American Economic Review*, 97(4):1047–1073.

Kőszegi, B. and Rabin, M. (2009). Reference-Dependent Consumption Plans. *American Economic Review*, 99(3):909–36.

Li, S. (2017a). Obviously Strategy-Proof Mechanisms. *American Economic Review*, 107(11):3257–87.

Li, S. (2017b). Replication data for: Obviously Strategy-Proof Mechanisms.

Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Projection Bias in Predicting Future Utility. *The Quarterly Journal of Economics*, 118(4):1209–1248.

Loomes, G. and Sugden, R. (1986). Disappointment and Dynamic Consistency in Choice under Uncertainty. *The Review of Economic Studies*, 53(2):271–282.

Masatlioglu, Y., Orhun, A. Y., and Raymond, C. (2017). Intrinsic Information Preferences and Skewness. Available at SSRN: https://ssrn.com/abstract=3232350.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, 10(1):6–38.

Meisner, V. and von Wangenheim, J. (2019). School choice and loss aversion. Available at SSRN: https://ssrn.com/abstract=3496769.

Rabin, M. and Weizsäcker, G. (2009). Narrow Bracketing and Dominated Choices. *American Economic Review*, 99(4):1508–43.

Rees-Jones, A. (2018). Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. *Games and Economic Behavior*, 108:317–330.

Rees-Jones, A., Shorrer, R. I., and Tergiman, C. (2019). Correlation Neglect in Student-to-School Matching. Available at SSRN: https://ssrn.com/abstract=3434662.

Rees-Jones, A. and Skowronek, S. (2018). An experimental investigation of preference misrepresentation in the residency match. *Proceedings of the National Academy of Sciences*, 115(45):11471–11476.

Roth, A. E. and Peranson, E. (1999). The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design. *American Economic Review*, 89(4):748–780.

Rydval, O., Ortmann, A., Prokosheva, S., and Hertwig, R. (2009). How certain is the uncertainty effect? *Experimental Economics*, 12(4):473–487.

Shorrer, R. I. and Sóvágó, S. (2018). Obvious Mistakes in a Strategically Simple College Admissions Environment: Causes and Consequences. Available at SSRN: https://ssrn.com/abstract=2993538.

Simonsohn, U. (2009). Direct Risk Aversion: Evidence From Risky Prospects Valued Below Their Worst Outcome. *Psychological Science*, 20(6):686–692.

Sonsino, D. (2008). Disappointment aversion in internet bidding-decisions. *Theory and Decision*, 64(2-3):363–393.

Spears, D. (2013). Poverty and probability: aspiration and aversion to compound lotteries in El Salvador and India. *Experimental Economics*, 16(3):263–284.

Sprenger, C. (2015). An Endowment Effect for Risk: Experimental Tests of Stochastic Reference Points. *Journal of Political Economy*, 123(6):1456–1499.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition.