NBER WORKING PAPER SERIES

BEHAVIORAL ECONOMICS IN EDUCATION MARKET DESIGN: A FORWARD-LOOKING REVIEW

Alex Rees-Jones
Ran Shorrer

Working Paper 30973
http://www.nber.org/papers/w30973

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2023

For helpful comments, we are grateful to Eduardo Azevedo, Bnaya Dreyfuss, Pablo Guillen, Yannai Gonczarowski, Ori Heffetz, Clémence Idoux, Vincent Meisner, Tayfun Sönmez, Sándor Sóvágó, Ao Wang, and Ilan Wolff. We thank Robert Hovakimyan for research assistance.The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Behavioral Economics in Education Market Design: A Forward-Looking Review
Alex Rees-Jones and Ran Shorrer
NBER Working Paper No. 30973
February 2023
JEL No. D47,D9

ABSTRACT

The rational-choice framework for modeling matching markets has been tremendously useful in guiding the design of school-assignment systems. Despite this success, a large body of work documents deviations from the predictions of this framework that appear influenced by behavioral-economic phenomena. We review these findings and the body of behavioral theories that have been presented as possible explanations. Motivated by this literature, we lay out paths for behavioral economists to be directly useful to education market design.

Alex Rees-Jones
University of Pennsylvania
The Wharton School
Department of Business Economics and Public Policy
3733 Spruce Street
Philadelphia, PA 19104-6372
and NBER
alre@wharton.upenn.edu

Ran Shorrer
Department of Economics
The Pennsylvania State University
Kern Building
University Park, PA 16802
USA
rshorrer@gmail.com

1 Introduction

Since the emergence of behavioral economics, prominent behavioral economists have envisioned an endgame for the field in which the distinction between "behavioral" economics and "regular" economics would cease to be meaningful. Under this ideal, once the field reached maturity, foundational non-rational phenomena from behavioral economics would be widely accepted and integrated productively into the study of standard economic questions. Were this to happen, economists with no fundamental interest in psychology would find it productive to regularly attend to psychological findings from behavioral economics due to their demonstrated value in understanding and analyzing the economic question at hand.

As of the time we are writing this article, this idealized endgame is clearly not the uniform state of behavioral economics. And yet, within specific fields and for specific findings, a more modest version of this endgame has been achieved. Within public finance, for example, issues related to tax salience or tax misperceptions have been well documented and reasonably widely accepted, serving as the foundational non-rational phenomena in the template above. Across a large body of work, researchers within public finance have studied how these phenomena can be integrated into standard tax policy analysis and how their presence can significantly impact central conclusions regarding optimal tax policy.[1] As another example, within health economics, issues related to suboptimal insurance choice have been well documented and reasonably widely accepted, serving as another foundational non-rational phenomenon for the template above. Again across a large body of work, researchers within health economics have studied how this phenomenon influences our thinking about insurance markets, and in particular the insurance prices that can be sustained in equilibrium and the degree of importance of adverse selection.[2] We view applications like these as success stories of behavioral economics. In either example, a reasonably compelling case has been made that "regular" economists within these fields should be aware of some behavioral phenomena. But more importantly, beyond merely reaching a point of agreement that certain mistakes exist, researchers in these fields regularly apply frameworks that accommodate these phenomena, and their presence is at times directly relevant to the economic bottom line.

[1] For a review of this literature, see Bernheim and Taubinsky (2018).

[2] For a review of this literature, see Chandra et al. (2019).

In this article, we argue that the field of *education market design* may be poised to become a similar success story. In our view, the presence of foundational behavioral mistakes has been clearly established. Despite this foundation, we do not believe this field has yet reached the point where most members view the presence of these mistakes as central to the day-to-day conduct of their research. And indeed, we readily concede that such skepticism can be held by a reasonable reader of this literature. Until recently, very few behavioral economists directly engaged with education market design, and despite recent progress there is clearly much work left to be done to demonstrate the value of behavioral economics in this particular domain. In some sense, the jury is still out on whether the important, final stage of the successful integration process will occur: the stage in which behavioral foundations are incorporated into tractable and widely adopted frameworks for addressing the standard questions of core interest to the field. However, several lines of recent research are encouraging. We will seek to outline these lines of research and to detail the hindrances that must be overcome to achieve the potential for success that we see as possible.

When we refer to *education market design*, we refer to the practice of designing centralized clearinghouses to assign students to schools, colleges, or broader educational opportunities.[3] Under a decentralized system, students directly apply to schools to seek admission, sometimes with a local school serving as a default assignment. These decentralized systems often face problems with coordination of market timing, congestion, inefficiencies in managing capacity, and inequalities arising from ill-resourced students having low-quality default schools. Decentralized systems also typically place burdens on families to learn about the various opportunities available and navigate the potentially complex application procedures. These concerns can be greatly reduced through a well-designed centralized assignment procedure. When such a procedure is adopted, the school district may present information in a systematic and organized way, collect information from students about the schools they would like to attend, collect information from schools about students' priority for admission (when necessary), and then apply some algorithm to determine the final assignment.

In some sense, the roll-out of centralized school-choice systems has long been influenced

[3]While there are at times important distinctions between matching to schools and matching to colleges—involving, for example, whether to treat the entity as a strategic actor—when talking in generality throughout this article we will at times use the term "schools" as a catch-all term for educational programs.

by concerns of imperfect rationality. In common ways of organizing an education market, students and their families face complex strategic decisions about how to apply. For example, a common incentive in matching markets is to avoid attempting to match to unattainable options. A student who fails to heed this advice and applies to an unattainable dream school may find that their favorite attainable options reached capacity before their application was considered. This possibility is notably present in the relatively common immediate acceptance mechanism (a.k.a. the Boston mechanism). Early papers in this literature emphasize that understanding these strategic motives may be hard for students and their families. Determining an optimal application strategy requires a large amount of information and a substantial degree of strategic sophistication—two things that might not be universally held by students or their parents. Furthermore, if some students submit sincere rankings of schools even when that is not strategically in their best interest, this would contribute to sorting students to schools based on attributes beyond need or merit, often with undesirable consequences. The desire to eliminate the necessary role of strategic reasoning—largely due to concerns that it is often imperfect in complex environments—served as a primary reason that market designers advocated so strongly for *strategy-proof* mechanisms. In a strategy-proof mechanism, the weakly dominant strategy is simply to honestly report preferences, and thus attempts to misrepresent one's true preferences are not incentivized. This property is notably held by the (student-proposing) deferred acceptance mechanism[4] that has served as the default recommendation in this literature.[5]

Over the past decade, a body of work has tested the premise that preference misrepresentation is eliminated in applications of strategy-proof mechanisms. Taken together, this literature paints a clear picture that misrepresentation remains. The bulk of this evidence comes from laboratory or lab-in-field experiments, which reproducibly show that some sub-

[4]More formally, this mechanism is strategy-proof for students. If schools are strategic actors, they may have an incentive to misrepresent their preferences. Throughout this article, when we refer to "strategy-proof mechanisms" we generally mean "strategy-proof for students."

[5]As well it should be! We encourage readers not to confuse our attention to remaining imperfections that exist in markets organized by the deferred acceptance mechanism with a claim that such markets do not typically outperform alternatives in very important ways. In particular, we believe that a major advantage of strategy-proof mechanisms is that they allow the market organizer to offer simple and trustworthy advice. This has a variety of benefits even if the advice is not always followed (for discussion, see Pathak, 2016; Sönmez, 2023).

jects engage with strategy-proof mechanisms in a suboptimal manner. In these experimental markets, this problem appears to be quantitatively important and unfortunately persistent in the face of advice, training, and experience. Despite our appreciation of experimental economics, we acknowledge that we must move beyond the lab to firmly establish that these issues affect field settings of direct policy interest. Unfortunately, this is typically difficult to do in a compelling manner because identifying mistakes requires knowledge of true preferences, and true preferences are typically unobserved and heterogeneous. However, these difficulties can at times be managed, and a small set of field studies now document suboptimal reporting strategies in large-scale education markets.

This persistent presence of suboptimal preference submissions is a foundation for behavioral economics within education market design. From the point of view of the idealized rational model, preference misrepresentation in markets organized by a strategy-proof mechanism is inexplicable. Behavioral economics offers a wide variety of potential explanations for this behavior. Across recent papers, candidate theories include failures of contingent reasoning, application of improper heuristics, misunderstanding of continuation value, report-dependent preferences, expectations-based reference dependence, incompletely constructed preferences, overconfidence, correlation neglect, and more. The behavioral market design literature has not settled on any one of these options as being "the" preferred option to replace the fully rational model in this domain, and indeed the presence of many competing models is a recurring difficulty of the behavioral-economic literature. However, should some of these models be shown to reproducibly outperform the fully rational default in the settings we directly care about, then they could be integrated into our theoretical and structural analyses.

As we have emphasized, the success of behavioral economics in this area will ultimately be measured by the value of the explanations it provides. In what ways might attending to behavioral economics productively influence the activities of mainstream economists in the education market design literature? We highlight four ways that we view as promising.

First, attention to behavioral economics may offer useful guidance on how to estimate preferences in these environments. Several recent papers offer concrete examples of ways in which this can be helpful. One productive line of research has explored how standard

estimation exercises can be conducted while weakening the assumption that preferences are submitted fully optimally, for example by replacing that assumption with the less restrictive assumption that any deviation from truthful reporting does not affect the stability of the final match. Another productive path forward involves exploring how the imposition of behavioral-economic assumptions can make an otherwise intractable problem tractable. Across examples like these, we see promise that the extra complications of behavioral-economic models can often be accommodated without large advances to current methods, and at times their accommodation can actually make estimation easier.

Second, attention to behavioral economics may inform our choice of mechanisms. Recent research has made progress on modeling features of mechanisms that make them comprehensible or that invite preference misrepresentation. Such models can introduce new tradeoffs into mechanism selection considerations, at times leading to the recommendation of otherwise disfavored market designs. Even in situations where optimizing for comprehensibility is not the goal, integrating behavioral-economic theories into structural frameworks may productively influence the favored mechanisms in counterfactual analyses.

Third, attention to behavioral economics may help guide the design of interventions in existing markets. We discuss several classes of interventions that have demonstrated value in the field, and how they relate to the behavioral-economic theories we have discussed.

Finally, attention to behavioral economics may influence our understanding of the process of mechanism adoption. While most applications of behavioral economics in education market design concern the behavior of applicants in a mechanism, a more recent literature has begun to explore the behavioral-economic factors influencing preferences over different mechanisms. Understanding these factors can help market designers predict when market organizers may choose mechanisms ill-suited for their goals. This understanding can also guide market designers to propose effective interventions that will be appreciated as such by those who must enact them.

We emphasize from the start that we do not mean to suggest that the papers we discuss are the only behavioral-economics-informed papers in this literature, nor that the productive paths forward that we consider are the only ones possible. Rather, our goal is to focus attention on the paths that appear clear to us and readily justifiable by current literature,

while appreciating that other productive work and pathways are possible. For excellent reviews of the behavioral and experimental market-design literature, with some emphasis on other paths forward and different emphasis in their coverage, see Chen et al. (2021) or Hakimov and Kübler (2021).

The remainder of the paper proceeds as follows. Section 2 discusses foundational evidence of suboptimal behavior from education market design. Section 3 presents a variety of behavioral-economic theories that help explain this evidence. Section 4 presents our summaries of paths for behavioral economics to influence the conduct of education market design. Section 5 concludes.

2 Foundations for Behavioral Economics in Education Market Design

2.1 Background: Initial Concerns about Strategic Sophistication

The literature on education market design builds from the long-standing literature on two-sided matching. This literature was initiated by Gale and Shapley (1962), who considered the problems of matching students to colleges in a stable manner.[6] In their model, they determined that a stable match was always possible, and constructively proved so by providing an algorithm that produced it: the deferred acceptance algorithm. While strategic (noncooperative) incentives in the deferred acceptance mechanism were not initially considered in Gale and Shapley (1962), they were later examined in the work of Dubins and Freedman (1981) and Roth (1982). These papers established that the student-proposing deferred acceptance mechanism is *strategy-proof* for students. Formally, this means that students engaging with this mechanism would weakly prefer their assignment from truthfully submitting their preferences to their assignment under any other "strategic" preference submission (holding

[6] A matching is *blocked* if there exists a student-college pair where the student prefers the college to her assignment and the college has free capacity or it prefers the student to at least one of the students it is assigned. A matching is *stable* if it is individually rational and not blocked. In the related context of priority-based school assignment, the situation where Alice envies the assignment of Bob to a school where Alice has higher priority is termed *justified envy* (Abdulkadiroğlu and Sönmez, 2003). Stable matchings eliminate justified envy.

fixed the submissions of other market participants).

The practical connections between this theoretical literature and the design of real education markets became starkly apparent shortly after these incentive issues were first documented. In a clear demonstration of this connection, Roth (1984) used this theoretical literature to help make sense of the interesting history of the system for matching medical students to internships and residencies. In the decades leading up to a centralizing reform in 1951, the decentralized procedure used in this market faced *market unraveling* (Roth and Xing, 1994). In attempts to secure the best students before they had been captured by competitors, internships filled their positions earlier and earlier with clearly deleterious effects. Desire to correct these problems led the various actors in this market to coordinate on the use of a centralized procedure. The initially announced procedure quickly drew criticism for the strategic incentives it presented. As noted in Roth (1984), some parties quickly recognized that "a student might suffer if he ... gave high rank to hospitals he preferred but had little chance of being matched with." This motivated changes to the procedure that were purported to resolve the issue. Through their trial and error, the administrators had essentially independently developed the deferred acceptance mechanism a decade before its presentation in Gale and Shapley (1962). However, it had unfortunately been implemented in the *school proposing*, rather than *student proposing*, configuration. As a result, the market was not strategy-proof for students despite initial claims to the contrary. Seeking to correct this issue (along with several others), the administrators of this system engaged with Roth in a redesign. This redesign, documented in Roth and Peranson (1999), came to serve as a focal case study of the potential value of matching theory to directly inform real-world markets.

Very shortly after this public demonstration of matching theory's value in the medical match, potential for similar value became apparent in policies regarding school choice. "School Choice: A Mechanism Design Approach" (Abdulkadiroğlu and Sönmez, 2003) used the tools of this literature to analyze student assignment systems. This analysis laid bare some of the shortcomings of existing systems in US cities.[7] Again, a central concern was the

[7]This comparatively expansive analysis of Abdulkadiroğlu and Sönmez (2003) built from the earlier analysis of Balinski and Sönmez (1999), which studied a particular mechanism applied in Turkish college admissions and illustrated its shortcomings compared to the student-proposing deferred acceptance mechanism.

potential need for strategic reasoning.[8] Considering the immediate acceptance mechanism (which at the time was employed in Boston, Minneapolis, Lee County of Florida, and other school districts), Abdulkadiroğlu and Sönmez noted that the "assignment mechanism gives very strong incentives to students and their parents to misrepresent their preferences by improving ranks of those schools for which they have high priority." They noted that the "difficult task of finding optimal admissions strategies" may be avoided by the deployment of either the deferred acceptance mechanism or the (also strategy-proof) top trading cycles mechanism (inspired by Shapley and Scarf, 1974). Publicization of this literature and these findings contributed to a wave of collaborations between school districts and economists. Successful collaborations in Boston (Abdulkadiroğlu et al., 2005b) and New York City (Abdulkadiroğlu et al., 2005a) served as initial focal case studies.

As researchers began to take a hand in the design of student assignment mechanisms, further concerns about the consequences of strategic complexity came to the forefront. In their analysis of preference submissions under Boston's implementation of the manipulable immediate acceptance mechanism, Abdulkadiroğlu et al. (2006) found evidence of heterogeneous strategic sophistication across families. Some families clearly attended to the availability of seats across different schools, while others appeared not to. Failure to account for seat availability when submitting preferences was costly: many unassigned students could have been assigned with a more strategic preference submission. This introduced a new concern about procedures that require strategic sophistication: not only is the requirement burdensome, but its burdens fall unequally on families depending on their degree of strategic savvy. This introduces equity considerations on its own, and those equity considerations could be made worse if strategic savvy were more common among more privileged families. These ideas were further advanced in the theoretical work of Pathak and Sönmez (2008), who formalized the notion of strategy-proof mechanisms "leveling the playing field."[9]

[8]Concerns about strategic reasoning are not the only ones emphasized in Abdulkadiroğlu and Sönmez (2003). Building on the earlier Balinski and Sönmez (1999), Abdulkadiroğlu and Sönmez (2003) emphasized a central role for normative, fairness-based considerations in selecting a school-choice mechanism. This provides another strong rationale for the deferred acceptance mechanism in these contexts: it Pareto dominates any other mechanism that eliminates justified envy (a natural fairness criterion).

[9]The potential for strategy-proof mechanisms to level the playing field was further supported in lab experiments. In experimental matching markets (following the template described in the next section), Basteck and Mantovani (2018) found that participants of low cognitive ability were disproportionately matched to

While this body of work put forth strong arguments in favor of using strategy-proof mechanisms, a major conceptual hurdle remained for directly applying this line of reasoning: preference-list length constraints rendered many applications of strategy-proof algorithms not truly strategy-proof. To illustrate with an example, consider again the medical residency match. In 2022, there were over 5,000 first-year post-graduate programs participating in this system. Literal application of a strategy-proof algorithm would require applicants to form rank-order lists of all programs that they prefer to remaining unassigned. However, in the current system, applicants must pay additional fees to submit lists with more than 20 programs, and listing more than 300 programs is not allowed.[10] This introduces strategic considerations for a medical student who truly prefers a large number of residencies to remaining unmatched. Constraints like these are common in applications of (otherwise) strategy-proof mechanisms in education markets.

While list-length constraints complicate the discussion of strategy-proofness, much of the logic of the discussion so far holds in their presence. Lists shorter than the constraint are weakly dominated unless they consist of all schools that the applicant prefers to the outside option, and misordering any schools on the preference list remains a weakly dominated strategy (Haeringer and Klijn, 2009). Thus, while truthfully revealing one's full preference ordering is no longer always incentivized (or possible), truthfully revealing one's preference ordering over all listed programs remains incentivized, and the presence of constraints is irrelevant when they do not bind. Going forward in this literature, when discussing "strategy-proofness" or the prediction of truthful preference reporting, this more narrow notion of "strategy-proofness" or truthfulness is often what is meant. Of course, when the constraint on list size is binding, there are incentives to be strategic about which schools are omitted from this list. For example, in the absence of strong outside options, there are incentives to include a "safety school" at the cost of a more desirable option. More broadly, the student faces an optimal portfolio problem as in the models of Chade and Smith (2006) and Ali and

low quality schools when matched by the immediate acceptance mechanism, with this difference mitigated under the deferred acceptance mechanism. Basteck and Mantovani (2022) replicated the finding of an unlevel playing field under immediate acceptance and explored some means of partially reducing the gap.

[10] In practice, few applicants rank more than 20 programs. These shorter preference lists are facilitated by an interview process that occurs before preferences are submitted. See Echenique et al. (2022) for a discussion of these issues.

Shorrer (2022).

Despite the introduction of some scope for gaming through strategic decisions of which programs to rank, there is still a sense in which constrained deferred acceptance is less manipulable than common algorithms that give undue weight to reported top choices. This idea was formalized in Pathak and Sönmez (2013), who developed these results in their examinations of school reforms that occurred in Chicago and England. These reforms did not arise through the direct involvement of economists or market designers, and yet reflected very similar reasoning to the prior reforms that had been conducted with that involvement. In England, concerns surrounding the complexity of optimal strategies led to the 2007 ban on the use of "first preference first" procedures (which include the immediate acceptance algorithm). As a result, school districts throughout the country transitioned to variants of (constrained) deferred acceptance. Similarly, in Chicago's selective high-school match, problems with high-scoring students remaining unmatched due to unstrategic preference submissions led to a change of the mechanism midway through the 2009 match. They abandoned the immediate acceptance mechanism for (constrained) serial dictatorship, another mechanism that would be strategy-proof without constraints. As Pathak and Sönmez document, both of these reforms involved changes that reduced the scope for profitable preference manipulation.

As this brief history illustrates, concerns about families' ability to execute complex optimal reporting strategies were central in the early history of education market design. These concerns were among the primary reasons cited for recommending strategy-proof mechanisms in education applications. Following this recommendation and encouraged by the initial successful reforms, a great many school districts modified their practices, leading to the current era in which applications of strategy-proof mechanisms are significantly more commonplace.

2.2 Preference Misrepresentation in Strategy-Proof Mechanisms

In the wake of these initial studies, a number of researchers became interested in whether strategy-proof mechanisms truly eliminate preference misrepresentation. On the one hand, under a strategy-proof mechanism, the optimal reporting strategy of truthtelling is as simple as can be. If families understand that and follow this simple strategy, further attention to

the drivers of reporting behavior would be of limited value. This line of reasoning is apparent in many of the subsequent papers in this literature. On the other hand, if problems persist—that is, if families continue to apparently misrepresent preferences despite an incentive not to do so—further study of the drivers of reporting behavior could be necessary and productive. A relatively large literature in experimental economics, and a smaller literature studying field applications of school-choice mechanisms, arose to test for that possibility. We discuss each literature separately below.

2.2.1 Lab Evidence

Modern practices for using lab experiments to test for preference manipulation in school-choice mechanisms were largely initiated by Chen and Sönmez (2006).[11] This experiment was run as part of a persuasive appeal to the Boston public school system to abandon their use of the manipulable immediate acceptance mechanism and replace it with a strategy-proof alternative. The authors sought to compare the performance of three focal mechanisms (deferred acceptance, immediate acceptance, and top trading cycles) according to their success in eliciting true preferences and the efficiency of their final assignments. Making these comparisons requires knowing subjects' true preferences, which cannot necessarily be inferred from reported preferences when subjects face incentives to manipulate their reports or when subjects might report suboptimally. This motivated their decision to deploy a lab experiment, and the key element of their experimental design: directly assigning preferences to subjects so they are known to the analyst.

In Chen and Sönmez (2006), and in the later papers of this literature that generally adopt similar designs, the key pieces of the experiment are as follows. First, experimental participants take the role of students in a matching market and the experimenters communicate their preferences over school assignments. These preferences are set by offering different payments for the experiment depending on the school where the participant is ultimately

[11] While Chen and Sönmez (2006) served as the template for much of this literature going forward, we do not mean to suggest that matching lab experiments did not predate it. Experimental tests of Gale and Shapley (1962) date back at least to Harrison and McCabe (1996). In additional related examples, Kagel and Roth (2000) used a lab experiment to study congestion and unraveling in an experimental analog to medical student matches, and Roth (2002) forcefully argued for the value of lab experimentation as a tool of market design.

admitted. For example, in the "designed environment" of Chen and Sönmez (2006), the market consisted of seven schools. The participant would receive \$16 for matching to her first choice, \$13 for matching to her second choice, \$11 for matching to her third choice, \$9 for matching to her fourth choice, \$7 for matching to her fifth choice, \$5 for matching to her sixth choice, and \$2 for matching to her last choice. Under the assumption that more money is preferred to less, this dictates students' preferences, and thus submitting a ranking of schools out of monetary order would be evidence of preference misrepresentation.[12] Second, the experimenters communicate to subjects how their priority for admission will be determined. Chen and Sönmez assigned priority by assigning subjects to the prioritized "walk zones" of different schools; at times in later experiments (typically geared towards college choice), this was replaced with priority determined by test scores. Third, the experimenters communicate to subjects how assignments will be made based on the submitted preferences and priorities. Chen and Sönmez did this by presenting step-by-step explanations of the operation of the immediate acceptance, deferred acceptance, or top trading cycles mechanisms (depending on experimental condition).

Focusing on their analysis of preference misrepresentation, Chen and Sönmez's results are straightforward. Across two environments considered,[13] the rate of truthful preference reporting under immediate acceptance was 14-28 percent, whereas under deferred acceptance it was 56-72 percent and under top trading cycles it was 43-50 percent. These results support the claim that strategy-proof mechanisms lead to much higher rates of truthful preference reporting. They also suggest greater success in eliciting true preferences with the deferred acceptance mechanism as opposed to the top trading cycles mechanism (at least in this environment). It is worth emphasis, however, that despite showing that preference misrepresentation declines under strategy-proof mechanisms, this experiment does not establish that it is eliminated. Across the two environments considered and the two strategy-proof

[12]Note that some later behavioral papers present theories under which a chance at more money is not necessarily preferred to a chance for less, for example due to a preference to manage one's expectations by not requesting low-probability high-payoff options (see, e.g., Dreyfuss et al., 2022b). In such a model, ranking schools in a different order than their monetary payoffs is not necessarily a mistake, and some interpretations of the phrase "preference misrepresenation" would be inappropriate. We return to these theories in Section 3.2.

[13]These environments differed by whether the payoffs were those reported above (the "designed environment") or were randomly generated (the "random environment").

mechanisms deployed, the rate of nontruthful reporting ranged from 28 to 57 percent—an improvement over the rates seen under the immediate acceptance mechanism, but still far from perfect. Perhaps justifiably so, because perfect compliance with incentives may be unreasonable to expect in an experiment presenting novel and complex matching procedures that must be learned quickly with lower incentives than exist in the field.[14] Perhaps due to these potential justifications, the initial discussion of this research focused on its demonstration that truthful reporting became more common, and thus moved in the right direction. Later research would further assess the remaining nontruthful reporting that exists in this and related paradigms.

The general paradigm of Chen and Sönmez soon became a testing ground for assessing the role of various considerations in education market design. For example, Pais and Pintér (2008) studied how truthtelling (and other) outcomes changed as the amount of information about priorities and other students' preferences was varied. While such information has a clear role in experimental behavior, the general pattern of greater but imperfect truthtelling under strategy-proof alternatives persisted.[15] In another related example, Calsamiglia et al. (2010) studied the effect of preference-list length constraints. Again, the results confirm that in unconstrained markets, greater but imperfect truthtelling occurs in strategy-proof mechanisms. Interestingly, under constraints, subjects were even more likely to preserve the ordinal ranking of listed options, but again did not do so universally. Featherstone and Niederle (2016) tested whether comparable rates of truthtelling could be sustained in a special environment where it serves as an ordinal Bayes-Nash equilibrium strategy for participants in the immediate acceptance mechanism. They found that it could be, and in the course of their study provided further evidence on imperfect truthtelling for participants in the deferred acceptance mechanism.

As the use of this experimental paradigm grew, some researchers began directing more at-

[14]Supporting this idea, the working paper version of Chen and Sönmez (2006) speculated that the efficiency benefits arising from these strategy-proof mechanisms "are likely to be more profound when parents are educated about the incentive compatibility of these mechanisms."

[15]Pais et al. (2011) study similar information issues in a broader setting. While they find higher levels of truthtelling than in Pais and Pintér (2008), they continue to find remaining nontruthful reporting even in strategy-proof mechanisms. Chen et al. (2016) redeploy the Chen and Sönmez (2006) paradigm with full information on others' preferences. They again find greater, but non-universal, truthtelling among the strategy-proof mechanisms (this time with the top trading cycles mechanism performing comparatively well).

tention to the remaining misrepresentation that appeared even in strategy-proof mechanisms. While many researchers contributed to this discussion, these ideas were most persistently pursued in a sequence of papers by Pablo Guillen and coauthors. In Guillen and Hing (2014), the authors questioned whether the residual preference misrepresentation that is observed in the top trading cycles mechanism is the result of "a minority of confused participants" versus a more general and systematic problem with strategic sophistication. If participants do not deeply understand strategy-proofness, then preference reporting "could be influenced by previously ignored environmental factors." The specific environmental factor considered in Guillen and Hing (2014) was third-party advice. The authors found that, when advice on the optimal strategy was presented, the rate of truthful reporting *declined*. This was true regardless of whether the presented advice was correct or incorrect. In a similar vein, Guillen and Hakimov (2017) showed that truthful preference reporting under the top trading cycles mechanism declined when subjects received information that other participants were misrepresenting their preferences—a factor that is not relevant from the perspective of the rational model. Findings like these could arise if subjects' truthful preference reporting was not based on their ability to deduce their incentives in a given mechanism, but instead on quick inferences about desirable behavior drawn from whatever cues are present. Guillen and Hakimov (2018) tested this idea when using the top trading cycles mechanism for an in-class topic-allocation task. Across three treatments, they varied whether they gave simple and straightforward advice that the optimal strategy is truthtelling, further description of the functioning of the mechanism, or both. The authors found that clearly advising students to tell the truth had a positive effect on the rate of truthtelling,[16] but further description of the mechanism reduced the rate of truthtelling (even when paired with advice on the optimal strategy).[17]

This line of reasoning in some sense culminated in Guillen and Veszteg (2021). This experiment was designed to challenge the idea that the relatively high rates of truthtelling seen

[16]This finding may appear slightly in conflict with Guillen and Hing's finding that even true advice reduced the rate of truthtelling. Relative to the true advice given in Guillen and Hing (2014), the advice on optimal strategies provided in Guillen and Hakimov (2018) more directly explains that truthtelling is weakly dominant.

[17]The role of advice was further explored in Ding and Schotter (2017), Ding and Schotter (2019), and Koutout et al. (2021).

in this experimental literature arose due to subjects' understanding of strategy-proofness, per se, as opposed to a tendency to submit the induced preferences as a low-effort default strategy. To make this point, the authors introduced two variants of existing mechanisms: the reverse top trading cycles and the reverse deferred acceptance mechanisms. These function nearly identically to their namesakes, but the dominant strategy is to report school preferences in *reverse* order from true preferences. The authors argued that, if optimal use of these mechanisms were truly driven by a deep understanding of incentives combined with simple optimal strategies, the rate of optimal use should not differ across the standard and reversed versions of these mechanisms. Contrary to that expectation, the authors found that only 16% and 26% of subjects played the dominant strategy in reverse deferred acceptance and reverse top trading cycles, respectively, as compared to 68% and 46% in the standard versions of these mechanisms.

Taken together, this body of work supports the idea that experimental subjects' understanding of incentives in these environments is tenuous. While these experimental tests reproducibly show that rates of truthful reporting increase when a strategy-proof mechanism is adopted, they do not provide strong evidence that the role of strategic sophistication is eliminated.

Of course, the problems with subjects' understanding of incentives in these experimental tasks could be more severe than the problems with understanding that exist in real-world matches. Compared to the subjects in these experiments, participants in field applications will often have greater incentives to learn about the mechanism and their optimal strategy. They will also often have more time and more materials available to help them learn, and potentially more trusted sources of advice.

To test whether preference misrepresentation is eliminated under such conditions, Rees-Jones and Skowronek (2018) deployed an experimental matching task to medical students who had just participated in the medical match discussed in Section 2.1, the National Resident Matching Program (NRMP). In this task, subjects had to submit their preferences over five potential residencies, offering payoffs of \$50, \$25, \$15, \$10, or \$7.5 conditional on assignment. Final assignments would be determined by applying the deferred acceptance mechanism to these reported preferences, using randomly generated test scores for priority.

The fact that this experimental task involved the same mechanism as the medical match was a salient component of the instructions. This connection was further emphasized by linking to materials on the NRMP website in the course of explaining the matching procedure.

Under the assumption that the knowledge brought to bear in the NRMP preference submission was still accessible in a transparently similar experimental task deployed shortly thereafter, the knowledge of optimal play in the deferred acceptance mechanism held by these students represents the state of knowledge of incentivized, intelligent, and informed market participants. If understanding of the market were near universal in the NRMP match, one might expect that participants in this experiment would nearly universally rank schools in order of their monetary payoffs. If, instead, some medical students went through the NRMP while harboring misconceptions about profitable manipulation strategies, then nontruthful preference reporting should persist in this experimental task. Results support the latter possibility: 23% of the 1,714 participants submitted a preference ordering other than the dominant strategy of truthtelling. While the possibility remains that behavior in experimental tasks fundamentally differs from behavior in "real" matches, this study suggests that the body of lab evidence on nontruthful reporting cannot be easily dismissed as a consequences of different populations and non-externally-valid training.

2.2.2 Field Evidence: Obvious Misrepresentations

Given the results of the experimental literature, there is clear motivation to test whether preferences are indeed submitted truthfully when strategy-proof school-choice mechanisms are deployed in the field. Unfortunately, field testing of nontruthful reporting is extremely challenging. Conceptually, the difficulty is simple to convey. Say a researcher observes that a student submits a preference order indicating that school A is preferred to school B. To flag this submission as nontruthful, the researcher must know that the student in fact prefers school B to school A. While there may be cases where one believes that most students *should* prefer school B, it will rarely be possible to completely eliminate the potential for idiosyncratic preferences to explain the preference submission.

Based on this fundamental challenge, field tests of truthful preference reporting have been largely limited to examinations of relatively special markets in which the idiosyncratic pref-

erence explanation can be managed. A key class of such settings was identified by Hassidim et al. (2021). These authors direct attention to matching environments that are designed to simultaneously determine admissions and financial aid. Some such environments exist in which variants of the deferred acceptance mechanism are applied. In these cases, applications to a school without a scholarship and with a scholarship are treated as distinct items to rank in one's preference order. While economists may balk at insisting that school A *should* be preferred to school B, most economists' respect for potential preference idiosyncrasies does not extend to balking at the notion that attending school A with a prestigious scholarship *should* be preferred to attending school A with no scholarship.[18] Preference submissions that imply violations of this ordering can thus serve as comparatively unobjectionable indicators of nontruthful preference submissions or "mistakes." Hassidim et al. (2021) label such submissions as containing an *obvious misrepresentation*.

This approach to flagging nontruthful preference submissions has been applied in three large-scale applications of the deferred acceptance mechanism: the Psychology Master's Match in Israel (Hassidim et al., 2021), and college admissions matches for Victoria, Australia (Artemov et al., 2020) and Hungary (Shorrer and Sóvágó, 2022). Artemov et al. (2022) provides a concise summary of these papers' findings. Conditional on listing programs that make identifying this class of mistakes possible, the rate of obvious misrepresentations varied from 17-35% across these settings. In short, even this relatively extreme type of preference misrepresentation appears somewhat common, which invites the concern that less obvious and extreme types of preference misrepresentation are more common still.

2.2.2.1 Correlates of Mistakes. While studying the causes of preference misrepresentation is fundamentally more challenging in the field than in the lab, this body of work documents two clear and important correlates: students' desirability from the perspective of schools and the selectivity of different schools. Supporting a role for students' desirability, Hassidim et al. (2021) and Artemov et al. (2020) both documented that students with weaker applications were more likely to submit obvious misrepresentations. Supporting a role for

[18]However, some potential for additional balking remains. In some circumstances, requesting a scholarship may come with stigma or be dissuaded by social norms (as in Aygün and Turhan, 2017). This must be ruled out to use apparent dislike of a scholarship as evidence of preference misrepresentation.

program selectivity, Shorrer and Sóvágó (2023) showed changes to reporting behavior following a national funding cut that reduced scholarships in some fields while leaving other fields relatively unscathed. When the funding cuts went into effect, obvious misrepresentations involving programs facing funding cuts more than quadrupled, suggesting a causal role of perceived selectivity. These findings were further supported by within-subject analyses, finding that students are more likely to commit an obvious misrepresentation for the programs they rank that have higher admissions cutoffs.

These findings suggest possibilities for contributing factors to misrepresentation. Some focal possibilities are:

- Students with less desirable applications are expected to be less cognitively able, which potentially directly relates to strategic sophistication.[19]
- Students considering an option where they are unlikely to match may apply heuristics that make sense in other matching domains.
- Students considering an option where they are unlikely to match may react to psychological motives outside of our standard models, such as avoidance of disappointment.
- Students considering an option where they are unlikely to match may rationally not rank them to avoid typically unmodeled application costs.

In both the field and experimental literature, there have been a number of attempts to test the candidate explanations above. By our reading, there are some compelling demonstrations of particular items being active in particular domains, and we view the current state of the literature to suggest that there is some potential for all of these stories to have an influence. We will further discuss theories related to these points in Section 3.

[19]This motivates some studies to attempt to predict misrepresentations while simultaneously measuring students' priority and cognitive ability. For example, Artemov et al. (2020) conducted such exercises while controlling for raw test scores (as a measure of cognitive ability) and a percentile-based measure of performance (as a measure of priority). Rees-Jones and Skowronek (2018) conducted such excises while controlling for performance on Raven's matrices (as a measure of cognitive ability) and randomly assigned test scores (as a measure of priority). These exercises support some direct role for cognitive ability, but also suggest some remaining role of priority once it is disentangled from cognitive ability.

2.2.2.2 Payoff Relevance of Mistakes. Moving beyond the baseline rate of of preference misrepresentation or "mistakes," a major topic of interest within this literature is the frequency of *payoff relevant* mistakes. Consider a case where a person who truly prefers school A over school B submits preferences that represent that B is preferred to A. If, for example, the student would match to B under either preference submission (holding all other preference submissions fixed), then this "mistake" is *ex post* inconsequential or *payoff irrelevant*. While truthfully submitting preferences is a weakly dominant strategy, in this case the "weakly" qualifier binds, and in that sense the out-of-order preference submission is not *ex post* mistaken. Of course, these *ex post* evaluations can make use of information that a student does not have at the moment of choice. If, at the moment of choice, the student believes he has a positive probability of admission at school A, then truthfully ranking his options leads to higher expected utility. If the student believes he has no chance of admission at school B, failing to rank B correctly has no subjective expected utility cost. Taken together, this line of reasoning makes clear that at least some apparently out-of-order preference submissions might not be evidence of misunderstanding of incentives. While all of the field papers cited above engage with this line of reasoning, it is made particularly forcefully in Artemov et al. (2020).[20]

Across the three large-scale matches studied with the obvious misrepresentation methodology, it appears that misranking of programs with scholarships is often payoff irrelevant: the rate of payoff relevance is bounded between 2-8% in Hassidim et al. (2021), 1-20% in Artemov et al. (2020), and 12-19% in Shorrer and Sóvágó (2022). On the one hand, these rates are sufficiently low that, for certain types of analysis, they may be only minimally consequential. This is the primary interpretation offered by Artemov et al. (2020). On the other hand, the costs associated with foregoing a scholarship are sufficiently large that attention to these issues might be worthwhile in some analyses. To illustrate, the average cost of a payoff-relevant mistake was over $5,000 in Hassidim et al. (2021) and over $3,000 in Shorrer and Sóvágó (2022). Large populations of students exposing themselves to 10% chances of multi-thousand-dollar losses can have non-trivial effects on aggregate welfare. In

[20]This point is further developed in the closely related papers Fack et al. (2019) and Artemov et al. (2022), which we will discuss further in Section 4.1.

addition, Shorrer and Sóvágó (2022) documented that the higher rate of payoff-relevant mistakes among higher SES applicants serves an arguably desirable function in the Hungarian college match, leading scholarship money to be better targeted towards those in higher need. This demonstrates that the impact on welfare analyses can involve changes in distribution, rather than mere level effects.

2.2.3 Field Evidence: Survey-Based Approaches

Returning to the general problem of identifying preference misrepresentation: in order to identify a case of preference misrepresentation, an analyst must observe a student reporting that school A is preferred to school B but know that school B is preferred to school A. While it is challenging to identify such situations with only data on reported preferences, it is comparatively easy to identify such situations if external data on true preferences is available. This motivates another class of approaches to testing for preference misrepresentation: surveying respondents about their true preferences or their reporting strategy.

In an example of this approach, Chen and Pereyra (2019) tested for preference manipulation in the Mexico city high school match, which was organized by serial dictatorship. In this environment, they used survey data to identify students whose first choice should be to attend one of the selective high school associated with the *Universidad Nacional Autónoma de México* (UNAM).[21] Among this population, they found that one fifth of students did not top-rank a UNAM school. They interpreted these findings as evidence of *self-selection*—i.e. students omitting desirable schools where they think they will not be admitted—and provided additional analyses in support of that hypothesis.

In another example of this approach, Larroucau and Rios (2020) sought to test whether preference misrepresentation contributes to the phenomenon of "short lists." Administrators of strategy-proof matches typically encounter non-trivial fractions of students who submit preference lists with few options—for example, by only listing two schools despite the ability to list more. The optimality of these submissions is often viewed with skepticism: while

[21]Due to a lack of survey data directly eliciting preferences over high schools, Chen and Pereyra identify these cases based on a survey question about the type of university the student would like to attend after college. They argue that students who would most like to attend UNAM for university should view a UNAM high school as their first choice.

the example student could conceivably view only the two listed schools as acceptable, it often seems plausible that he instead has failed to list some schools that are preferred to remaining unmatched.[22] To test for this possibility, Larroucau and Rios used data from a survey administered to participants in the Chilean college match. In the survey, students were asked about their true first choice in the hypothetical case that they could be admitted anywhere. Focusing on the 90% of students for whom the preference-length constraint did not bind (who the standard model predicts are truthful), Larroucau and Rios found that only 43% rank their most preferred option first. Again, there was evidence that students omit programs where their chance of admission is "too low." These analyses support not treating "short lists" as truthful preference reports.

The examples just discussed involve comparing preferences inferred from survey data to preferences inferred from reports made to a match. An alternative approach to identifying preference misrepresentation using survey data is perhaps more direct: simply asking respondents whether they misrepresent their preferences. Illustrating such an approach, Rees-Jones (2018b) asked applicants to the National Resident Matching Program whether their submitted preferences matched their true preferences; 17% of respondents self-assessed their reporting strategy as nontruthful, with 5% identifying their deviation from truthful reporting to strategic considerations. Similarly, Hassidim et al. (2021) found that 20% of their surveyed participants in the Israeli Psychology Match reported that they misrepresented the ranking of some program in their submitted preferences.

Surveys have well-documented shortcomings relative to direct examinations of behavior, and yet they can offer information about market participants' motivations that would generally be inaccessible through other means. While rarely ideal for quantitative purposes, data from open-ended survey prompts can at times provide extremely clear evidence of the drivers of preference misrepresentation. For example, Gross et al. (2015) reports results from surveys and interviews with parents in Denver and New Orleans. At the time, each city used a constrained implementation of a strategy-proof algorithm. And yet, in these cities parents would regularly submit preference lists with very few schools, in line with the general problem of "short lists" discussed above. Gross et al. found that

[22] Of course, such reasoning would not be compelling in cases where students have strong outside options.

> ... some parents listed fewer choices because they were trying to strategically "outwit" the matching algorithm. Parents in both cities said that they had heard that listing the maximum number of choices was the best strategy, but few of them could reconcile this message with their assumptions about how the system worked. They assumed that "the system" was somehow gaming *their* preferences, and so they decided to list only one or two schools because they feared that listing more schools would send a signal to "the system" that they were happy attending a low-demand school. In other words, they worried that if they showed "the system" that they would be willing to accept a low-demand school—even if it was their fourth or fifth choice—"the system" would skip over their first choices and just give them the low-demand school to fill seats.

While economic analyses rarely directly use qualitative data of this variety, in our experience it is extremely valuable for guiding later quantitative analysis.

2.3 Summary

Since the emergence of education market design as a field, the potential for preference manipulation has been a central concern. In common manipulable mechanisms, determining the optimal preference reporting strategy is sufficiently demanding that some degree of suboptimality should be expected. This introduces a variety of challenges for economic analysts. For example, it implies that models assuming optimal preference reporting may fail to correctly characterize behavior and welfare, and that approaches to inferring preferences based on assumptions of optimal reporting may at times go awry.

A common line of reasoning suggests that these concerns are obviated by the use of strategy-proof mechanisms. Within such a mechanism, optimal reporting is simple, which perhaps could suggest that everyone could be expected to do it. If true, this would be extremely convenient for market designers: it would imply that models assuming optimal behavior function perfectly in strategy-proof environments despite their well appreciated imperfections elsewhere. In short, within strategy-proof environments, our quest for a model of behavior to use for policy analysis would be complete.

Unfortunately, as we have documented, a large literature now decisively refutes this common line of reasoning. Across a large body of work using a wide variety of methods, substantial fractions of market participants are reproducibly found to violate the predictions of the idealized, fully rational framework. Strategy-proof mechanisms have many benefits, but we should not count the complete elimination of apparently suboptimal behavior among them. This leaves open the possibility that other modeling approaches may better explain existing behavior, and may usefully guide the analysis of, or implementation of, school assignment systems.

3 Behavioral-Economic Explanations of Preference Misrepresenation

The central proposition of behavioral economics is that economic modeling might be improved by the integration of insights from psychology. In this section, we summarize the primary psychologically founded explanations that have been put forth to explain the apparently suboptimal behavior discussed in Section 2.2. These models largely fall into three categories: models of incorrect processing of incentives, models of non-standard preferences, and models of biased assessments of admissions probabilities. We discuss each category below, along with specific theories within each category and explanations of the field behaviors they can rationalize.

3.1 Models of Incorrect Processing of Incentives

Strategy-proof mechanisms are commonly lauded for the simplicity of their optimal strategy. And yet, while the strategy of truthtelling is simple, proving that truthtelling is incentivized is not. To illustrate this point, recall that the proof of the strategy-proofness of the deferred acceptance mechanism was deemed sufficiently non-trivial to merit publication in operations research and mathematics journals (Dubins and Freedman, 1981; Roth, 1982). These proofs are often presented as non-trivial exercises to second-year PhD students in market design field classes. We view it as very natural to imagine that some match participants might lack

the analytical rigor to independently deduce the incentive properties of these mechanisms.

Given these considerations, difficulties in processing strategic incentives provide a natural explanation for preference manipulation. We present several related theories below.

3.1.1 Failure of Contingent Reasoning

A prominent framework for modeling understandable mechanisms comes from the recent work of Li (2017). In this work, Li introduces the notion of *obvious dominance*: "a strategy S_i is *obviously dominant* if, for any deviating strategy S_i', starting from the earliest information set where S_i and S_i' diverge, the best possible outcome from S_i' is no better than the worst possible outcome from S_i." A mechanism that admits an equilibrium in obviously dominant strategies is called *obviously strategy-proof* (OSP).

One motivation for this classification comes from the behavioral economics literature on *failures of contingent reasoning* (for a review, see Niederle and Vespa, 2022). In a variety of domains, making contingent reasoning necessary has been shown to harm decision quality. In general, determining that one strategy dominates another requires applying contingent reasoning. It need not be the case that every possible outcome from playing the dominant strategy is preferred to every possible outcome of the alternative strategy. Rather, within every contingency, the dominant strategy's outcome is preferred. In contrast, for an obviously dominant strategy, it *is* the case that every possible outcome from playing the obviously dominant strategy is preferred. In that sense, OSP mechanisms protect a player who can be confused by inappropriately making cross-contingency comparisons.

Li shows that the predictions of this framework align with common intuitions that, e.g., ascending clock auctions are simpler to understand than second-price sealed-bid auctions. He also provides experimental evidence that supports these intuitions, considering three pairs of "equivalent" games, each with an *OSP* implementation and a *strategy-proof but not OSP* implementation. One pair of games involves the strategy-proof serial dictatorship mechanism that at times is used in school-choice applications (and that functions as a special case of the deferred acceptance mechanism). In Li's experiment, preference misrepresentation is less common in the (sequential) OSP implementation of serial dictatorship than in the (static) non-OSP implementation.

If preference misrepresentations are driven by contingent-reasoning failures, design and deployment of OSP mechanisms may be helpful. We will return to discussion of this possibility in Section 4.2.

3.1.2 Application of Improper Heuristics.

As we discussed in Section 2.1, the immediate acceptance mechanism introduces strong incentives to have attainable first choices. The deferred acceptance mechanism eliminates this strategic incentive. And yet, the literature discussed in Section 2.2 recurringly finds evidence that participants in the deferred acceptance mechanism fail to pursue high-value/low-probability options. One way to characterize this finding is as showing that some individuals facing the *deferred acceptance* mechanism act as if they face the *immediate acceptance* mechanism, or more broadly act as if avoiding rejections is incentivized when it is not. This suggests a particularly simple explanation for this behavior: students are intentionally strategizing in a manner that is misguidedly optimized for a mechanism that is not strategy-proof.

Reliance on incorrect strategies could arise through several natural channels. First, this could happen if students misunderstand the explanation of the matching procedure and literally confuse one mechanism for the other. This possibility is analogous to game-form-misconception-based explanations of failure to truthfully report valuations in the Becker, DeGroot, and Marschak (1964) mechanism (Cason and Plott, 2014). Second, this could happen if students do not fully attempt to understand the matching procedure and instead merely apply rules that served them well in the past. The tendency to rely on such heuristics is among the the most common themes in research on judgment and decision making (for extensive coverage, see Gilovich et al., 2002), and seems particularly compelling here. Because matching markets akin to the the immediate acceptance mechanism are relatively common and applications of the deferred acceptance mechanism are so rare, individuals who attempt to guide their strategy with experience in seemingly similar markets could very naturally make these mistakes.

This possible explanation for preference misrepresentations has been discussed at times in this literature (see, e.g., Hassidim et al., 2017a, 2021). And yet, to our knowledge, no paper has made testing this explanation its focal goal. We encourage future work in this direction,

although we warn that the current absence of this research is likely due to the substantial difficulty of disentangling this explanation from others (including some presented in this section).

3.1.3 Misunderstanding Continuation Value.

When participating in constrained versions of strategy-proof mechanisms, students must decide on a subset of all schools to include in their preference submissions. Students are making static decision—each deciding what preference order to submit—but it can be useful to frame these decisions as reflecting dynamic optimization problems. In this dynamic framing, the algorithm will sequentially attempt to match students to schools in the order indicated in their preference submissions. When considering the school to rank in a particular position, students should consider both the utility they will realize if admitted to that school and the expected utility if they are rejected. Using terminology from dynamic optimization, the expected utility following rejection of the currently considered option is the *continuation value*, and represents the utility that the student expects to receive if the algorithm attempts to make a match among the schools the student has ranked lower.

Wang et al. (2021) presents and tests a theory in which preference submissions are harmed by a failure to properly appreciate continuation values. This offers an explanation for preference submissions that are overly cautious. To illustrate this theory, consider a student who must submit a list of three schools to be processed by the deferred acceptance algorithm. In the framework of Wang et al., this student ranks schools in a manner suggested by the *directed cognition model* of Gabaix et al. (2006).[23] As summarized by Gabaix et al., the key feature of this model is that "at each decision point, agents act as if their next set of search operations were their last opportunity for search." In the Wang et al. framework, this translates to the student always acting as if his continuation value is the expected utility of ending the process unmatched. The true continuation value is more favorable than this student assumes: for example, upon being rejected by his first choice, he could conceivably match to his second or third listed schools. Failure to appreciate this leads to overly conser-

[23] Ajayi and Sidibe (2022) also consider the directed cognition model (specified differently) as it applies to engagement with matching mechanisms. They consider it among several search models presented to explain behavior in the Ghanaian high-school match.

vative preference submissions: the student acts as if he is in no way hedged against rejection by his first and second choices, and thus is more avoidant of rejection risk than he should be.

Wang et al. motivate their interest in this model by documenting puzzling behaviors in the preference submissions of students in Ningxia, China's college admissions system. In this application of the constrained deferred acceptance mechanism, a plurality of students chose colleges that offer them nearly assured admission as their first choice. Additionally, a majority of students exhibited "competetiveness reversals," which entail listing a less selective college as more preferred to a more selective college. While these two findings need not be evidence of mistakes on their own, a variety of supporting analyses suggest that mistakes have a role. The directed cognition model helps to rationalize these behaviors, and is further supported by a complementary survey experiment. The authors use these results to motivate estimating a mixed-type structural model and find that approximately half of students are better explained by the directed cognition model than by a fully rational alternative.

3.2 Models of Non-Standard Preferences

A second class of behavioral-economic explanations of preference misrepresentation are derived from models of non-standard preferences.

To illustrate, consider a student participating in a market organized by a strategy-proof mechanism. This student prefers going to school A over school B, but submits the preference order $B \succ A$. This is typically discussed as a misrepresentation or mistake in the literature summarized in Section 2.2. However, it could be that this submission is fully deliberate and, in some sense, rational. If this student is influenced by psychological factors that depend on the preference submission that is made, the psychological benefits of submitting $B \succ A$ might outweigh the expected direct payoff improvement from instead reporting $A \succ B$. Some behavioral economic models predict such a tradeoff, as we illustrate in the examples below.

3.2.1 Direct Concern about Getting a Highly Ranked Choice.

Consider a student in a market with five schools that are clearly ranked by quality. Imagine that this student has very low priority, and is essentially guaranteed to match to the fifth (i.e., the worst) school. The benefit of a strategy-proof mechanism is that it eliminates any negative *strategic* consequences from ranking the four better schools higher than school five. Notice, however, that submitting truthfully generates a *psychological* consequence through an aversive pattern of feedback. If this student submits truthful preferences and ultimately matches to school five, the student gets clear feedback that he was rejected by the four better schools. If, instead, this student ranks school five first, he can avoid ever facing those rejections. This may facilitate the student believing (or representing to others) that his ultimate match was truly his favorite, and that he could have gone elsewhere if he had just applied differently.[24] Ego-management stories like this are commonly explored in behavioral economic models (see, e.g., the "Ego Utility" model of Kőszegi, 2006), and seem particularly natural in the school-choice context.

Motivated by considerations like these, Meisner (2022) provides a theory of matching under *report dependence*. Agents in this model "dislike rejections and enjoy the confirmation from getting what they declare desirable." As Meisner demonstrates, preference manipulation can persist even for arbitrarily small degrees of report dependence. Furthermore, truthful preference reporting is guaranteed to be optimal only in the (implausible) circumstance that more desirable schools are always more likely to grant admission. Kloosterman and Troyan (2022) provide experimental support for the role of report dependence.

This theory can serve as a natural explanation of the tendency for individuals to "self select" out of applying to schools with low probabilities of admissions (as was found by Chen and Pereyra, 2019, and others). This also provides an explanation for the recurring finding that preference misrepresentation is more common among students of low ability or low priority.

[24] Anecdotally, motives like this appear common in the residency match. Medical students are regularly advised to adopt the story that their ultimate match was their first choice, regardless of the truth of that claim. This perhaps contributes to doctors' tendency to list attainable first choices at a surprising rate.

3.2.2 Expectations-Based Reference Dependence.

Prospect Theory (Kahneman and Tversky, 1979) is the most widely adopted model developed in the behavioral economics literature (see O'Donoghue and Sprenger, 2018, for a recent review). Two of the primary features of Prospect Theory are the assumptions of *reference-dependence* (which dictates that consumption amounts are considered as gains or losses relative to a reference point) and *loss aversion* (which dictates that marginal utility is distinctly higher over losses versus gains). As the term "loss aversion" suggests, a general prediction of this model is that individuals seek to avoid losses beyond the typical degree of avoidance in standard expected utility models.

Dreyfuss et al. (2022b) consider the preference-submission decision problem of an agent with expectations-based loss aversion (EBLA), as arises in the Prospect Theory variant developed by Kőszegi and Rabin (2006, 2007, 2009). As Dreyfuss et al. summarize, there are two key forces in the EBLA framework that help explain preference misrepresentation. First is an endowment effect for schools that offer high probabilities of admission. Under this theory, "a student who is likely to get matched with a school will feel a loss when matched with any other school (even a better one)." This offers a motive for the student to move a likely match up in their reported preferences to avoid such feelings of loss. Second, avoiding creating an expectation of matching with high-value/low-probability schools helps avoid a sense of loss in the likely situation that the student does not match with them. This offers a motive for the student to move such schools down in their reported preferences or to omit them altogether. In relation to the results of Section 2.2, this can serve as an explanation for both the existence of preference misrepresentation[25] and for the fact that it is more common when admission is less likely.

This modeling framework was further developed by Meisner and von Wangenheim (2023), who characterized rationalizable preference submissions and the choice-acclimating Bayes-Nash equilibria that arise in markets of EBLA actors. Meisner and von Wangenheim (2021)

[25]Note that "preference misrepresentation" may be viewed as a misnomer under this theory. If loss aversion is viewed as a preference and not a bias, then ranking schools out of order of their straightforward valuations may indeed be preference maximizing. In that case, the term "preference misrepresentation" should be understood as meaning that reported preferences do not represent the preferences that the individual would express if matches could be made through choices from menus.

and Dreyfuss et al. (2022a) also consider the potential for sequential mechanisms to reduce the role of the expectations-management incentives induced by EBLA. We return to this issue in our discussion of behaviorally guided mechanism choice in Section 4.2.

3.2.3 Unknown Preferences and Learning

Like many economic models, the standard models used in market design assume that agents know their own preferences. Literal compliance with this assumption would require students to understand the experience available in all schools within their district, or all colleges where they could apply. In many contexts, so many options are available that researching all of them would be prohibitively costly. The presence of consideration costs can offer natural explanations for some common types of preference misrepresentation. For example, a student might report preferences out of order based on an incompletely informed assessment of different schools. As another example, "short lists" might be submitted to avoid the costs of developing informed preferences over a larger list.

A significant body of recent research has examined the consequences of these costs for education market design (see, e.g., Bade, 2015; Immorlica et al., 2020; Artemov, 2021; Bucher and Caplin, 2021; Chen and He, 2021, 2022; Grenet et al., 2022; Noda, 2022). Much of this research need not be considered behavioral economic in nature, and instead might be better characterized as applying the rational-choice framework with more attention to modeling the costs inherent in the application process. However, not all findings can be interpreted through that lens. For example, in the experimental environment studied in Chen and He (2021), students overinvested in information relative to the rational benchmark. Furthermore, students' *curiosity*—measured by their willingness to pay for noninstrumental information—predicted this overinvestment.

Non-instrumental demand for information, as well as the broader process whereby preferences are constructed, are both topics of longstanding literatures in psychology or economics (for relevant reviews, see Slovic, 1995; Golman et al., 2017; Sharot and Sunstein, 2020). The findings above suggest potential for these literatures to usefully inform market design, and we encourage further pursuit of such work.

3.3 Models of Incorrect Processing of Probabilities

A third class of behavioral-economic explanations of preference misrepresentation are derived from models of biased probability assessments.

Biases in probability assessments are one of the most broadly studied phenomena in behavioral economics (see Benjamin, 2019, for an extensive review). This literature suggests that problems with probability forecasts are common and at times severe. If that premise is accepted, it serves as yet another reason to appreciate the fact that strategy-proof mechanisms make attention to admissions probabilities unnecessary. A person who knows their valuation of different match options, but has inaccurate assessments of admissions probabilities, could simply report their preferences truthfully without worrying that they have pursued a suboptimal strategy.

Note, however, that assessments of probability of admission can remain relevant even in the fully rational model. If a student has no chance of acceptance at a school, there is no benefit to ranking this school in a preference submission. Furthermore, if a student will surely match to one of their top N ranked schools, there is no benefit to continuing to rank schools beyond the N^{th}. In either case, while truthtelling is a weakly dominant strategy, strategies that omit impossible matches would still be optimal.[26] This argument illustrates that probability perceptions can remain relevant even in an idealized setting. These issues become even more relevant if ranking each school is costly, or if preference-list length constraints bind.

Because of this remaining role of probability assessments, the wealth of different models of biased probability evaluation may be relevant in understanding preference reporting behavior. We highlight two examples below.

3.3.1 Overconfidence.

Across a variety of contexts, researchers in economics and psychology have documented some tendency for overconfidence. The dominant paradigm, laid out in Moore and Healy (2008),

[26]Of course, this claim relies on correct probability perceptions. If, for example, a student elects not to rank a school where they think a match is impossible, but in fact it is merely unlikely, then omitting this school is costly in expectation.

involves three related but distinct phenomena:

- *Overestimation:* Thinking that you are better than you are.
- *Overplacement:* Thinking that your position relative to others is better than it is.
- *Overprecision:* Thinking that your forecasts are more accurate than they are.

This taxonomy of overconfidence helps make clear the ways in which overconfidence can become relevant to preference reporting. Overplacement could lead students to believe that they are higher in a schools' preference ordering (or in the distribution of priorities) than they truly are. This could naturally lead students to undervalue "safety school" applications, and could guide students to apply to low-probability options despite a binding preference-length constraint. Overprecision could lead to misvaluation of both "safety" and "longshot" options. When considering such schools, a student would be unduly confident that their admission was assured or impossible, respectively.

In a lab experiment, Pan (2019) considered the role of overconfidence in an environment where students must rank schools prior to completing a test that determines their priority. When a match is conducted using the immediate acceptance mechanism, students should be careful to have attainable top-ranked choices. However, overconfidence can lead them to view unattainable options as attainable, leading to strategically unwise submissions and instability in the final assignment. Pan (2019) documents these outcomes and shows that mismatch was reduced by replacing the immediate acceptance mechanism with the strategy-proof serial dictatorship mechanism. This study clearly demonstrates a role of overconfidence in the manipulable immediate acceptance mechanism.

Important roles for overconfidence appear in several recent field evaluations of matching procedures. Kapor et al. (2020) studies behavior in the New Haven high-school match governed by the immediate acceptance mechanism.[27] They document that preference submissions regularly appear strategically unwise, and using survey data they document that mistaken assessments of admissions probabilities contribute to these apparent mistakes. Arteaga

[27]More precisely, in different years considered in Kapor et al.'s study, New Haven used both the immediate acceptance mechanism and the closely related "New Haven" mechanism. Simulations in Kapor et al. (2020) suggest that these mechanisms result in the same placement for 99% of applicants.

et al. (2022) study the Chilean centralized primary and secondary school match, which applies the deferred acceptance mechanism. They document that approximately 20 percent of applicants submitted a "risky" rank-order list—one where the applicant faces a probability of being unassigned greater than 30 percent. Surveys suggest that these decisions are partially misguided, with applicants appearing to overestimate their chance of assignment. Fabre et al. (2021) studies behavior under constrained deferred acceptance in the Chilean college match. They too found that students underestimate the probability of not being assigned. More generally, they found that students attenuate the probability of extreme events (including the probability of admission to "reach" schools). Both Arteaga et al. (2022) and Fabre et al. (2021) presented interventions aimed at correcting students' misperceptions, which we discuss further in Section 4.3. Bobba and Frisancho (2022) elicited students' beliefs about their performance in a centralized exam that would determine their admission priority to high school in Mexico City. They found discrepancies between students' beliefs and their actual score, and overly optimistic (perhaps overconfident) beliefs are more common.

Taken together, this body of work suggests a relatively clear connection between overconfidence and otherwise-puzzling field behaviors, and suggests productive paths to guiding students to better choices in such cases.

3.3.2 Correlation Neglect.

A body of research within behavioral economics presents evidence of *correlation neglect*: a tendency for individuals to treat correlated signals as if they were independent (see, e.g., Enke and Zimmermann, 2019). Rees-Jones et al. (2020) explore the implications of this theory for interactions with the constrained deferred acceptance algorithm. This theory offers an explanation for preference submissions with an unappreciated high risk of failure to match.

To see the potential for correlation neglect to be relevant, consider a simple example drawn directly from Rees-Jones et al. (2020). Imagine a student considering three schools offering utility payoffs of 3, 2, and 1 if the student matriculates (and a payoff of zero if the student does not match). Further assume that these schools will evaluate the student for admission based on a currently unknown priority score, to be drawn uniformly from the

integers between 0 and 99. Assume the thresholds for admission at the three school are 50, 45, and 0, ranked in order of the school's desirability. Finally, assume that this market is organized by a constrained deferred acceptance mechanism, permitting the student to submit a list of only two schools.

While this student could conceivably be tempted to submit the preference that lists the best two schools, such a submission strategy would be unwise because the admissions decisions are so correlated. To illustrate, notice that the second school on the preference submission will only be considered conditional on rejection by the first-choice school. In this case, we could infer that the student's priority score is below 50. Conditional on that fact, the probability of acceptance at the second-best school is a mere 10%. The student is better off listing the worst school as his second choice: even though it offers half the payoff, the fact that it is available with certainty compensates in expected value terms.[28] If, however, the student acts as if admissions were evaluated independently, he would expect a 55% chance of admission at the second-best school after rejection by the best school. Applying to this school would then mistakenly be viewed as appealing, and a student making this mistake would underestimate his chance of remaining unmatched.

This example illustrates that correlation-neglectful agents will be overly aggressive in filling slots after the first in a preference submission, leading to an undervaluation of "safety school" options. At this extreme, this overaggression can lead students to list schools where admission is impossible conditional on rejection by the first choice—for example, because the test-score threshold is higher. Patterns like these have been documented in several field settings in which preferences are submitted before test scores are available.[29] Rees-Jones et al. (2020) provide a series of lab experiments that support a role of correlation neglect in environments meant to resemble these field settings. These findings offer guidance on the assignment procedures that make correlation-neglectful preference reporting especially problematic.

While the predictions of correlation neglect operate through misassessment of correlated

[28]In fact, listing the worst school as the second choice leads to a gamble over outcomes that second-order stochastically dominates the gamble induced by submitting the middle school as second choice.

[29]Examples include nationwide secondary school matches in Ghana and Kenya and the University and Colleges Admissions Service of the UK. For discussion, see Ajayi (2014), Lucas and Mbiti (2012), or the summaries provided in Rees-Jones et al. (2020).

probabilities, it is possible that this is an as-if representation derived from higher-level failures of reasoning. For example, correlation neglect can occur among agents with failures of contingent reasoning (as in Section 3.1.1), and can be modeled as a failure in calculating continuation values (as in Section 3.1.3). Further research into the deep foundations of correlation neglect may help guide efforts to combat it.

3.4 Managing the Multiplicity of Behavioral-Economic Theories

As illustrated by the examples above, behavioral economics offers a variety of plausible explanations for preference misrepresentations, many with empirical support. This suggests a clear path for the integration of behavioral economics in mainstream market design: these models could be further vetted, and those that are strongly supported in contexts of direct interest could be integrated into the standard theoretical toolkit.

A challenge to pursuing this path comes from the great variety of theories that have been put forward. A common criticism of behavioral economics is that it involves many disparate models rather than a single unifying framework. Absent a single unifying framework, theorists must use undesirable discretion to determine which model of behavioral forces to use in their analysis.

In some ways, this common criticism is often overblown. The alternative to behavioral economics is not a single, unified model. The rational choice framework may also be understood as a large collection of individual models with similar researcher discretion required. When writing rational theories, we (as a research community) understand that attention to intertemporal trade-offs can be important, and yet often model static environments when the intertemporal component seems inessential. We understand that uncertainty can be important, and yet often model deterministic environments when risk seems inessential. Similar contrasts exist for imperfect vs. perfect information, costless vs. costly search, and many other dimensions of rational choice. In short, there is nothing unusual about a researcher having to make determinations about the essential elements of a modeling exercise, and then choosing among different developed models according to that determination.

In other ways, this common criticism is justified. While we argue that there is nothing unusual about applying discretion to select the appropriate model, we note that much more

discretion is needed when selecting from the menu of behavioral models. The problem with behavioral models is not that there are many of them, but rather that there are *so many* of them. In most cases, determining a desirable rational model requires assessing a comparatively managable check-list of considerations. The number of candidate psychological considerations is much larger. What's worse, there is less developed evidence and modeling experience to draw from to inform which models to choose. Worse yet, there is plenty of reason to suspect that different models may describe different individuals in the same environment.

One potential path to influence of behavioral economics in education market design is to firmly establish a useful role for a manageably small number of behavioral theories. This would require identifying types of behavior that are sufficiently common in certain environments to explain good fractions of observed preference misrepresentations. It would also require clear theoretical and empirical guidance on the circumstances under which the presence of this common behavioral-economic consideration affected bottom-line market design considerations, and the circumstances under which the behavioral-economic consideration may be safely ignored. The literature reviewed in this section shows that efforts to guide such decisions are well underway for several candidate theories. And yet, there is substantial work to do before theorists could feel that their adoption of a behavioral feature was as natural and informed as their adoption of different components of the standard rational-choice framework.

Another path to potentially broader influence is to motivate the development of tractable frameworks that incorporate broad classes of psychological considerations. As a concrete example, this could involve noticing that many of the behavioral theories discussed above predict that students will avoid listing a school with low admission probability as a top choice. For some questions, the analyst must know the reason these submissions are foregone in order to assess the consequences for market design. For other questions, merely knowing preferences are submitted in this way may be enough, and working with a theory that explicitly models many possible explanations for that behavior may be unnecessary. In such cases, the analyst can make use of this behavioral prediction without taking the difficult step of fully determining the relative contribution of different models in generating it.

The benefits and drawbacks of this type of approach have been thoroughly studied in the *behavioral sufficient statistics* approach that is now common in public finance. The general idea of the standard sufficient statistics approach, as articulated by Chetty (2009), is to bridge the gap between structural and reduced-form methods. In such an approach, one begins from a full structural model. The researcher then works within the context of that model to express welfare effects as a function of objects amenable to direct measurement. In the public finance context, this could mean writing welfare not as a function of primitives of utility functions, but instead as a function of measurable elasticities.[30] While expressing one's policy outcomes of interest in this way is not always possible, it sometimes is, and attempts to pursue such a representation often yield insights that are missed by purely structural exercises. Furthermore, a line of research building from Chetty et al. (2009) has demonstrated that this type of approach is perhaps especially fruitful when behavioral forces are included in the model. In some cases, behaviors with a variety of behavioral-economic explanations, perhaps all relevant simultaneously, can be simply accommodated without the loss of tractability that would occur if the primitives of each contributing theory needed to be fully specified and measured.[31]

We believe that attempts to pursue a similar path in education market design might be fruitful, and a similar degree of ultimate influence may be possible. The connection may initially seem tenuous given the very different, discrete-mathematics-based framework typically adopted in market design. However, recent works building from the large-market framework of Azevedo and Leshno (2016) show deep connections between the operation of matching markets and standard price theory. In such a large-market framework, connection

[30]While we emphasize the benefits of such an approach, it is not without costs. For example, as is well understood in public finance, if we have not modeled the fundamental determinants of elasticities, then the elasticity that serves as a sufficient statistic might change across contexts in a manner not predicted by the stripped-down model. Connecting this to the market design issue we consider, if we do not model the reason that students forgo risky first choices we may fail to predict how this behavior will change across environments. This limits the value of this approach when attempting to make policy predictions in novel environments. It also means that the results of the model do not necessarily inform the most productive ways to intervene (since interventions may have different impacts on different potential models contributing to the behavior of interest).

[31]For examples, see Allcott and Taubinsky (2015), Rees-Jones (2018a), Taubinsky and Rees-Jones (2018), Allcott et al. (2019), or Rees-Jones and Taubinsky (2020). The potential of behavioral sufficient statistics to be broadly incorporated into public finance theory is especially prominently displayed in Farhi and Gabaix (2020).

to these behavioral sufficient statistics approaches may prove to be relatively natural.

4 Implications of Behavioral Economics for the Conduct of Education Market Design

As we have reviewed in the prior section, behavioral economics offers a variety of explanations for the persistent preference misrepresentation observed in centralized education markets. Many of these explanations engage with motivations or cognitive limitations that fall outside of the idealized, fully rational framework, but that capture plausible and empirically supported accounts of behavior in these markets.

Say that the relevance of some of these behavioral accounts is accepted. What does this then imply for conduct of education market designers? In this section, we consider the endgames of behavioral economics for market design. We first focus our discussion on the potential for behavioral economics to influence three common decisions that market designers must make: decisions about estimation strategies, decisions about desirable mechanisms to deploy, and decisions about auxiliary interventions to assist students in centralized markets. We then consider the potential for behavioral factors to influence the individuals who must choose a mechanism or approve its use.

4.1 Influencing Estimation Strategies

One way in which behavioral economics could inform education market design is by influencing the approaches adopted for preference estimation.

Strategy-proof mechanisms greatly facilitate inference on preferences. Under the assumption that strategy-proof mechanisms induce all students to follow the weakly dominant strategy of truthtelling, the administrators of these mechanisms have access to unusually rich preference data. While typical economic datasets indicate the option chosen from a menu, preference submissions in these markets indicate the full preference ordering of the universe of options.[32] This is extremely useful to analysts aiming to estimate preference

[32]Indeed, the unusual availability of such data was what led one of us to initially be exposed to this field. Benjamin et al. (2014) studied behavior in the medical residency match to compare the results from

models and conduct policy simulations. A rich literature has developed and applied such revealed-preference methods in this domain (see Agarwal and Somaini, 2020, for a review).

The suboptimal reporting strategies documented in Section 2.2 violate the common assumptions in these revealed-preference analyses. Of course, this revealed-preference framework can still be extremely useful despite such imperfections. Like many researchers, we interpret this framework not as a literal account of all behavior, but as a useful approximation that enables progress on difficult but important problems. Given that attitude, we advise against the destructive framing of behavioral-economic results: the framing in which their purpose is to establish that the rational framework is incorrect. Rather, we view the role of behavioral economics to be constructive, for example by providing guidance on the most robust estimation strategies or on the structure that structural models should contain. In the remainder of this subsection, we discuss progress along these lines.

4.1.1 Estimating Preferences in the Presence of Inconsequential Mistakes

The recent works of Fack et al. (2019) and Artemov et al. (2020, 2022) develop a framework for accommodating certain classes of mistakes in preference estimation. In situations where mistakes are payoff irrelevant (or nearly so), these papers illustrate the merits of relying on assumptions about the stability of the final match rather than the assumption of full truthtelling.

To illustrate the conceptual approach, consider a student in a large market with five overdemanded schools labeled A through E. The student's preferences are $A \succ B \succ C \succ D \succ E$. Further assume that admission at these schools is determined by a test score. While the score thresholds for admissions depend on this year's pool of applicants, imagine that the distributions of preferences and scores are sufficiently consistent year-to-year that the thresholds do not vary much. Based on his score and these past thresholds, this student knows that he will not be admitted to schools A or B, but he expects to be admitted by C, D, or E if his application is considered for a match. Call these the *feasible schools*. Because

revealed-preference and survey-based methods of inference on preference parameters. Interactions with medical students in the course of running this survey led to some doubts that all were truthtelling, which motivated the collection of further survey data to assess robustness of inference that assumes fully optimal reporting. See Rees-Jones (2018b) for further discussion.

school C is strictly preferred to D and E, and because D and E are feasible schools, assigning this student to school C is the only way to avoid justified envy and instability.

In this set-up, we may consider the merits of two assumptions that could be used as a basis for estimation. The first is the *truthtelling assumption*: that submitted preferences reveal true preferences. The second is the *stability assumption*: that the result of the match yields a stable outcome with respect to true preferences. Notice that the set of preference submissions that satisfy the truthtelling assumption is a (typically much smaller) subset of the set of preference submissions that satisfy the stability assumption. For example, consider the following potential preference submissions, with misordered items flagged in bold:

$$\mathbf{B} \succ \mathbf{A} \succ C \succ D \succ E$$
$$C \succ D \succ E \succ \mathbf{A} \succ \mathbf{B}$$
$$C \succ D \succ E$$

In these examples, the misordering or omission of schools A and B would violate the truthtelling assumption but not the stability assumption. As related examples, consider:

$$A \succ B \succ C \succ \mathbf{E} \succ \mathbf{D}$$
$$A \succ B \succ C$$

Again, the misordering or omission of schools D and E would violate the truthtelling assumption but not the stability assumption.

The empirical content of the stability assumption is that a student's preference submission ranks the school where they would ultimately stably match over all other feasible alternatives. In the context of our example, if this student commits any "mistake" in preference submission that preserves school C being ranked over schools D and E, the truthtelling assumption would be violated while the stability assumption would remain valid. This permits payoff-irrelevant mistakes and forbids payoff-relevant mistakes, each in the sense discussed in Section 2.2.2.

Fack et al. (2019) formalize this argument and illustrate its consequences for preference

estimation. While the assumption of full truthtelling justifies using all information present in the reported preference list, the assumption that stability is preserved justifies using more common discrete-choice machinery, modeling only the most preferred item from a given choice set (here, the feasible set). Fack et al. also present a model in which nontruthful reports arise due to avoidance of application costs and notes that such a model supports their assumptions of stability. Returning to our example, if our student definitively knew that schools C, D, and E formed his feasible set and faced costs of application or consideration, he would face a clear incentive to indicate that C was ranked over D and E and to avoid incurring costs for any other comparisons. In short, such a model illustrates incentives to make payoff-irrelevant mistakes and illustrates incentives to avoid payoff-relevant mistakes.

While we have discussed these results in the context of literal stability for ease of exposition, these intuitions extend to markets with *approximate* stability in the sense that they are asymptotically stable as the market grows large. In such a model, mistakes are not required to be payoff irrelevant but instead are required to have small and asymptotically vanishing payoff consequences. Parts of this argument appear in Fack et al. (2019), and they are further refined in Artemov et al. (2022).

In short, this framework provides an approach to preference estimation that is robust to the presence of payoff-irrelevant mistakes. In environments where all mistakes are payoff irrelevant, this in some sense provides a complete solution to the problem at hand. In environments where some mistakes are payoff relevant and some are not—as many of the models discussed in Section 3 predict—this approach has the benefit that its foundational assumption will be violated in fewer observations than would an assumption of full truthtelling. While this has clear intuitive appeal, the literature has not yet provided detailed guidance on the magnitude of estimation biases under these alternative assumptions in the presence of payoff-relevant mistakes. We encourage further work along these lines.

4.1.2 Adding Psychological Structure to Structural Approaches

Another natural way of accommodating behavioral economics in preference estimation is to directly include psychological structure in the models one estimates. This practice is well underway, with a number of papers referenced in Section 3 including such estimation

exercises.[33]

The motivation for using behavioral findings in this way will depend directly on the successful advancement of behavioral theories. That argument is straightforward and well understood: if the research community decides a particular model is "the right one" for evaluating a certain environment, then that model will be the one to estimate. We now draw attention to a complementary reason why behavioral-economic models could at times be viewed as desirable: because they at times greatly facilitate tractable analysis.

This potential is clearly demonstrated in the recent work of Idoux (2022). This paper aims to assess the relative contribution of preferences and priorities as determinants of segregation in New York City middle schools. To illustrate the exercise, consider a school district with two schools. Imagine that, after running the deferred acceptance algorithm, the district assigns most students of one race to school A and most students of another race to school B. The resulting segregation could arise from two sources: the priority score used to determine admission could differ across racial groups, or students' preferences over schools could differ across racial groups. Distinguishing between these possibilities is important for guiding policy. If segregation arises due to the manner in which priority is assigned by the schools, it raises a worry that (potentially implicit) biases imposed by the school district may contribute. Detecting and correcting these biases could serve as an immediate path to remedy. If segregation arises due to different preferences, it suggests that interventions must take a different form. It also raises the possibility that some interventions could hurt the students they are meant to help by assigning them to less preferred schools.

To understand the importance of these different contributing factors, one needs access to information about students' priorities and students' preferences. With access to New York City's administrative data, a researcher could directly observe priorities and preference submissions. However, as of the time of Idoux's study, students could rank no more than 12 schools. Applying now-familiar results, this means that an analyst who assumes undominated preference submission can infer the preference ordering of those 12 schools, but this still provides relatively little guidance on their preferences over the full set of 450

[33]For an example of behavioral-economic structural analysis of experimental data, see Dreyfuss et al. (2022b). For an example of behavioral-economic structural analysis of field data, see Wang et al. (2021).

schools that could be ranked. More importantly, nearly all students submit a list of fewer than 12 schools—in line with the common "short list" problem discussed in Section 2.2.3. An analyst who assumes undominated preference reporting would have to infer that these students viewed all unranked schools as unacceptable, which can significantly bias preference estimation when schools are omitted for other reasons (Larroucau and Rios, 2020).

The modeling framework of Agarwal and Somaini (2018) provides a powerful approach to estimate full preferences under the assumption of optimal strategic preference submission. Idoux (2022) sought to apply this approach to recovering preferences, but was hindered by its computational demands in situations with many available schools, comparatively long preference lists, and the use of a single tie-breaking rule that drives correlation in admission outcomes.[34]

One might worry that the same factors that make this exercise challenging for a computer also make it challenging for the families of New York City middle schoolers. While it is compelling to assume that they are attempting to optimize, and perhaps are approximately optimizing, they may well be doing so in a manner with fewer computational requirements. Attempting to determine the maximization approach that they apply could therefore have benefits, not only by providing a more accurate model of how their submissions are determined, but also by providing a method that could be swiftly computationally executed.

Along these lines, Idoux develops a model in which applicants do not consider all possible preference lists. Modeled applicants instead follow a sequential process of consideration, starting with their top-ranked school, that does not fully internalize the impact of each sequential choice on continuation values.[35] This heuristic helps explain some patterns of apparent misoptimization in preferences submissions (including "short lists") and ultimately facilitates executing a variant of the Agarwal and Somaini (2018) strategy. The preference estimates generated from this approach suggest that the segregation observed in New York

[34]Larroucau and Rios (2020) provide guidance on implementing Agarwal and Somaini (2018) in the computationally simpler case where multiple, independent tiebreakers are used.

[35]This model falls in the class of theories describing failure to understand continuation values discussed in Section 3.1. The model of Idoux (2022) reflects an optimistic assessment of continuation values: assuming that they involve proceeding through a sequence of applications that is currently viewed as optimal, without accounting for how avoidance of application costs will dissuade future search. In contrast, the model of Wang et al. (2021) reflects a pessimistic assessment of continuation values: assuming that they involve immediately accepting an outside option.

City middle schools is determined by preferences and priorities to roughly similar degrees.

While behavioral economics is sometimes characterized as adding extra complications to already complicated exercises, examples like Idoux (2022) demonstrate that this is not always the case. We view this study as an illustration that behavioral-economic theories have potential to jointly improve a model's realism while also making it more computationally tractable. We explicitly do not encourage adopting behavioral-economic theories purely for their computational properties, without regard to whether their assumptions are in fact reasonable improvements on the status quo. However, identifying the cases in which well supported behavioral-economic models make estimation become tractable precisely leads us to the cases where behavioral economics may be most pivotal.

4.2 Influencing Choice of Mechanisms

Another way in which behavioral economics could inform education market design is by influencing our approach to choosing mechanisms. Conceptually, if we come to understand the features of mechanisms that cause behavior to deviate from the fully rational benchmark, that understanding can guide the mechanisms that we ultimately design. Most directly, this could involve attempts to make a mechanism maximally understandable or attempts to make preference reporting maximally truthful. But more broadly, the (not fully rational) behavioral responses induced by design decisions could be accounted for when assessing any other criteria used for mechanism selection. In this subsection we discuss several lines of research that pursue this line of reasoning.

4.2.1 Selecting Simple-to-Understand Mechanisms

Since the early days of education market design, market designers have attempted to deploy simple-to-understand mechanisms. To illustrate, consider the reform of the school assignment system in Boston—an initial reform that served as a beachhead for market designers' engagement with school choice. The process of this reform is thoroughly documented in Abdulkadiroğlu et al. (2006). In initial meetings with the Boston Public Schools (BPS) strategic planning team, Abdulkadiroğlu et al. emphasized the Boston mechanism's vulnerability to

manipulation. This led the committee to entertain the notion of switching to a strategy-proof alternative. Abdulkadiroğlu et al. presented two strategy-proof alternatives: the deferred acceptance mechanism and the top trading cycles mechanism. Assuming truthful preference reporting, the deferred acceptance mechanism has the advantage that it eliminates justified envy, whereas the top trading cycles mechanism has the advantage of Pareto efficiency. Due to concern about efficiency, a taskforce of community members initially recommended that the top trading cycles mechanism be adopted. In spite of this initial recommendation, BPS ultimately adopted the deferred acceptance mechanism. In a presentation explaining this choice, officials cited the lack of transparency of the top trading cycles mechanism:

> ... trading of priorities could lead families to believe they can still benefit from strategizing, as they may be encouraged to rank schools to which they have priority, even if they would not have put it on the form if the opportunity for trading did not exist. The behind the scenes mechanized trading makes the student assignment process less transparent. (BPS, 2005)

Similarly, Abdulkadiroğlu et al. (2006) quote the Boston Public Schools superintendent explaining that

> Top Trading Cycles is less transparent—and therefore more difficult to explain to parents—because of the trading feature executed by the algorithm, which may perpetuate the need or perceived need to "game the system."

Members of the school district appeared to believe that, while both the deferred acceptance mechanism and the top trading cycles mechanism were strategy-proof, *understanding* that the top trading cycles mechanism is strategy-proof is more difficult.[36] In a similar anecdote, Pathak (2016) cited the difficulty of explaining the top trading cycles mechanism as one reason why New Orleans replaced it with the deferred acceptance mechanism after only one year.

[36]The concerns expressed by these officials have support in the experimental findings of Chen and Sönmez (2006), which was used as part of Abdulkadiroğlu et al.'s persuasive appeal. Chen and Sönmez found lower rates of truthful reporting under the top trading cycles mechanism compared with the deferred acceptance mechanism.

As this account illustrates, making strategy-proofness easy to see was a clear design goal even in the earliest stages of the education market design literature. And yet, further progress in this dimension was slowed by the difficulty of formally defining the problem. Analytically characterizing what makes strategy-proofness easy to see is clearly challenging, and the idealized rational actor framework is not amenable to pursuing the question.

The clear need for such theories contributed to the positive reception of the "obvious strategy-proofness" framework of Li (2017). At its core, this framework reflects a straightforward approach to behaviorally informed market design: using insights from behavioral economics to analytically characterize the features of mechanisms that are easy to understand, and then treating implementation by a mechanism that holds such features as an explicit design goal. As we previously discussed in Section 3.1.1, when the dominant strategy is "obviously dominant," it can be identified as desirable even by an individual incapable of a certain type of contingent reasoning. If failures of contingent reasoning drive misunderstanding of mechanisms, OSP mechanisms may provide a solution.[37]

Unfortunately, while the OSP framework offers useful guidance on mechanism choice for auctions, its potential for application in the context of matching mechanisms is limited. As documented by Ashlagi and Gonczarowski (2018) and Thomas (2021), stable matching algorithms are not OSP-implementable outside of restrictive special cases with highly aligned priorities. Troyan (2019) shows that a similar result holds when considering OSP implementation of the top trading cycles algorithm. Assuming that we continue to honor the design goals that led to advocacy for stable algorithms or for the top trading cycles algorithm, these papers suggest that OSP mechanisms will be restricted to markets that can be organized by a serial dictatorship (or slight variations thereof). While serial dictatorship can at times be used in education market settings, it often cannot.

Bó and Hakimov (2020b) provide a relaxation of obvious strategy-proofness with greater applicability to matching markets. This approach is motivated by the observation that, in the domain of object allocation, OSP mechanisms can be presented in a particular form: agents sequentially choosing whether to pick an available object, rather than submitting rank

[37]However, Pycia and Troyan (2019) demonstrate that even OSP mechanisms can require complex forward-looking reasoning. They propose a refinement of obvious strategy-proofness that characterizes mechanisms that do not require capacity for future planning.

orderings that represent their preferences (Pycia and Troyan, 2019). Bó and Hakimov posit that this "pick an object" (PAO) form, rather than obvious strategy-proofness per se, drives much of the simplicity of these mechanisms. To support this claim, Bó and Hakimov note that dynamic implementations of the deferred acceptance mechanism—which have PAO form and are not OSP—lead to more truthtelling than the standard static implementation (Klijn et al., 2019; Bó and Hakimov, 2020a). This motivates their formal study of PAO mechanisms, and allocation rules that can be implemented as robust ordinal perfect Bayesian equilibria via PAO mechanisms. The set of such allocation rules is a strict superset of (non-bossy) OSP-implementable rules, and includes deferred acceptance and top trading cycles. This framework thus has immediate potential to influence the implementation of allocation rules in common use.

Interestingly, convergent support for sequential assignment procedures has begun to accumulate from a variety of sources, and with a variety of behavioral foundations not necessarily grounded in understanding.[38] Grenet et al. (2022) found that batched, sequential offers improve students' expected utility when students must incur costs to learn their own preferences. Meisner and von Wangenheim (2021) and Dreyfuss et al. (2022a) theoretically established that sequential, school-proposing deferred acceptance eliminates the expectations-management considerations that lead expectations-based loss-averse agents to misrepresent their preferences.

While a variety of rationales lead to some support for sequential mechanisms, the deep reason for their comparatively favorable elicitation of preferences is not yet fully understood. If sequential implementations help guide information acquisition, perhaps similar benefits could be achieved without changing the mechanism—for example, through information interventions (see Immorlica et al., 2020, for a related discussion). If sequential implementations help guide the consideration of relevant contingencies, it again may be possible to achieve similar benefits without changing the mechanism—for example, through explanations of

[38]Luflade (2018) provides potentially non-behavioral reasoning for valuing sequential implementation. In her study of the Tunisian university match, she found that its multi-round implementation helps students navigate the binding constraints placed on preference list length. Even fully rational agents could benefit from information on admissions chances that is revealed across rounds. However, Luflade additionally showed that part of the benefit of this procedure derives from students' biased subjective assessments of admissions probabilities.

mechanisms that accentuate relevant contingencies. The presentation of menu descriptions suggested by Gonczarowski et al. (2022) may achieve that goal.

As of the time of this writing, this literature is best viewed as a work in active progress. Theories of simplicity are proliferating rapidly, as are studies of their theoretical relation to behavioral-economic theories and their practical relation to the common mechanisms used in education market design. Guidance towards sequential implementations may already inform tradeoff considerations of mechanism choice. We look forward to seeing what other broadly supported guidance this literature produces as it continues its rapid development.

4.2.2 Guiding Counterfactual Analyses

Choices between different mechanisms often entail tradeoffs, and in many environments the welfare consequences of these tradeoffs are theoretically ambiguous. When theory alone does not dictate the desirable choice of mechanism, market designers commonly use existing markets as guiding case studies. In a given application, the market designer will only observe the allocation that arises from the specific mechanism that is used. However, under typical rationality assumptions—dictating a standard structure to preferences and optimal preference reporting—the preference reports in that market might be used to infer all market participants' true preferences. Maintaining rationality assumptions, the market designer may then calculate how market participants would submit preferences under alternative mechanisms. With access to these counterfactual preference reports, the market designer may then calculate the final allocations that would have been realized under the alternative systems. Following this procedure enables a comparison of the welfare that could be attained under the different candidate mechanisms.

In this standard procedure for counterfactual analysis, rationality assumptions play a critical role: they allow the market designer to infer true preferences and to subsequently forecast counterfactual behavior. Despite this critical role, rationality, per se, is not essential. If the market designer has access to an alternative behavioral model that is sufficiently well specified to make counterfactual forecasts, this model could be used to the same ends. If that alternative behavioral model more accurately accounts for behavior and welfare, the accuracy of the resulting counterfactual analysis could then be improved.

The potential for productive relaxation of rationality assumptions is well demonstrated in the work of Kapor et al. (2020). In this paper, the authors aimed to assess a central tradeoff considered in education market design. When choosing between the immediate and deferred acceptance mechanisms, a market designer may face a tradeoff between strategic simplicity and efficiency. The need to strategically manipulate preference reports in the immediate acceptance mechanism is often viewed as undesirable, but optimal manipulation strategies encode information about cardinal utilities that is not conveyed by truthful reporting in the deferred acceptance mechanism. Through this channel, the immediate acceptance mechanism can yield welfare gains in equilibrium (Abdulkadiroğlu et al., 2011).

Kapor et al. (2020) studied an application of the immediate acceptance mechanism in the New Haven high-school match, and wished to assess the welfare consequences of a transition to deferred acceptance. Their key behavioral-economic concern was inaccurate beliefs about admissions probabilities, as could be motivated by the theories of Section 3.3. Optimal preference reports under immediate acceptance depend on students' probabilities of assignment under different strategies. However, if subjective perceptions of these probabilities were misguided—as Kapor et al. document in a survey of participating families—then preference submissions predicated upon these perceptions would be suboptimal. Imposing an assumption of optimal preference reporting and correct probability assessments, Kapor et al. found *higher* welfare under immediate acceptance. When they relaxed the assumption of correct probability assessments, they determined that this initial conclusion is misleading: the welfare losses stemming from incorrect probability estimates led to *lower* welfare under immediate acceptance.

This paper provides a template for the relaxation of full rationality assumptions while conducting empirically guided counterfactual analyses. Put simply, the approach is to specify the candidate behavioral model, estimate its relevant components making use of both administrative data and auxiliary surveys, and then proceed with the standard template for counterfactual analysis taking the behavioral model as given.

While this paper focuses on the relaxation of a single element of the idealized rational model—the assumption of accurate probability assessments—its template for analysis could be applied to many other behavioral theories. In particular, we believe it could allow for

assessments of the welfare consequences of preference misrepresentation in strategy-proof environments. While we encourage this pursuit, we warn that doing such analyses well will be demanding. The welfare effects of preference misrepresentation depend critically on who misrepresents[39] and on the exact nature of their misrepresentation, and both of these dimensions are currently incompletely understood. What's more, calculating the welfare effect requires taking a normative stance on these preference misrepresentations. While many misrepresentations seem to be clear mistakes, some other types are debatable. For example, under the expectations-based reference dependence framework of Dreyfuss et al. (2022b), the researcher must specify whether reference dependence is a bias or a deep preference—an unresolved question within the behavioral economic literature.[40] For all of these reasons and more, principled welfare analysis involving behavioral-economic actors in strategy-proof markets is challenging to compellingly execute. However, examples like that in Kapor et al. (2020) provide encouragement that such challenges can indeed be overcome.

4.3 Guiding Interventions

Yet another way in which behavioral economics could inform education market design is by guiding us towards useful interventions in existing markets.

For much of recent history, large swaths of the behavioral economics literature have studied interventions designed to assist behavioral-economic agents. Principles for the design of such interventions were famously laid out in Thaler and Sunstein's "Nudge: Improving Decisions about Health, Wealth, and Happiness." As defined by Thaler and Sunstein, a *nudge* is an intervention that can help guide individuals to better decisions without materially changing incentives or restricting choices. Intervening in such a way is appealing because it minimally affects someone who optimally pursues their incentives, while at the same time assisting individuals behaving suboptimally. This satisfies a rough "do no harm" principle that appeals to Thaler and Sunstein's ideals as libertarian paternalists.

Recent literature in market design illustrates a productive role of nudges. We emphasize

[39] See Rees-Jones (2017) for a demonstration of this dependence.

[40] Once again, there is potential to learn from how similar issues have been handled within the behavioral public finance literature. The approach to modeling *normative ambiguity* advanced by Goldin and Reck (2018, 2022) could conceivably be usefully applied in this different domain.

two classes of such interventions that we believe have relatively broad utility and appeal.

One class of relevant nudges consists of simple, automated warnings that a participant is submitting suspicious preferences. To illustrate, consider an intervention that was deployed in the Israeli Psychology Master's Match.[41] Recall from Section 2.2.2 that participants in this match can rank programs with and without funding, and thus can reveal an *obvious misrepresentation* by indicating that they prefer admission without a scholarship to admission with a scholarship. Given the clear worry that these submissions can be costly mistakes, the system has been designed to dissuade them. Upon submitting a preference report that omits some contracts for a given program (e.g., failing to rank a program with funding despite ranking it without), a pop-up window appears. The pop-up draws attention to the omitted program and asks the participant to confirm whether they want to submit the ranking anyway. Interventions like these are easy to deploy and are minimally costly for optimizing students, but have some clear potential to help the students who need it. A similar intervention has since been deployed in the American Genetic Counseling Admissions Match (Peranson, 2019).

Another class of relevant nudges consists of simple communications about match probabilities. Recall from Section 3.3 that a large literature in behavioral economics predicts overconfidence in probability assessments, and that several papers have documented potentially related overoptimism about match probabilities. Concern about these issues led to the development of the "smart matching platform" that was studied in Arteaga et al. (2022) and deployed in the deferred-acceptance-based centralized school assignment system in Chile.[42] This intervention was targeted to applicants whose predicted risk of being unassigned (based on their applications portfolio and data from previous years) was over 30%. Such applicants are bearing enough risk that an analyst might suspect a mistake. These applicants received personalized information about the probability that they would end up being unassigned and were advised to add more schools to their portfolio. This approach combines several features that have been shown to be pivotal for the success of informational interventions including timely provision, computational burden reduction, and information from a trusted

[41] For details of the design of this match, see Hassidim et al. (2017b).

[42] They also examine a smaller scale (but still impressive) intervention in the New Haven public school match.

source (Mani et al., 2013; Fernandes et al., 2014; Dynarski et al., 2021). However, it also has the feature of being a minimal burden for a student who was knowingly accepting a high risk of remaining unmatched—such a student could simply ignore this nudge. In practice, this intervention was not ignored: it caused 22% of applicants to add schools to their application list, leading to a 58% decrease in nonplacement risk. Providing non-personalized risk information was approximately half as effective. Fabre et al. (2021) similarly finds that showing Chilean college applicants personalized information about admissions probabilities reduces their chance of nonplacement.

Interventions like these would be unnecessary for students who conform to the idealized rational actor model, and yet examples like those above demonstrate that they appear quite effective. Further development of behavioral models might guide us to other, similarly effective types of interventions. Pursuing this possibility, and formally testing interventions suggested by particular behavioral models, is a clear path forward in this literature. However, while we would expect such exercises to bear fruit, we also suspect that the most effective interventions will be ones that interact with broad ranges of models (like those summarized above), rather than ones narrowly tailored to very specific behavioral considerations.

4.4 Influencing our Understanding of the Process of Mechanism Adoption

Yet another way in which behavioral economics could inform education market design is by influencing our understanding of how groups come to adopt centralized mechanisms.

Market designers have long appreciated that influencing an existing market requires coming to understand the fundamental issues its organizers care about.[43] Making such a determination often benefits from examining the stated goals or mandates of the decisions makers. At times, however, it becomes clear that the mechanisms the decisionmakers have put in place are ill suited to their goals from the perspective of standard market design. This could suggest that the market organizers are themselves confused or subject to behavioral-economic considerations like those we have considered in this review.

[43]For a detailed guide to intervening in such markets, emphasizing the importance of determining the organizers' root concerns and communicating about them clearly and simply, see Sönmez (2023).

The literature on the deployment of centralized education markets is replete with anecdotes suggesting imperfect rationality among market organizers. Despite these anecdotes, this imperfect reasoning of market organizers has rarely been framed as a topic for formal and systematic academic study. A notable exception arises in the recent literature on "reserve systems." We review that literature to illustrate the potential of work along these lines.

4.4.1 An Example: Reserve Systems

Reserve systems are a means to advance the admission of a given group while preserving existing measures of priority. To illustrate, consider the reserve system that once existed in the Boston Public School (BPS) system, and that was studied by Dur et al. (2018). In 1999, BPS ended the use of racial and ethnic criteria in their assignment procedures. As the system was reformed, two opposing positions were pursued by different constituencies within the school district. One position was that the new system should feature unrestricted opportunity for families to choose their school. This would give the opportunity for students in underserved communities to choose to be schooled elsewhere, allowing some inequities to be mitigated. An opposing position was that the new system should cater to "neighborhood schooling." By filling schools with students from a fixed geographic area, this could potentially help encourage local community building and local investment in the school. As a compromise, the school district adopted a reserve system: they would apply a centralized matching procedure that respects students' preferences, but would reserve 50% of seats in each school for students living within the walk zone.

The degree of advantage that this system grants to walk-zone students depends on the manner in which reserves are processed. If the school first fills 50% of seats with the highest-priority walk-zone students, then fills the remaining 50% with the highest-priority remaining students, this approach functions as a *minimum guarantee*. Consider the case where walk-zone students' priority is such that they would fill fewer than 50% of seats without the reserve. In the first step, the school will fill 50% of their seats with the highest priority walk-zone students. This would include all walk-zone students who would be admitted without a reserve and some who would not be. In the second step, where any student could be considered, the set of remaining walk-zone students come from the bottom of the distribution of priorities,

and all would have insufficient priority for admission. In contrast, if walk-zone students' priority is such that they would fill more than 50% of seats without the reserve, the presence of the reserve has no effect and the same set of students is admitted. This configuration of the reserve thus guarantees a minimum number of walk-zone admissions, but has no additional function if the minimum does not bind. If the school instead first fills 50% of seats with the highest-priority students (regardless of walk-zone status), and then fills the remaining 50% with the highest-priority remaining walk-zone students, this approach has different effects. Rather than taking up reserve seats, the highest priority walk-zone students would now fill seats in the first, open-to-all, stage. The reserve seats are thus effectively reserved for lower priority walk-zone students. The consequence is that this configuration continues to advantage walk-zone students even if more than 50% would have been assigned without the reserve. Processing reserve seats last will always lead to weakly more walk-zone admissions than processing them first.

To economists trained in statistical reasoning, the line of logic above might seem natural. However, a variety of theories from behavioral economics suggest that the distinctions between these configurations could be misunderstood or missed entirely. Some of the difference in outcomes arises due to selection, which is often ignored by experimental subjects (Enke, 2020). Some of the difference in outcomes arises due to changes in the probability that an arbitrary student eligible for an unreserved seat is a walk-zone students, which could trigger problems for individuals subject to base-rate neglect (for a review, see Benjamin, 2019).

The potential for reserve systems to confuse some users now has substantial empirical support. In the BPS context, Dur et al. (2018) documented that the market organizers did not realize the importance of reserve order at the time the system was deployed. The programmer in charge of enacting the decision arbitrarily decided to implement it with reserved seats processed first. Dur et al. found that this resulted in a system that provided few additional seats to walk-zone students. The reserve system was implemented as a compromise between unrestricted choice and neighborhood schooling, but these findings revealed that the system was essentially no compromise at all. Shortly thereafter the system was again reformed to rectify this issue and replace the approach with something more "transparent." Similar stories, also with suggestive evidence of confusion, were later documented

in the histories of reforms to the U.S. H-1B visa reserves (Pathak et al., 2022a) and in the constitutionally mandated reserve systems in India (Sönmez and Yenmez, 2022).

While it seems natural to attribute much of the histories above to misunderstanding, that interpretation is admittedly somewhat speculative. To provide more direct evidence that reserve systems are misunderstood, Pathak et al. (2022b) presents a direct experimental evaluation of ability to process these systems' consequences. In an online experiment deployed to a representative survey of Americans, Pathak et al. present scenarios in which the subject must choose which of two mechanisms to adopt. The mechanisms differ in the number of seats reserved for their group, and in the order in which seats are processed. Subjects have a financial incentive to choose the system most favorable to their group. Their responses reveal some natural comparative statics—for example, the probability of choosing a mechanism increases as more seats are reserved—but there is little evidence that any subjects fully understand the role of processing order. Choice patterns suggest that approximately 40% of subjects think order is irrelevant. Such widespread misunderstanding of the role of processing order provides a natural explanation for the apparent confusion seen in the field.

4.4.2 Challenges and Promising Paths Forward

Studying the process by which mechanisms are chosen brings a different class of decisions into the purview of behavioral market design. Essentially all results discussed in this review pertain to the decisions made by applicants within a centralized market. The results of this section, however, pertain to individuals assessing different ways to organize a market. We view the rarity of formal behavioral study of such evaluations as a major hole in this literature, and a major opportunity for future research.

One challenge to pursuing this research is a strong and widespread intuition that the individuals choosing market designs are not subject to behavioral-economic biases. When considering the person making the decision, many researchers imagine an academic market designer like Al Roth, Parag Pathak, or Tayfun Sönmez (or at least someone who speaks with them and follows their advice). In reality, however, it is nearly always an individual or group without ready access to such expertise, without technical training in market design,

often ill-equipped to process formal market design theory, and often facing political-economy constraints to satisfy the desires of an uninformed and untrained constituency. Put simply, for behavioral economics to influence the way in which mechanisms are adopted in education markets, one does not need to believe that Al Roth is "behavioral." Rather, one just needs to believe that members of the school board may be behavioral, or that the parents in their constituencies that exert influence on the process are behavioral. In short, while conflicting intuitions are common, we think that studying the behavioral economics of mechanism choice is well justified and welfare-relevant.

A challenge to experimentally pursuing these questions is ensuring that one studies the relevant population. Behavioral-economic lab experiments are often criticized because the biases revealed by student subjects may not be held by "real" decisionmakers. This criticism is slightly less worrying than usual when the study is intended to measure how students interact with school-choice mechanism: in some sense, the typical lab experimental population is well suited for this purpose. For studies relevant to how mechanisms are chosen, these lab populations are much farther from the ideal. The best versions of such studies likely involve lab-in-field approaches—for example, deploying one's study directly to principals, school board members, or parents in a parent-teacher association. While these are harder to deploy than lab experiments, the rapid transition of lab experimental economics out of the lab and onto online platforms makes such studies easier than they used to be. We strongly encourage such work going forward, and hope to do some ourselves.

5 Conclusion

The literature studying the design of centralized education markets is an unambiguous success story of economics. It has yielded a long line of research that developed substantive advances in both theory and empirics, and accordingly the pioneers of this area have been granted some of the professions' highest honors. More importantly, the field has had tremendous direct and positive impact on the world.

Members of this field have long shown concern about the ability of market participants to optimally engage with school-choice algorithms. Supporting that concern, the literature

reviewed in this paper makes it clear that behavior violating the fully rational model is commonplace. This provides a clear point-of-entry for behavioral economists, and in recent years behavioral economists have capitalized on this opportunity. It is now clear that educational market design provides an interesting environment for the development of behavioral-economic theories. What remains to be seen is whether those theories and the insights they generate will be deeply influential to the conduct of "standard" market design. We think there is great promise, and we hope the paths forward that we have detailed in this paper help it to be achieved.

References

Abdulkadiroğlu, Atila, Yeon-Koo Che, and Yosuke Yasuda (2011): "Resolving Conflicting Preferences in School Choice: The "Boston Mechanism" Reconsidered," *American Economic Review*, 101(1), 399–410.

Abdulkadiroğlu, Atila, Parag A. Pathak, and Alvin E. Roth (2005a): "The New York City High School Match," *American Economic Review, Papers and Proceedings*, 95, 364–367.

Abdulkadiroğlu, Atila, Parag A. Pathak, Alvin E. Roth, and Tayfun Sönmez (2005b): "The Boston Public School Match," *American Economic Review, Papers and Proceedings*, 95, 368–371.

——— (2006): "Changing the Boston School Choice Mechanism," NBER Working Paper 11965.

Abdulkadiroğlu, Atila and Tayfun Sönmez (2003): "School Choice: A Mechanism Design Approach," *American Economic Review*, 93, 729–747.

Agarwal, Nikhil and Paulo Somaini (2018): "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism," *Econometrica*, 86, 391–444.

——— (2020): "Revealed Preference Analysis of School Choice Models," *Annual Review of Economics*, 12, 471–501.

Ajayi, Kehinde (2014): "Does School Quality Improve Student Performance? New Evidence from Ghana," Working Paper.

Ajayi, Kehinde and Modibo Sidibe (2022): "School Choice Under Imperfect Information," Working Paper.

Ali, S. Nageeb and Ran I. Shorrer (2022): "The College Portfolio Problem," Working Paper.

Allcott, Hunt, Benjamin B. Lockwood, and Dmitry Taubinsky (2019): "Regressive Sin Taxes, with an Application to the Optimal Soda Tax," *The Quarterly Journal of Economics*, 134, 1557–1626.

Allcott, Hunt and Dmitry Taubinsky (2015): "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market," *American Economic Review*, 105, 2501–38.

Arteaga, Felipe, Adam J. Kapor, Christopher A. Neilson, and Seth D. Zimmerman (2022): "Smart Matching Platforms and Heterogeneous Beliefs in Centralized School Choice," *The Quarterly Journal of Economics*, 137, 1791–1848.

Artemov, Georgy (2021): "Assignment Mechanisms: Common Preferences and Information Acquisition," *Journal of Economic Theory*, 198, 105370.

Artemov, Georgy, Yeon-Koo Che, and Yinghua He (2020): "Strategic 'Mistakes': Implications for Market Design Research," Working Paper.

Artemov, Georgy, Yeon-Koo Che, and YingHua He (2022): "Stable Matching with Mistaken Agents," *Journal of Political Economy Microeconomics, forthcoming*.

Ashlagi, Itai and Yannai A. Gonczarowski (2018): "Stable Matching Mechanisms are not Obviously Strategy-Proof," *Journal of Economic Theory*, 177, 405–425.

Aygün, Orhan and Bertan Turhan (2017): "Large-Scale Affirmative Action in School Choice: Admissions to IITs in India," *American Economic Review*, 107, 210–13.

Azevedo, Eduardo M. and Jacob D. Leshno (2016): "A Supply and Demand Framework for Two-Sided Matching Markets," *Journal of Political Economy*, 124(5), 1235–1268.

Bade, Sophie (2015): "Serial Dictatorship: The Unique Optimal Allocation Rule when Information is Endogenous," *Theoretical Economics*, 10, 385–410.

Balinski, Michel and Tayfun Sönmez (1999): "A Tale of Two Mechanisms: Student Placement," *Journal of Economic Theory*, 84, 73–94.

Basteck, Christian and Marco Mantovani (2018): "Cognitive Ability and Games of School Choice," *Games and Economic Behavior*, 109, 156 – 183.

——— (2022): "Aiding Applicants: Leveling the Playing Field Within the Immediate Acceptance Mechanism," *Review of Economic Design*, 1–34.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak (1964): "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, 9, 226–232.

Benjamin, Daniel J. (2019): "Errors in Probabilistic Reasoning and Judgment Biases," in *Handbook of Behavioral Economics: Applications and Foundations*, North-Holland, vol. 2, chap. 2, 69–186.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones (2014): "Can Marginal Rates of Substitution be Inferred from Happiness Data? Evidence from Residency Choices," *American Economic Review*, 104, 3498–3528.

Bernheim, B. Douglas and Dmitry Taubinsky (2018): "Behavioral Public Economics," in *Handbook of Behavioral Economics: Applications and Foundations*, North-Holland, vol. 1, chap. 5, 381–516.

Bó, Inácio and Rustamdjan Hakimov (2020a): "Iterative versus Standard Deferred Acceptance: Experimental Evidence," *The Economic Journal*, 130, 356–392.

——— (2020b): "Pick-an-Object Mechanisms," *arXiv preprint arXiv:2012.01025*.

Bobba, Matteo and Veronica Frisancho (2022): "Self-Perceptions about Academic Achievement: Evidence from Mexico City," *Journal of Econometrics*, 231, 58–73.

BPS (2005): "Recommendation to Implement a New BPS Assignment Algorithm," Presentation to the Boston School Committee by Carleton Jones, May 11.

Bucher, Stefan F. and Andrew Caplin (2021): "Inattention and Inequity in School Matching," NBER Working Paper 29586.

Calsamiglia, Caterina, Guillaume Haeringer, and Flip Klijn (2010): "Constrained School Choice: An Experimental Study," *American Economic Review*, 100, 1860–74.

Cason, Timothy N. and Charles R. Plott (2014): "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing," *Journal of Political Economy*, 122, 1235–1270.

Chade, Hector and Lones Smith (2006): "Simultaneous Search," *Econometrica*, 74, 1293–1307.

Chandra, Amitabh, Benjamin Handel, and Joshua Schwartzstein (2019): "Behavioral Economics and Health-Care Markets," in *Handbook of Behavioral Economics: Applications and Foundations*, North-Holland, vol. 2, chap. 6, 459–502.

Chen, Li and Juan Sebastián Pereyra (2019): "Self-Selection in School Choice," *Games and Economic Behavior*, 117, 59–81.

Chen, Yan, Peter Cramton, John A. List, and Axel Ockenfels (2021): "Market Design, Human Behavior, and Management," *Management Science*, 67, 5317–5348.

Chen, Yan and YingHua He (2021): "Information Acquisition and Provision in School Choice: An Experimental Study," *Journal of Economic Theory*, 197, 105345.

——— (2022): "Information Acquisition and Provision in School Choice: A Theoretical Investigation," *Economic Theory*, 74, 293–327.

Chen, Yan, Yingzhi Liang, and Tayfun Sönmez (2016): "School Choice Under Complete Information: An Experimental Study," *The Journal of Mechanism and Institution Design*, 1, 45–82.

Chen, Yan and Tayfun Sönmez (2006): "School Choice: An Experimental Study," *Journal of Economic Theory*, 127, 202–231.

Chetty, Raj (2009): "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Review of Economics*, 1, 451–488.

Chetty, Raj, Adam Looney, and Kory Kroft (2009): "Salience and Taxation: Theory and Evidence," *American Economic Review*, 99, 1145–77.

Ding, Tingting and Andrew Schotter (2017): "Matching and Chatting: An Experimental Study of the Impact of Network Communication on School-Matching Mechanisms," *Games and Economic Behavior*, 103, 94–115.

——— (2019): "Learning and Mechanism Design: An Experimental Test of School Matching Mechanisms with Intergenerational Advice," *The Economic Journal*, 129, 2779–2804.

Dreyfuss, Bnaya, Ofer Glicksohn, Ori Heffetz, and Assaf Romm (2022a): "Deferred Acceptance with News Utility," NBER Working Paper 30635.

Dreyfuss, Bnaya, Ori Heffetz, and Matthew Rabin (2022b): "Expectations-Based Loss Aversion May Help Explain Seemingly Dominated Choices in Strategy-Proof Mechanisms," *American Economic Journal: Microeconomics*, 14, 515–55.

Dubins, Lester E. and David A. Freedman (1981): "Machiavelli and the Gale-Shapley Algorithm," *American Mathematical Monthly*, 88, 485–494.

Dur, Umut, Scott D. Kominers, Parag A. Pathak, and Tayfun Sönmez (2018): "Reserve Design: Unintended Consequences and the Demise of Boston's Walk Zones," *Journal of Political Economy*, 126, 2457–2479.

Dynarski, Susan, CJ Libassi, Katherine Michelmore, and Stephanie Owen (2021): "Closing the Gap: The Effect of Reducing Complexity and Uncertainty in College Pricing on the Choices of Low-income Students," *American Economic Review*, 111, 1721–56.

Echenique, Federico, Ruy González, Alistair J. Wilson, and Leeat Yariv (2022): "Top of the Batch: Interviews and the Match," *American Economic Review: Insights*, 4, 223–38.

Enke, Benjamin (2020): "What You See Is All There Is," *The Quarterly Journal of Economics*, 135, 1363–1398.

Enke, Benjamin and Florian Zimmermann (2019): "Correlation Neglect in Belief Formation," *The Review of Economic Studies*, 86, 313–332.

Fabre, Anaïs, Tomás Larroucau, Manuel Martinez, Christopher Neilson, and Ignacio Rios (2021): "Application Mistakes and Information Frictions in College Admissions," Working Paper.

Fack, Gabrielle, Julien Grenet, and Yinghua He (2019): "Beyond Truth-Telling: Preference Estimation with Centralized School Choice and College Admissions," *American Economic Review*, 109, 1486–1529.

Farhi, Emmanuel and Xavier Gabaix (2020): "Optimal Taxation with Behavioral Agents," *American Economic Review*, 110, 298–336.

Featherstone, Clayton R. and Muriel Niederle (2016): "Boston versus Deferred Acceptance in an Interim Setting: An Experimental Investigation," *Games and Economic Behavior*, 100, 353–375.

Fernandes, Daniel, John G Lynch Jr, and Richard G Netemeyer (2014): "Financial Literacy, Financial Education, and Downstream Financial Behaviors," *Management science*, 60, 1861–1883.

Gabaix, Xavier, David Laibson, Guillermo Moloche, and Stephen Weinberg (2006): "Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model," *American Economic Review*, 96, 1043–1068.

Gale, David and Lloyd S. Shapley (1962): "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 69, 9–15.

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman (2002): *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press.

Goldin, Jacob and Daniel Reck (2018): "Rationalizations and Mistakes: Optimal Policy with Normative Ambiguity," *AEA Papers and Proceedings*, 108, 98–102.

——— (2022): "Optimal Defaults with Normative Ambiguity," *The Review of Economics and Statistics*, 104, 17–33.

Golman, Russell, David Hagmann, and George Loewenstein (2017): "Information Avoidance," *Journal of Economic Literature*, 55, 96–135.

Gonczarowski, Yannai A., Ori Heffetz, and Clayton Thomas (2022): "Strategyproofness-Exposing Mechanism Descriptions," *arXiv preprint arXiv:2209.13148*.

Grenet, Julien, YingHua He, and Dorothea Kübler (2022): "Preference Discovery in University Admissions: The Case for Dynamic Multioffer Mechanisms," *Journal of Political Economy*, 130, 1427–1476.

Gross, Betheny, Michael DeArmond, and Patrick Denice (2015): "Common Enrollment, Parents, and School Choice: Early Evidence from Denver and New Orleans." *Center on Reinventing Public Education*.

Guillen, Pablo and Rustamdjan Hakimov (2017): "Not Quite the Best Response: Truth-telling, Strategy-proof Matching, and the Manipulation of Others," *Experimental Economics*, 20, 670–686.

——— (2018): "The Effectiveness of Top-Down Advice in Strategy-Proof Mechanisms: A Field Experiment," *European Economic Review*, 101, 505–511.

Guillen, Pablo and Alexander Hing (2014): "Lying Through Their Teeth: Third Party Advice and Truth Telling in a Strategy Proof Mechanism," *European Economic Review*, 70, 178–185.

Guillen, Pablo and Róbert F Veszteg (2021): "Strategy-proofness in Experimental Matching Markets," *Experimental Economics*, 24, 650–668.

Haeringer, Guillaume and Flip Klijn (2009): "Constrained School Choice," *Journal of Economic Theory*, 144, 1921–1947.

Hakimov, Rustamdjan and Dorothea Kübler (2021): "Experiments on Centralized School Choice and College Admissions: a Survey," *Experimental Economics*, 24, 434–488.

Harrison, Glenn and Kevin McCabe (1996): "Stability and Preference Distortion in Resource Matching: An Experimental Study of the Marriage Problem," in *Research in Experimental Economics*, ed. by Vernon Smith, JAI Press, 53–129.

Hassidim, Avinatan, Déborah Marciano, Assaf Romm, and Ran I. Shorrer (2017a): "The Mechanism Is Truthful, Why Aren't You?" *American Economic Review*, 107, 220–24.

Hassidim, Avinatan, Assaf Romm, and Ran I. Shorrer (2017b): "Redesigning the Israeli Psychology Master's Match," *American Economic Review*, 107, 205–09.

——— (2021): "The Limits of Incentives in Economic Matching Procedures," *Management Science*, 67, 951–963.

Idoux, Clémence (2022): "Integrating New York City Schools: The Role of Admission Criteria and Family Preferences," Discussion Paper 2022.14, Blueprint Labs.

Immorlica, Nicole, Jacob Leshno, Irene Lo, and Brendan Lucier (2020): "Information Acquisition in Matching Markets: The Role of Price Discovery," SSRN Working Paper 3705049.

Kagel, John and Alvin E. Roth (2000): "The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment," *The Quarterly Journal of Economics*, 115, 201–235.

Kahneman, Daniel and Amos Tversky (1979): "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 47, 263–91.

Kapor, Adam J., Christopher A. Neilson, and Seth D. Zimmerman (2020): "Heterogeneous Beliefs and School Choice Mechanisms," *American Economic Review*, 110, 1274–1315.

Klijn, Flip, Joana Pais, and Marc Vorsatz (2019): "Static versus Dynamic Deferred Acceptance in School Choice: Theory and Experiment," *Games and Economic Behavior*, 113, 147–163.

Kloosterman, Andrew and Peter Troyan (2022): "Rankings-Dependent Preferences: A Real Goods Matching Experiment," Working Paper.

Koutout, Kristine, Andrew Dustan, Martin Van der Linden, and Myrna Wooders (2021): "Mechanism Performance Under Strategy Advice and Sub-Optimal Play: A School Choice Experiment," *Journal of Behavioral and Experimental Economics*, 94, 101755.

Kőszegi, Botond (2006): "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, 4, 673–707.

Kőszegi, Botond and Matthew Rabin (2006): "A Model of Reference-Dependent Preferences," *The Quarterly Journal of Economics*, 121, 1133–1166.

——— (2007): "Reference-Dependent Risk Attitudes," *American Economic Review*, 97, 1047–1073.

——— (2009): "Reference-Dependent Consumption Plans," *American Economic Review*, 99, 909–36.

Larroucau, Tomás and Ignacio Rios (2020): "Do "Short-List" Students Report Truthfully? Strategic Behavior in the Chilean College Admissions Problem," Working Paper.

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

Lucas, Adrienne M. and Isaac M. Mbiti (2012): "The Determinants and Consequences of School Choice Errors in Kenya," *American Economic Review*, 102, 283–88.

Luflade, Margaux (2018): "The Value of Information in Centralized School Choice Systems," Working Paper.

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao (2013): "Poverty Impedes Cognitive Function," *science*, 341, 976–980.

Meisner, Vincent (2022): "Report-Dependent Utility and Strategy-Proofness," *Management Science, forthcoming*.

Meisner, Vincent and Jonas von Wangenheim (2021): "School Choice and Loss Aversion," CESifo Working Paper 9479.

Meisner, Vincent and Jonas von Wangenheim (2023): "Loss Aversion in Strategy-Proof School-Choice Mechanisms," *Journal of Economic Theory*, 207, 105588.

Moore, Don A. and Paul J. Healy (2008): "The Trouble with Overconfidence," *Psychological Review*, 115, 502.

Niederle, Muriel and Emmanuel Vespa (2022): "Cognitive Reasoning: Failures of Contingent Thinking," Working Paper.

Noda, Shunya (2022): "Strategic Experimentation with Random Serial Dictatorship," *Games and Economic Behavior*, 133, 115–125.

O'Donoghue, Ted and Charles Sprenger (2018): "Reference-Dependent Preferences," in *Handbook of Behavioral Economics: Applications and Foundations*, North-Holland, vol. 1, chap. 2, 1–77.

Pais, Joana and Ágnes Pintér (2008): "School Choice and Information: An Experimental Study on Matching Mechanisms," *Games and Economic Behavior*, 64, 303–328.

Pais, Joana, Ágnes Pintér, and Róbert F. Veszteg (2011): "College Admissions and the Role of Informaton: An Experimental Study," *International Economic Review*, 52, 713–737.

Pan, Siqi (2019): "The Instability of Matching with Overconfident Agents," *Games and Economic Behavior*, 113, 396 – 415.

Pathak, Parag A. (2016): "What Really Matters in Designing School Choice Mechanisms," Advances in Economics and Econometrics, 11th World Congress of the Econometric Society.

Pathak, Parag A., Alex Rees-Jones, and Tayfun Sönmez (2022a): "Immigration Lottery Design: Engineered and Coincidental Consequences of H-1B Reforms," *Review of Economics and Statistics, forthcoming.*

——— (2022b): "Reversing Reserves," *Management Science, forthcoming.*

Pathak, Parag A. and Tayfun Sönmez (2008): "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism," *American Economic Review*, 98(4), 1636–1652.

——— (2013): "School Admissions Reform in Chicago and England: Comparing Mechanisms by their Vulnerability to Manipulation," *American Economic Review*, 103(1), 80–106.

Peranson, Jonah (2019): "Design and Implementation of the Genetic Counseling Admissions Match," in *Proceedings of MATCH-UP 2019, 5th International Workshop on Matching Under Preferences*, Match-UP, 21.

Pycia, Marek and Peter Troyan (2019): "A Theory of Simplicity in Games and Mechanism Design," CEPR Discussion Paper No. DP14043.

Rees-Jones, Alex (2017): "Mistaken Play in the Deferred Acceptance Algorithm: Implications for Positive Assortative Matching," *American Economic Review*, 107, 225–29.

——— (2018a): "Quantifying Loss-Averse Tax Manipulation," *The Review of Economic Studies*, 85, 1251–1278.

——— (2018b): "Suboptimal Behavior in Strategy-Proof Mechanisms: Evidence from the Residency Match," *Games and Economic Behavior*, 108, 317 – 330.

Rees-Jones, Alex, Ran I. Shorrer, and Chloe Tergiman (2020): "Correlation Neglect in Student-to-School Matching," NBER Working Paper 26734.

Rees-Jones, Alex and Samuel Skowronek (2018): "An Experimental Investigation of Preference Misrepresentation in the Residency Match," *Proceedings of the National Academy of Sciences*, 115, 11471–11476.

Rees-Jones, Alex and Dmitry Taubinsky (2020): "Measuring "Schmeduling"," *The Review of Economic Studies*, 87, 2399–2438.

Roth, Alvin E. (1982): "The Economics of Matching: Stability and Incentives," *Mathematics of Operations Research*, 7, 617–628.

——— (1984): "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of Political Economy*, 92, 991–1016.

——— (2002): "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics," *Econometrica*, 70, 1341–1378.

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

Roth, Alvin E. and Xiaolin Xing (1994): "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions," *American Economic Review*, 84, 992–1044.

Shapley, Lloyd and Herbert Scarf (1974): "On Cores and Indivisibility," *Journal of Mathematical Economics*, 1, 23–28.

Sharot, Tali and Cass R Sunstein (2020): "How People Decide What They Want to Know," *Nature Human Behaviour*, 4, 14–19.

Shorrer, Ran I. and Sándor Sóvágó (2022): "Dominated Choices in a Strategically Simple College Admissions Environment," Working Paper.

——— (2023): "Dominated Choices under Deferred Acceptance Mechanism: The Effect of Admission Selectivity," Working Paper.

Slovic, Paul (1995): "The Construction of Preference." *American psychologist*, 50, 364.

Sönmez, Tayfun (2023): "Minimalist Approach to Institution Redesign: A Framework for an Aspiring Market Designer," Working Paper.

Sönmez, Tayfun and M. Bumin Yenmez (2022): "Affirmative Action in India via Vertical, Horizontal, and Overlapping Reservations," *Econometrica*, 90, 1143–1176.

Taubinsky, Dmitry and Alex Rees-Jones (2018): "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment," *The Review of Economic Studies*, 85, 2462–2496.

Thaler, Richard H. and Cass R. Sunstein (2009): *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Penguin.

Thomas, Clayton (2021): "Classification of Priorities Such That Deferred Acceptance is OSP Implementable," in *Proceedings of the 22nd ACM Conference on Economics and Computation*, New York, NY, USA: Association for Computing Machinery, EC '21, 860.

Troyan, Peter (2019): "Obviously Strategy-Proof Implementation of Top Trading Cycles," *International Economic Review*, 60, 1249–1261.

Wang, Ao, Shaoda Wang, and Xiaoyang Ye (2021): "Cognitive Distortions in Complex Decisions: Evidence from Centralized College Admission," Working Paper.