

Stable Matching with Mistaken Agents*

Georgy Artemov[†]

Yeon-Koo Che[‡]

YingHua He[§]

April 28, 2022

Abstract

Motivated by growing evidence of agents’ mistakes in strategically simple environments, we propose a solution concept—robust equilibrium—that requires only an asymptotically optimal behavior. We apply it to study large random matching markets operated by the applicant-proposing Deferred Acceptance (DA). Although truth-telling is a dominant strategy, almost all applicants may be non-truthful in robust equilibrium; however, the outcome must be arbitrarily close to the stable matching with truth-telling agents. Our results imply that one can assume truthful agents to study DA outcomes, theoretically or counterfactually. However, to estimate preferences of mistaken agents, one should assume stable matching but not truth-telling.

JEL Classification Numbers: C70, D47, D61, D63.

Keywords: Strategic mistakes, payoff relevance of mistakes, robust equilibria, truth-telling, stable-response strategy, stable matching.

*This paper partially subsumes “Strategic Mistakes: Implications for Market Design Research” by the same authors. We thank Xingye Wu, who has provided excellent research assistance for the theory part, and Julien Grenet for his generous help with the Monte Carlo simulations. We are grateful to the seminar/conference participants at ANU, Asia-Pacific IO Conference, ASSA Meeting, Barcelona GSE Summer Forum, Boston College, Deakin, Duke, Columbia, Conference on Economic Design, “Econometrics Meets Theory” Conference at NYU, European Meeting on Game Theory, “Dynamic Models in Economics” Workshop at NUS, Hitotsubashi, Higher School of Economics, Game Theory and Management Conference, MIT, NBER Market Design Group Meeting, PET Conference, Workshop “Matching in Practice”, Paris School of Economics, Stony Brook Conference on Game Theory, UC Irvine, University of Queensland, and Waseda for their comments. Authors acknowledge support from the Australian Research Council (DP160101350) and the University of Melbourne (Artemov); National Research Foundation of Korea (NRF-2020S1A5A2A03043516, Che); National Science Foundation (SES-1851821, Che; SES-1730636, He).

[†]Department of Economics, University of Melbourne, Australia. Email: georgy@gmail.com

[‡]Department of Economics, Columbia University, USA. Email: yeonkooche@gmail.com.

[§]Department of Economics, Rice University, USA. Email: yinghua.he@rice.edu.

1 Introduction

Strategy-proofness—or making it a dominant strategy to truthfully reveal one’s own preferences—is an important desideratum in market design (Abdulkadiroglu and Sonmez, 2003). It makes straightforward for an agent to act in one’s best interest, thus minimizing the scope for mistakes. It also equalizes the playing field by eliminating the need for strategizing. Furthermore, strategy-proofness aids empirical research by making agents’ choices easy to interpret.

However, there is growing evidence that strategic mistakes are common, even in strategy-proof environments.¹ Laboratory experiments (see, e.g., Chen and Sönmez, 2002) and studies of high-stake real-life mechanisms (Chen and Pereyra, 2019; Hassidim, Romm, and Shorrer, 2020; Artemov, Che, and He, 2020; Shorrer and Sóvágó, 2018; Rees-Jones, 2017) show that a significant fraction of participants mis-report their preferences in a commonly used strategy-proof mechanism, the applicant-proposing Deferred Acceptance (DA).

Such mistakes may pose a challenge to the market-design research, both theoretical and empirical. Most theoretical studies on strategy-proof mechanisms assume that agents play their unique dominant strategy: truth-telling (TT). Do the documented mistakes, i.e., non-truthful behaviors, imply that the theoretical predictions in the literature are incorrect? Moreover, much of the empirical literature relies on similar assumptions that ignore mistakes. Does it mean that the estimates from that literature are biased? These are the questions that our paper aims to answer.

We study large economies operated by DA and take centralized college admissions as our leading example. Each college strictly ranks applicants by some score, and every applicant knows her own score before submitting a rank-order list (ROL) of colleges. A wide variety of markets in the field operate in this fashion; for example, the aforementioned papers study admissions to secondary schools, universities, or post-graduate programs in four different countries.

Our equilibrium concept is motivated by the empirical finding that, although mistakes are frequent, only a small fraction of them have any payoff consequences. In general, mistakes are difficult to identify in the field because applicant preferences are unknown to researchers. Recent studies (Artemov, Che, and He, 2020; Hassidim, Romm, and Shorrer, 2020; Shorrer and Sóvágó, 2018), which are summarized in Table 1, focus on pairs of education programs that differ only in a financial component (e.g., scholarship vs. no scholarship) and hence whose preference order can be unambiguously determined. In these studies, 17–35% of applicants make a mistake: when reporting their ordinal preferences to the mechanism, they rank a program *without* a scholarship *above* the identical program with a scholarship (column 3). However, among the applicants with an identified mistake, only 1–20% would have a different assignment if each applicant’s mistake is corrected

¹In keeping with the literature, we use the term “mistake” to refer to the play of a *dominated strategy*, regardless of whether it entails an actual payoff loss (which depends on other individuals’ actions). When a mistake does lead to a payoff loss, we say it is a *payoff-relevant* mistake.

unilaterally (columns 5 and 7).

Table 1: Mistakes and Payoff-relevance across Different Studies

	Size of the relevant sample (1)	Identified mistakes		Payoff-relevant mistakes			
		Freq. (2)	share: $\frac{(2)}{(1)}$ (3)	upper bound		lower bound	
				Freq. (4)	share: $\frac{(4)}{(2)}$ (5)	Freq. (6)	share: $\frac{(6)}{(2)}$ (7)
Artemov, Che, and He (2020) College admissions in Australia	2,915	1,009	35%	201	20%	14	1%
Hassidim, Romm, and Shorrer (2020): Graduate admissions in Israel	672	130	19%	10	8%	3	2%
Shorrer and Sóvágó (2018) College admissions in Hungary	92,777	15,653	17%	1,479	9%	669	4%

Notes: All studies identify instances where an applicant appears to prefer an education program without a scholarship to the same program with a scholarship (i.e., she ranks the former higher than the latter in her application or applies only to the former). We call these instances *identifiable mistakes*. Column (1) is the number of applicants that can possibly make an identifiable mistake. Columns (2) and (3) show the number of applicants who made an identifiable mistake. If unilaterally correcting an applicant’s mistake leads to a different outcome for the applicant, the mistake is payoff-relevant. As evaluating payoff-relevance requires knowledge of true applicant preferences, these studies find an upper (columns 4 and 5) and a lower (columns 6 and 7) bounds.

Motivated by such an empirical pattern, our *robust equilibrium* allows for mistakes with small payoff consequences. We study a sequence of economies, in which applicant types are randomly drawn, the number of applicants and the total number of seats grow proportionally, and the number of colleges is fixed. A robust equilibrium requires that payoff consequences of mistakes vanish when the economy grows large.

Our first main result shows that a dramatic departure from TT—all but a vanishing fraction of applicants submitting untruthful ROLs—is supported as a robust equilibrium (Theorem 1). This is consistent with the behavior reported in Table 1, where applicants flip the order of some colleges in their ROL or omit “out-of-reach” colleges even when they can include more colleges in an ROL at no cost.

Nevertheless, as shown in our second main result (Theorem 2), not all is lost due to non-truthful behavior. Despite the behavioral multiplicity and ambiguity, under mild conditions, *all* robust equilibria yield a virtually unique outcome for a sufficiently large economy. Equilibrium outcomes of the sequence of economies are asymptotically stable and converge to the outcome that would arise from truth-telling behavior.

Asymptotic stability requires that, with probability approaching one as the economy grows, an applicant is assigned to her best feasible college. A college is feasible to an applicant if her score is above the college’s cutoff, which is the lowest score among the college’s accepted applicants. While allowing for mistakes means that stability is no longer guaranteed, Theorem 2 suggests that virtually all applicants obtain their best feasible college.

At first glance, asymptotic stability may appear to be an unsurprising consequence of a robust equilibrium. For example, one may conjecture that, as payoff losses vanish,

fewer and fewer applicants suffer a loss. However, a robust equilibrium allows that an arbitrarily large number of applicants suffer a vanishing loss and do not obtain their best feasible college, contradicting asymptotic stability. One may further conjecture that, in large enough economies, the applicants are “price-takers:” the cutoffs are almost certain and virtually unaffected by an applicant’s deviation. Hence, for most applicants who suffer a loss, the loss must be discrete. A vanishing loss then implies that nearly all applicants must obtain their best feasible college in a robust equilibrium. However, we show in Example 1 that the price-taking assumption may not hold: in an arbitrarily large economy, a unilateral deviation of an applicant can lead to significant changes in cutoffs.

The price-taking behavior is restored—at least to the extent necessary for our results—in economies where applicants are randomly sampled from a full-support distribution and report truthfully with some (arbitrarily low) probability (Propositions 1 and 2). The propositions use the observation that, even though cutoffs are not well-behaved, the demands for individual colleges respond smoothly to a deviation. Further, with random sampling, if the change in demand is small in expectation, the change in cutoffs must be small, delivering the results. The propositions are the key to proving Theorems 1 and 2. As the economy grows, the uncertainty about cutoffs vanishes and the set of feasible colleges become apparent to the applicants. For their ROL, applicants then know which colleges they can safely omit (hence Theorem 1) and which colleges they must include (hence Theorem 2).

Our Theorems 1 and 2 are reassuring news for the theoretical literature on DA. To the extent that outcomes are more important than applicant behavior, the existing results that rely on applicants’ truth-telling behavior are largely robust to applicants’ mistakes, at least in large economies.

Our results also yield important implications for empirical research. Strategy-proofness is sometimes taken literally in interpreting applicants’ ROLs. Specifically, one assumes that applicants are weakly truth-telling (WTT); see, for example, Hällsten (2010) and Kirkebøen (2012). WTT hypothesizes that an applicant ranks her most-preferred colleges truthfully, but may not rank all acceptable colleges. Theorem 1 calls such an approach into question. When an applicant omits a more-preferred out-of-reach college, WTT infers that that college is less preferred than any college listed in the applicant’s ROL, leading to biased estimates. At the same time, an alternative approach that assumes stability of the matching is justified in large enough economies by Theorem 2. This approach only makes inference about feasible colleges, but “refuses” to infer any preferences over others. We illustrate the empirical implications of our theorems in Monte Carlo simulations. WTT estimates have low variance, but they are substantially biased when applicants omit colleges with which they are never matched.

Even though our results advise caution in relying on TT for preference estimation, they support using TT for the counterfactual analysis of policies. That is, our Theorem 2 justifies the approach that uses estimated applicant preferences and assume TT in simulating the outcome, as long as preference estimates are consistent. Even though applicants may make mistakes, the counterfactual outcome is well approximated by the truth-telling

outcome. It may be tempting to skip the estimation step and assume that an applicant submits the same ROL in both regimes, given that DA is strategy-proof. If applicants play robust equilibrium, this assumption is not theoretically justified: if previously “out-of-reach” colleges become “within-reach” for an applicant under the counterfactual, we should not expect her to submit the same ROL. Our Monte Carlo simulations illustrate this point: assuming the same ROLs across two regimes can mis-predict the assignment of 40% of the applicants.

Other Related Literature. Our paper joins the long line of research recognizing that agents may not play rationally even in a strategically straightforward environment (e.g., [Li, 2017](#); [Dreyfuss, Heffetz, and Rabin, 2019](#)).² Our solution concept, robust equilibrium, is consistent with the previous literature rationalizing deviations from equilibrium behavior, e.g. rational inattention ([Sims, 2003](#); [Matejka and McKay, 2015](#)) and quantal response equilibria ([McKelvey and Palfrey, 1995](#)), which postulates that the higher the payoff consequences of a mistake are, the rarer the mistake is. While based on the same ideas, our solution concept is designed for a different goal. We are interested in the implications of mistakes and therefore are agnostic about why agents make mistakes. Compared to these existing concepts, robust equilibrium imposes less structure, and is thus more permissive, on the types of mistakes allowed. At the same time, it is more tractable for our large market analysis and admits a sharp prediction.³

As we are interested in studying the implications of strategic mistakes, our motivation and results are similar to [Kalai \(2004\)](#) and [Deb and Kalai \(2015\)](#). They also study approximate Bayesian equilibrium and show that it implies ex-post stability. Critically, they assume that the effect any participant can unilaterally have on an opponent’s payoff is uniformly bounded and decreases with the number of participants in the game. This assumption is tantamount to assuming “price-taking” (or “cutoff-taking”) behavior and does not hold in our setting even in an arbitrarily large economy (Example 1). Instead, we derive the result endogenously through an elaborate asymptotics of large random economies.

Our setting is related to Section IV.B of [Azevedo and Leshno \(2016\)](#) (AL, hereafter). They assume that colleges are overdemanded (i.e., the total college capacity is less than the total number of applicants) and that the gradient of demand is invertible. These assumptions may not hold in our setting and our results do not rely on them. Further, AL perform price-theory analysis of stable matchings without a game-theoretic framework. Yet, if applicants are allowed to strategize, they may not be price-takers even in a large economy. A richer game-theoretical setup makes some of the key results in AL inapplicable and requires a different proof strategy (for details, see [Appendices C and E](#)). We thus use

²[Dreyfuss, Heffetz, and Rabin \(2019\)](#) argue that such behavior may be intentional when applicants exhibit expectations-based loss aversion.

³The rational inattention model and quantal response equilibria, as formalized in the papers cited above, are generally intractable for a rich choice environment like matching where a choice takes the form of a rank-order list.

a novel technique, exploiting the lattice structure of stability and the properties of DA.

Our results provide a theoretical foundation for the stability-based approach to preference estimation. We are not the first to do so. [Fack, Grenet, and He \(2019\)](#) (FGH, hereafter) explain deviations from TT by positive application costs (i.e., costs associated with including a college in an ROL), while assuming fully rational applicants. However, FGH’s results imply that the colleges in an applicant’s ROL are ordered truthfully. This contradicts [Hassidim, Romm, and Shorrer \(2020\)](#) and [Li \(2017\)](#), who observe a non-negligible number of applicants with non-truthful order. Our robust equilibrium accommodates such a behavior.

Even though robust equilibrium imposes relatively little restrictions on applicant behavior in a large economy, thus accommodates applicant behavior that FGH do not, our prediction is sharper than FGH’s with respect to the equilibrium outcome: Theorem 2 shows that *every* regular robust equilibrium leads to asymptotic stability. In contrast, FGH show that there exists *one* such sequence of equilibria. This is an important difference because a mere existence does not guarantee that such an outcome would be observed in the field. Our uniqueness result (Theorem 2) provides a stronger justification for using stability for estimation. The proof of the theorem is also more challenging than in FGH and uses different techniques.

The rest of the paper is organized as follows. We first describe the model primitives in Section 2. Section 3 presents the analysis of applicant behavior and outcome under our solution concept. In Section 4, we provide the asymptotics of cutoffs with unilateral deviations. We conclude in Section 5.

2 Model Primitives

Consider an economy, F^k , in which k applicants compete for admissions to a finite set of colleges, $\mathbf{C} = \{c_1, \dots, c_C\}$, $C \geq 2$, under the applicant-proposing Deferred Acceptance algorithm ([Gale and Shapley, 1962](#)). Throughout, we refer to this algorithm simply as DA. A formal definition of DA can be found in Appendix A.

Each applicant has a type $\theta = (\mathbf{u}, \mathbf{s}) \in \Theta = [\underline{u}, \bar{u}]^C \times [0, 1]^C$ for some $0 \leq \underline{u} < \bar{u}$, where $\mathbf{u} = (u_1, \dots, u_C)$ is a vector of von-Neumann Morgenstern utilities of attending colleges, and $\mathbf{s} = (s_1, \dots, s_C)$ is a vector of scores representing the colleges’ priorities, such that an applicant with a higher score has a higher priority at a college. We assume that the outside option for an applicant has a zero utility, implying that all colleges are acceptable. This assumption simplifies the notation, but is not critical for our results. A vector $\mathbf{u}$ induces ordinal preferences over colleges, denoted by a rank-order list (ROL) of colleges, $\rho(\theta)$, of length C .

Colleges rank applicants by their scores; college capacities are a C -vector $k \cdot \mathbf{S}^k = [k \cdot \mathbf{S}]$, where $\mathbf{S} = (S_1, \dots, S_C)$, $0 < S_c < 1$ for all c , is a fixed vector and $\lceil \mathbf{x} \rceil$ is the vector of

integers nearest to $\mathbf{x}$ (rounded down in case of a tie).

The economy F^k is random in that applicant types are drawn independently from a full-support probability measure η over Θ ;⁴ the resulting empirical measure is denoted η^k . In this matching game, applicant types are private information, while η and all other information about the economy is common knowledge. Such a specification corresponds to the matching games summarized in Table 1 as well as many others many others in which admissions are based on privately known scores (for more examples, see Table 1 in [Fack, Grenet, and He, 2019](#)).⁵

We are interested in applicant behaviors and outcomes in “sufficiently large” economies and thus study the asymptotics of behaviors and outcomes in a sequence of random economies $\{F^k\}_{k \in \mathbb{N}}$. As $k \rightarrow \infty$, the number of applicants and college capacities increase proportionally, while the number of colleges is fixed. The sequence of economies $\{F^k\}$ converges in the sense that η^k converges in probability to η and that $\mathbf{S}^k$ converges to $\mathbf{S}$.

In each random economy, an applicant’s action is to choose an ROL from the set of possible ROLs $\mathcal{R}$. Applicant i ’s *strategy* is a measurable function $\sigma_i : \Theta \mapsto \Delta(\mathcal{R})$. One example is truth-telling, or TT, $\sigma_i(\theta) = \rho(\theta)$, which is a dominant strategy under DA ([Dubins and Freedman, 1981](#); [Roth, 1982](#)). A *strategy profile* for $\{F^k\}_{k \in \mathbb{N}}$, denoted by σ , is an infinite vector of individual strategies $\sigma = (\sigma_1, \sigma_2, \dots)$. Strategy profile σ requires that applicant i in F^k and applicant i in $F^{k'}$ adopt the same strategy σ_i , while their choice of ROL, which depends on their randomly drawn type θ_i , may differ. We say σ is *regular* if there exists $\gamma > 0$ such that for each i and each $\theta \in \Theta$, $\sigma_i(\theta)$ assigns probability of at least γ to playing $\rho(\theta)$.⁶ We denote the truncation of a strategy profile for the economy F^k , which omits the strategies of applicants not in F^k , by $\sigma^k = (\sigma_1, \dots, \sigma_k)$.

We assume that colleges are passive in the game. DA uses applicants’ submitted ROLs, their scores, and college capacities to calculate an outcome. An *outcome*, or a *matching*, is defined as a mapping $\mu : C \cup \Theta \rightarrow 2^\Theta \cup (C \cup \Theta)$ satisfying the usual two-sidedness and consistency requirements. A *stable matching* is also defined in the usual way to satisfy individual rationality and no-blocking.⁷ Stability is known to be an important property

⁴One important special case is the serial dictatorship, in which colleges’ preferences are given by a single score, i.e., $s_1 = \dots = s_C$. In that case, the full-support assumption holds with a reduced dimensionality of support; applicants’ scores are one-dimensional numbers in $[0, 1]$. Our results still hold.

⁵In these settings, applicants know their scores but do not know the scores and preferences of other applicants. Applicants may form a belief about the “typical” distribution of scores and preferences (captured by η), but are also aware that the particular distribution they face, η^k , may differ from the typical distribution. Note that our model does not apply to the setting where priorities are induced by lotteries, such as school choice in New York City ([Abdulkadiroglu, Pathak, and Roth, 2009](#); [Abdulkadiroglu, Agarwal, and Pathak, 2017](#); [Che and Tercieux, 2019](#)).

⁶A regular strategy need not mean that every applicant reports truthfully with some probability. We can “purify” it by defining a richer type space, with a “truthful” type who always adopts TT.

⁷Individual rationality requires that no participant (an applicant or a college) receives an unacceptable match, which holds by assumption here. No blocking means that no applicant-college pair exists such that the applicant prefers the college over her match and the college has either a vacant position or admits another applicant whom the college ranks below that applicant.

to ensure the longevity of a matching mechanism (Roth, 1991). When all applicants are TT (i.e., submitting $\rho(\theta)$) under DA, the resulting matching is stable (Gale and Shapley, 1962).

Given an outcome μ , we define a cutoff vector, $\mathbf{p} = (p_c)_{c \in C}$, such that college c 's cutoff p_c is the lowest score among c 's assigned applicants, $\mu(c)$, if its capacity is reached, or zero otherwise. When an applicant's score at college c , s_c , satisfies $s_c \geq p_c$, the college is *feasible* to her. An outcome is stable if everyone is assigned her most-preferred feasible college. DA ensures stability with respect to submitted ROLs as well as market clearing in the sense that no college admits more applicants than its capacity. When we consider a random economy F^k operated by DA, the cutoffs, which depend on applicants' realized types via σ^k , are random. We denote random cutoffs by $\mathbf{P} = (P_c)_{c \in C}$.

Our analysis will also consider any arbitrary cutoff vector, $\mathbf{p}$, and let applicants *demand* their highest-ranked feasible colleges in their ROLs. When doing so, we cannot guarantee market clearing, as the demand for some colleges may exceed their capacity.

3 Analysis of Robust Equilibria

To accommodate the play of dominated strategies documented in empirical studies, we introduce the following solution concept:⁸

DEFINITION 1. *A strategy σ forms a **robust equilibrium** if, for any $\epsilon > 0$, there exists $K \in \mathbb{N}$ such that, for each $k > K$, σ^k is an interim ϵ -Bayes Nash equilibrium of a k -random economy F^k —namely, σ gives each applicant within ϵ of the highest possible (supremum) payoff she can receive from any strategy when all the others employ σ .*

Note that robust equilibrium relaxes the exact Bayesian Nash solution concept by allowing for mistakes that are payoff insignificant in a large economy.⁹ Such a relaxation is necessary to accommodate mistakes in a finite economy. If cutoffs were known with certainty, a non-TT strategy, such as ranking only the most preferred feasible college with respect to the known cutoffs, may do just as well as TT. However, such a strategy is not optimal in a finite random economy because cutoffs are random and any non-TT strategy may result in a payoff loss with a positive probability. Thus, to explain departures from a dominant strategy, one must relax the exact optimality of applicant behavior. At the

⁸A number of authors adopted a similar ϵ -based solution concept to analyze approximate equilibrium behavior (see Kalai (2004), Deb and Kalai (2015), Azevedo and Budish (2018), and Che and Tercieux (2019), for instance).

⁹In this sense, our concept of robustness differs from another notion of “robustness,” or “incentives in the large” (see Che and Kojima (2010), Liu and Pycia (2016), Azevedo and Budish (2018), Che and Tercieux (2019), and Pycia (2019), for example). This latter concept refers to the property of a mechanism (rather than a solution concept) which provides asymptotic incentives for agents to report truthfully, even though truth-telling may not be an exact equilibrium behavior in a finite economy. By contrast, the current notion permits possible deviations from truth-telling even when it is a dominant strategy.

same time, a solution concept cannot be arbitrary, so some discipline must be placed on the extent to which payoff loss is tolerated. The robust equilibrium concept fulfills these requirements.

Does the robustness concept imply that most applicants report their preferences truthfully in a large economy? We show that it is not the case even for a slightly weaker requirement of weak truth-telling (WTT), which allows applicants to truncate $\rho(\theta)$. To show that WTT may not hold, we construct a robust equilibrium in which all but a vanishing fraction of applicants adopt *non*-WTT strategies.

To begin, we define a *stable-response strategy* (SRS) against an arbitrary, non-random cutoff vector $\mathbf{p}$ as *any* strategy whereby an applicant demands the most preferred feasible college given $\mathbf{p}$ (i.e., she ranks that college ahead of all other feasible colleges). The set of SRSs is typically large. She could skip infeasible colleges, rank them ahead of feasible ones, or flip their order relative to her true preferences. For a specific example, suppose that $\mathbf{C} = \{1, 2, 3, 4\}$, an applicant's true preference order is 1-2-3-4, and 2, 3 and 4 are feasible for her. Then, out of the 65 ROLs she can choose from, 21 are SRS, including ROLs 2-4-3-1, 2-4-1-3, 2-1-4-3, 1-2-4-3, and 2-4-3 which do not even respect the true preference order among the ranked colleges. For each type $\theta = (\mathbf{u}, \mathbf{s})$, there exists at least one SRS that violates WTT.¹⁰

In our first theorem, for a given vector of cutoffs $\mathbf{p}$, we allow every applicant to be non-truthful except for the applicants in the set

$$\Theta^\delta(\mathbf{p}) := \{(\mathbf{u}, \mathbf{s}) \in \Theta \mid \exists j \in \mathbf{C} \text{ s.t. } |s_j - p_j| \leq \delta\}$$

who are required to play TT. Note that everyone in $\Theta^\delta(\mathbf{p})$ has a score at some college close to that college's cutoff.

THEOREM 1. *There exists $\mathbf{p} \in [0, 1]^C$ such that, for any arbitrarily small $(\delta, \gamma) \in (0, 1)^2$, the following strategy forms a robust equilibrium: in each k -random economy,*

- *all applicants with types $\theta \in \Theta^\delta(\mathbf{p})$ play TT and*
- *all applicants with types $\theta \notin \Theta^\delta(\mathbf{p})$ randomize between TT (with probability γ) and an SRS strategy against $\mathbf{p}$ that violates both WTT and TT (with probability $1 - \gamma$).*

Since (δ, γ) is arbitrary, the following striking conclusion emerges.

COROLLARY 1. *There exists a robust equilibrium in which every applicant plays a non-WTT strategy (hence, a non-TT strategy) with probability arbitrarily close to one.*

¹⁰This can be shown as follows. If an applicant's most preferred college is infeasible (i.e., its cutoff at $\mathbf{p}$ is above the applicant's score at that college), then she can simply drop that college and rank order the remaining colleges truthfully. The resulting strategy is SRS but fails WTT. If an applicant most preferred college is feasible, then she can rank that college at the top of her ROL but rank the remaining colleges untruthfully (in relative rankings). Again, the resulting strategy is an SRS but violates WTT.

To the extent that a robust equilibrium is a reasonable solution concept, Theorem 1 implies that we should not be surprised to observe a non-negligible fraction of participants making “mistakes”—more precisely, playing dominated strategies—even in a strategy-proof environment. Importantly, even among the colleges that an applicant includes in her ROL, the order may not respect her true preferences. This result also raises some concerns about the empirical methods relying on WTT—any particular strategy relying on ROL data for that matter—as an identifying restriction.

If strategic mistakes undermine the prediction of applicant behavior, do they also undermine the stability of the outcome? This is an important question on two accounts. First, if mistakes jeopardize stability in a significant way, the rationale for using DA—to ensure a stable matching—should be called into question. Second, stability is widely used as an empirical identification assumption (see Fox, 2009; Agarwal, 2015; Fox and Bajari, 2013; Chiappori and Salanié, 2016; Fack, Grenet, and He, 2019, for instance). Our second theorem shows that mistakes captured by robust equilibrium leave the stability property of DA largely unscathed. We begin by defining a notion of approximate stability in large economies.

DEFINITION 2. *A strategy σ is **asymptotically stable** if the fraction of applicants assigned their most preferred feasible colleges (given the realized cutoffs) in economy F^k under σ^k converges in probability to one as $k \rightarrow \infty$.¹¹*

We now state the main theorem:

THEOREM 2. *Any regular robust equilibrium is asymptotically stable.*

While Theorem 2 already provides some justification for stability as an identification assumption for a sufficiently large economy, a question arises as to whether the concept of robust equilibrium would predict the same outcome as would emerge had all applicants reported their preferences truthfully. Our answer is in the affirmative:¹²

COROLLARY 2. *For a sequence of economies $\{F^k\}_k$, consider two sequences of outcomes: $\{\mu_\sigma^k\}_k$, generated by any regular robust equilibrium strategy σ , and $\{\mu_{TT}^k\}_k$, generated by TT. The fraction of applicants who receive their TT outcome while adopting σ (i.e., $\mu_\sigma^k(\theta) = \mu_{TT}^k(\theta)$) converges in probability to one.*

Although Theorem 1 questions TT as a *behavioral prediction*, Corollary 2 supports TT as a means for predicting an *outcome*. In this sense, the corollary validates the vast theoretical literature on DA that assume truth-telling. This result also suggests that when

¹¹More formally, we require that for any $\epsilon > 0$ there exists $K \in \mathbb{N}$ such that in any k -random economy with $k > K$, with probability of at least $1 - \epsilon$, at least a fraction $1 - \epsilon$ of all applicants are assigned their most preferred feasible colleges given the equilibrium cutoffs P^k .

¹²This result is reminiscent of the upper hemicontinuity of Nash equilibrium correspondence (see Fudenberg and Tirole, 1991, for instance). The current result is slightly stronger, however, since it implies that a sequence of ϵ -BNE (which is weaker than BNE) converges to an exact BNE as the economy grows large.

one evaluates the *outcome* of a counterfactual scenario involving DA, one can simply assume that applicants report their preferences truthfully in that scenario, as we do in Section F.

Taken together, our two theorems provide very different implications for the behavior and outcome under DA. On the one hand, the behavioral prediction exhibits multiplicity and possibly a drastic departure from truth-telling. On the other, the prediction in terms of outcome is virtually unique, and the outcome would be virtually the same as if all applicants reported their preferences truthfully. This latter finding should ultimately be reassuring about the performance of DA.

In the next section, we present novel results on cutoff convergence that significantly extends the existing results in Azevedo and Leshno (2016) and may be of interest in their own right. The two propositions are also a key step in proving Theorems 1 and 2. A reader who is more interested in an in-depth discussion of the implications for empirical studies may refer to Appendix F.

4 Cutoff Convergence

Applicants' ROLs give rise to demand for college seats. Random cutoffs $\mathbf{P}_{(i)}^k(\boldsymbol{\sigma}) \in [0, 1]^C$ clear random demands given applicants' drawn types and their mixed strategies. Non-random cutoffs $\bar{\mathbf{p}}(\boldsymbol{\sigma})$ clear the limit of *expected* demands. The formal definitions of these cutoffs can be found in Appendix C. Throughout the paper, we use $\|\cdot\|$ to denote the sup norm; i.e., for any $\mathbf{x}, \mathbf{x}' \in [0, 1]^C$, $\|\mathbf{x} - \mathbf{x}'\| := \sup_c |x_c - x'_c|$.

PROPOSITION 1. *Let $\boldsymbol{\sigma}$ be any γ -regular symmetric strategy profile $\boldsymbol{\sigma} = (\sigma, \dots, \sigma)$. Then*

$$\sup_{0 \leq i \leq k} \|\mathbf{P}_{(i)}^k(\boldsymbol{\sigma}) - \bar{\mathbf{p}}(\boldsymbol{\sigma})\| \xrightarrow{p} 0 \text{ as } k \rightarrow \infty.$$

Proposition 1 ensures that cutoffs are concentrated around $\bar{\mathbf{p}}(\boldsymbol{\sigma})$ with high probability even when there is a unilateral deviation. Note that it is established uniformly for all applicants, rather than for a fixed applicant. In the construction of a robust equilibrium strategy, Theorem 1 takes $\mathbf{p}$ to be $\bar{\mathbf{p}}(\boldsymbol{\rho})$, which are the cutoffs that would emerge in the limit if everyone is TT. Proposition 1 ensures that the random cutoffs from the constructed strategy converge to $\bar{\mathbf{p}}(\boldsymbol{\rho})$, even though almost no one is TT. These observations lead to Theorem 1.

Proposition 1 may be interesting in its own right as it generalizes AL's result on cutoff convergence with truth-telling applicants (part 2 of their Proposition 3). First, we do not require that $\sum_{c=1}^C S_c < 1$ (over-demanded systems); dropping this requirement may be practically important, as many matching markets, such as school choice, are not over-demanded. We also do not require that the derivative of limit expected demand is invertible. It is not clear whether this property holds under an arbitrary regular symmetric strategy $\boldsymbol{\sigma}$. Second and more importantly, in keeping with the field evidence, we allow for non-truthful

reporting and study the convergence of cutoffs following a unilateral deviation to TT, a necessary step in equilibrium analysis.

Proposition 1 only considers symmetric strategies, while our Theorem 2 allows for arbitrary regular strategies. Unfortunately, Proposition 1 does not hold for arbitrary asymmetric regular strategies.¹³ However, we can show the cutoff convergence for a subsequence of economies.

PROPOSITION 2. *Let σ be any γ -regular strategy profile. Then there exist a subsequence $\{F^{k_\ell}\}_\ell$ such that*

$$\sup_{0 \leq i \leq k_\ell} \|P_{(i)}^{k_\ell}(\sigma) - \bar{p}(\sigma)\| \xrightarrow{p} 0 \text{ as } \ell \rightarrow \infty.$$

Proposition 2 shows that, under mild conditions, the random cutoffs $P_{(i)}^{k_\ell}$ induced by σ converge to cutoffs $\bar{p}$, also induced by σ , in the limit. This is a critical ingredient in the proof of Theorem 2. The next step in the proof focuses on applicants whose utilities from each college are not too similar and whose scores are not too close to the limit cutoffs $\bar{p}$. In a large enough economy, almost all applicants fit this description. We then prove Theorem 2 by contradiction and assume σ is a regular robust equilibrium but not an SRS against $\bar{p}$. If the scores of applicant i are not too close to $\bar{p}$, then, by Proposition 2, they are not too close to $P_{(i)}^{k_\ell}$ in the subsequence either, with high probability. It implies that, in a large economy in the subsequence, applicant i can unilaterally deviate from σ and obtain a better assignment. Hence, σ is not a robust equilibrium, a contradiction.

5 Conclusion

Motivated by field evidence on mistakes, we theoretically argue, using a robust equilibrium concept, that an *outcome* of DA can be reliably predicted, but not participants' *behavior*. Moreover, in a sufficiently large economy, the outcome approximates well the one that would have emerged if everyone plays the dominant strategy. While this result justifies the vast theoretical literature that assumes truthful reporting behavior to analyze outcomes of DA, it calls into question empirical methods that take truthful reporting as a literal behavioral prediction. Our theory suggests that alternative approaches focusing on the stability property of the outcome may be robust to applicant mistakes. These implications are relevant to the estimation of participant preferences and counterfactual analysis.

Our paper focuses on environments where applicants know their scores according to which colleges rank them. However, the general insights can be extended to other settings, for example, where applicants are ranked by colleges according to a post-application lottery. [Che, Hahm, and He \(2022\)](#), an ongoing project, provide such an extension.

¹³To see this, consider a strategy profile σ such that blocks of applicants of an increasing length alternate between TT and submitting an empty ROL. These blocks can be constructed so that there are two subsequences of cutoffs: one is bounded below by a positive number with probability close to 1 and the other converges to 0. This implies non-convergence. Such an example is spelled out in detail in Appendix B.

Appendix to

Stable Matching with Mistaken Agents

Georgy Artemov

Yeon-Koo Che

YingHua He

February, 2022

List of Appendices

Appendix A: Definition of the Deferred-Acceptance Mechanism	14
Appendix B: An Example of Non-convergent Cutoffs	14
Appendix D: Preliminary Theoretical Results	18
Appendix E: Proofs of Theorems	26
Appendix G: Monte Carlo Simulations	37

A Definition of the Deferred-Acceptance Mechanism

The applicant-proposing Deferred-Acceptance (DA) mechanism uses each college's capacity and ranking over applicants as well as applicants' submitted ROLs to calculate a matching. It proceeds as follows:

Round 1. Every applicant applies to her first choice. Each college holds the highest-ranked applicants up to its capacity and rejects the rest, if any.

Generally, in

Round $m > 1$. Every applicant who is rejected in Round $(m - 1)$ applies to the next choice college on her ROL if there is one. Each college pools together new applicants and those held from Round $(m - 1)$; it holds the highest-ranked applicants up to its capacity and rejects the rest, if any.

The process terminates after any Round m when no rejections are issued. Each college is then matched with the applicants it is currently holding.

B An Example of Non-convergent Cutoffs

In this appendix, we construct a sequence of economies that satisfies the conditions of Proposition 2, in particular, full support of types, yet the sequence of cutoffs induced by a regular strategy does not converge. Note that this regular strategy is not a robust equilibrium.

In the example, we will use $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ to denote ceiling and floor functions. Consider a market with k applicants and two colleges, a and b . Each college has capacity $\lfloor k/4 \rfloor$. Applicants draw their types independently and uniformly from $\{a-b, b-a\} \times [0, 1]^2$.

We will consider two strategies: ρ , which is TT, and $\hat{\sigma}$ which prescribes submitting an empty ROL with probability $1 - \gamma$ and TT with probability γ , where γ is close to zero. Let $k_{2m} = \lceil 1000/\gamma^{2m} \rceil$, for $m \in \mathbb{N}$ (note that γ^{2m} is γ to the power of $2m$). The strategy profile σ is constructed as follows. Applicants from 1 to k_1 play $\hat{\sigma}$. For any m , applicants from $k_{2m-1} + 1$ to k_{2m} play ρ and applicants from $k_{2m} + 1$ to k_{2m+1} play $\hat{\sigma}$.

We consider two subsequences of economies: $\{F^{k_{2m}}\}_{m \in \mathbb{N}}$ and $\{F^{k_{2m+1}}\}_{m \in \mathbb{N}}$.

For any economy $F^{k_{2m}}$ from the first subsequence, applicants with indices between $k_{2m-1} + 1$ and k_{2m} play TT. Their total number is $\lceil 1000/\gamma^{2m} \rceil - \lceil 1000/\gamma^{2m-1} \rceil \geq \lfloor 1000(1 - \gamma)/\gamma^{2m} \rfloor$. As there are other applicants with lower indices who play TT, the fraction of applicant who are TT is more than $(1 - \gamma)$. Given the type distribution and capacities, when the economy size grows, in this subsequence, the cutoffs at each college tends to be no less than $(1 - \gamma)/2$ with probability close to one.

For any economy $F^{k_{2m+1}}$ in the second subsequence, applicants with indices between $k_{2m} + 1$ and k_{2m+1} constitute $(1 - \gamma)$ fraction of all applicants. They submit an empty

ROL with probability $1 - \gamma$. Thus, there are fewer than $\lfloor k/4 \rfloor$ applicants at each college, with probability close to 1. Thus, with probability close to 1, $P_a^{k_{2m+1}} = P_b^{k_{2m+1}} = 0$.

Given that, for these two subsequences, the cutoffs are either above $(1 - \gamma)/2$ or equal to 0 with probability close to 1, the sequence of cutoffs $\mathbf{P}^k$ does not converge in probability. Note that there is a convergent subsequence, in line with Proposition 2. This example also illustrates why AL's result cannot be applied. In particular, this setting does not have a symmetric strategy or an overall excess demand given the strategy profile.

C Proof Sketches: Cutoff Convergence and Asymptotic Stability

A simple but flawed intuition for Theorems 1 and 2 could be as follows. In a large economy, the distribution of types is close to η . One may hope that the cutoff distributions in a robust equilibrium of a large random economy are sufficiently concentrated around truth-telling cutoffs. With vanishing uncertainty about cutoffs, the payoff risk of playing non-TT strategies can be negligible. This would imply that there is a non-TT robust equilibrium, as Theorem 1 states. Further, if there is virtually no gain in playing TT instead of non-TT, then the outcome must be virtually the same under two strategies, implying asymptotic stability (Theorem 2).

There are multiple reasons why this intuition is flawed. First, we allow almost all applicants to play non-TT strategies in both Theorems 1 and 2; thus, there is no reason why cutoffs need to be concentrated around those arising under TT. Second and more worryingly, a unilateral deviation may significantly change the cutoffs, even in an arbitrarily large economy. We demonstrate the second point in the following example that builds on examples in Sotomayor (2008) and Haeringer and Klijn (2009).

EXAMPLE 1. Consider an economy F^k with $C = 3$ colleges, $\mathbf{a}, \mathbf{b}$ and $\mathbf{z}$, each with $n \in \mathbb{N}$ seats. There are three types of applicants, α , β , and ζ , with n applicants of each type.^{C.1} The ordinal preferences of the applicants and how colleges rank them are given below:

α :	$\mathbf{b}-\mathbf{a}-\mathbf{z}$	$\mathbf{a}$:	$\alpha-\zeta-\beta$
β :	$\mathbf{a}-\mathbf{b}-\mathbf{z}$	$\mathbf{b}$:	$\beta-\zeta-\alpha$
ζ :	$\mathbf{a}-\mathbf{b}-\mathbf{z}$	$\mathbf{z}$:	arbitrary

At each college, the top-ranked applicant type has a score in the interval $[2/3, 1]$, the middle-ranked type in $[1/3, 2/3]$, and the bottom-ranked type in $[0, 1/3]$.

Consider strategy profile σ^k such that applicants α and β play TT and that ζ ranks college $\mathbf{z}$ only ($\sigma(\zeta) = \mathbf{z}$). Every applicant is assigned the top-ranked college in her ROL.

^{C.1}We abuse the notion of “type” here. In the model, a type is a combination of cardinal utilities and scores.

The cutoffs P_a^k and P_b^k , induced by σ^k , are determined by applicants β and α , respectively, and both are below $1/3$.

Suppose that one applicant of type ζ deviates to TT. The deviation leads to a rejection chain where the type- β applicant with the lowest score at college **a** is rejected by **a** and applies to **b**; as a result, the type- α applicant with the lowest score at college **b** is rejected by **b** and applies to **a**, causing the rejection of another type- β applicant, and so on. At the last step of the rejection chain, all seats at colleges **a** and **b** are taken by applicants with scores above $2/3$. This illustrates that a single unilateral deviation leads to an abrupt change in P_a^k and P_b^k : before the deviation, the cutoffs are in the interval $[0, 1/3)$, while after the deviation, they are in $[2/3, 1]$. This is true for any n . Thus, when analyzing an applicant's deviation, we cannot assume that she is a “price-taker” in large economies: after a unilateral deviation, she will face cutoffs that may be significantly different from those induced by σ^k .

One may have hoped that the strategy we construct in the example, whereby ζ only ranks the least desirable college even though the other two colleges are feasible for her, is not a robust equilibrium. On the contrary, it is a robust equilibrium in a sequence of economies, where each economy is constructed as above, with $n \in \mathbb{N}$. A unilateral deviation of a type- ζ applicant to TT leads to a change in the cutoff but no change in her assignment, hence not a profitable deviation. Thus, the example cannot be dismissed as irrelevant to the analysis of robust equilibria.

Example 1 can be made more general by allowing an unequal number of applicants of each type. For instance, if there are more than n of type- α and type- β applicants, a different strategy—e.g., when some type- α and type- β applicants rank a single college—will still result in a unilateral deviation abruptly changing the cutoffs.

Propositions 1 and 2 show that the phenomenon described in Example 1 does not happen in the economies we study. We will end this section by sketching proofs of these propositions; for that, we need some additional notations.

Fix σ . The strategy profile σ^k induces a random ROL, R_j , for each applicant $j \in \{1, \dots, k\}$. For any $\mathbf{p} \in [0, 1]^C$, we define a per capita profile of (random) demands for colleges—henceforth, simply called demand— $\mathbf{D}^k(\mathbf{p}; \sigma) = (D_c^k(\mathbf{p}; \sigma))_{c \in C}$; the demand for college c is given by

$$D_c^k(\mathbf{p}; \sigma) := \frac{1}{k} \sum_{j=1}^k \mathbb{I} \left\{ c \in \arg \max_{\text{w.r.t. } R_j} \{c' \in C : s_{j,c'} \geq p_{c'}\} \right\},$$

where $\arg \max_{\text{w.r.t. } R_j}$ picks the highest-ranked college in R_j from the set of feasible colleges $\{c' \in C : s_{j,c'} \geq p_{c'}\}$ and $\mathbb{I}\{\cdot\}$ is an indicator function. Similarly, we define a demand profile $\mathbf{D}_{(i)}^k(\mathbf{p}; \sigma) = (D_{(i),c}^k(\mathbf{p}; \sigma))_{c \in C}$ that arises when applicant i employs truthful reporting ρ and all other applicants $j \neq i$ continue to use σ_j . For notational convenience, we use $\mathbf{D}_{(0)}^k(\mathbf{p}; \sigma) = \mathbf{D}^k(\mathbf{p}; \sigma)$ to denote the demand arising from the original strategy σ^k . Let

$\overline{D}_{(i)}^k(\mathbf{p}; \boldsymbol{\sigma}) := \mathbb{E} \left[D_{(i)}^k(\mathbf{p}; \boldsymbol{\sigma}) \right]$, where the expectation is taken over the random draws of applicants' types and the random ROLs arising from $\boldsymbol{\sigma}_{(i)}^k$ being (possibly) mixed.

Random cutoffs $\mathbf{P}_{(i)}^k(\boldsymbol{\sigma}) \in [0, 1]^C$, defined by DA with ROLs prescribed by $\boldsymbol{\sigma}$, clear the (random) demand system $\mathbf{D}_{(i)}^k(\mathbf{p}; \boldsymbol{\sigma})$. Cutoffs $\overline{\mathbf{p}}(\boldsymbol{\sigma})$ clear the (non-random) demand system $\overline{\mathbf{D}}(\mathbf{p}; \boldsymbol{\sigma})$. Appendix D.1 provides an argument that $\overline{\mathbf{p}}(\boldsymbol{\sigma})$ exists for any regular strategy. When there is no ambiguity, we use $\overline{\mathbf{p}}$ instead of $\overline{\mathbf{p}}(\boldsymbol{\sigma})$.

We can now sketch the proofs of Propositions 1 and 2; they are identical except for some initial steps. We first establish some preliminary results about demands. We fix an arbitrary regular $\boldsymbol{\sigma}$ and, in the rest of this section, suppress $\boldsymbol{\sigma}$ in the notation of demands when there is no confusion. We study the sequence of (random) demand vectors $\{\mathbf{D}_{(i)}^k(\mathbf{p})\}_k$ induced by $\boldsymbol{\sigma}$ for any fixed cutoffs $\mathbf{p}$. The fact that an individual applicant's influence on the per capita demand vector is vanishing in the limit leads to a version of the law of large numbers: for each vector $\mathbf{p}$, $\mathbf{D}_{(i)}^k(\mathbf{p})$ converges pointwise to its expectation $\overline{\mathbf{D}}_{(i)}^k(\mathbf{p})$ in probability as $k \rightarrow \infty$ (see McDiarmid, 1989). Functions $\{\overline{\mathbf{D}}_{(i)}^k(\mathbf{p})\}_k$ are Lipschitz-continuous and there is a subsequence $\{\overline{\mathbf{D}}_{(i)}^{k_\ell}(\mathbf{p})\}_\ell$ that converges to some Lipschitz-continuous function $\overline{\mathbf{D}}(\mathbf{p})$ (the Arzela-Ascoli theorem). Then, using an argument in the spirit of the Glivenko-Cantelli theorem, we show that, along this subsequence, the random demand $\mathbf{D}_{(i)}^{k_\ell}(\mathbf{p})$ converges uniformly to $\overline{\mathbf{D}}(\mathbf{p})$ in probability.

Recall that cutoffs $\mathbf{P}_{(i)}^{k_\ell}$ and $\overline{\mathbf{p}}$ clear demands $\mathbf{D}_{(i)}^{k_\ell}(\cdot)$ and $\overline{\mathbf{D}}(\cdot)$, respectively. As both are market-clearing cutoffs, the difference $\|\mathbf{D}_{(i)}^{k_\ell}(\mathbf{P}_{(i)}^{k_\ell}) - \overline{\mathbf{D}}(\overline{\mathbf{p}})\|$ is essentially zero; it may be non-zero only because the number of seats is rounded in finite economies or because some colleges have vacant seats. Given the results so far, this difference is close to $\|\overline{\mathbf{D}}(\mathbf{P}_{(i)}^{k_\ell}) - \overline{\mathbf{D}}(\overline{\mathbf{p}})\|$. Due to the full-support assumption and regularity of $\boldsymbol{\sigma}$, $\|\overline{\mathbf{D}}(\mathbf{P}_{(i)}^{k_\ell}) - \overline{\mathbf{D}}(\overline{\mathbf{p}})\|$ can be small only if $\|\mathbf{P}_{(i)}^{k_\ell} - \overline{\mathbf{p}}\|$ is small, as there can be applicants who demand different colleges under $\mathbf{P}_{(i)}^{k_\ell}$ and $\overline{\mathbf{p}}$. Thus, we can conclude that $\|\mathbf{P}_{(i)}^{k_\ell} - \overline{\mathbf{p}}\|$ is small. The formal argument relies on the fact that there is a unique stable outcome in the limit economy (by Theorem 1 in Azevedo and Leshno (2016)) and also uses inductive argument on the steps of applicant-proposing and college-proposing DAs. This concludes the proof.

One may wonder if our propositions could alternatively be proven by applying AL's Proposition 3, for example, by constructing a hypothetical economy in which "true" ordinal preferences are ROLs generated by $\boldsymbol{\sigma}$. AL's results are not applicable when we study unilateral deviations or non-symmetric regular strategies because AL require i.i.d. draws of ROLs. Appendix B provides an example. Even if we were to study the convergence of no-deviation cutoffs $(\mathbf{P}_{(0)}^k)$ under a symmetric strategy $\boldsymbol{\sigma} = (\sigma, \sigma, \dots)$ —thus ensuring i.i.d. draws of ROLs—AL's Proposition 3 would still not imply our Proposition 1 for two reasons. First, AL's analysis requires a continuum economy with a well-defined measure of types. In our setup, $\boldsymbol{\sigma}$ may involve an arbitrary mixing. Thus, in the hypothetical continuum

economy, the measure of types, which now depends on both η and σ , requires a law of large numbers on the continuum economy to be well-defined. That has usual conceptual difficulties (see, e.g., Judd, 1985). Second, their result requires an overdemand system and the invertibility of $\partial D(\bar{p})$, which are not guaranteed in our setting.

D Preliminary Theoretical Results

Consider the continuum economy $E = [\eta, S]$ with the full support assumption $\frac{1}{C}(\bar{u} - \underline{u})^C \eta(\theta) > \xi$ for all $\theta \in \Theta$ for some $\xi > 0$.

Recall that $D_{(i)}^k(\mathbf{p}; \sigma)$ is the demand induced by σ in economy F^k when applicant i deviates to TT (with $i = 0$ standing for no deviation). $\bar{D}_{(i)}^k(\mathbf{p}; \sigma)$ is the expectation of demands $D_{(i)}^k(\mathbf{p}; \sigma)$ and $\bar{D}_{(i)}(\mathbf{p}; \sigma)$ is the limit of $\bar{D}_{(i)}^k(\mathbf{p}; \sigma)$. We omit σ from the expression of the demands in the following. By construction, $D_{(i),c}^k(\mathbf{p})$ are non-increasing in p_c and non-decreasing in p_{-c} for any $c \in C$ and $0 \leq i \leq k$.

We now formally describe the outcome of an applicant-proposing deferred acceptance algorithm (DA) in the k -random economy by defining the *DA cutoffs* of k -random economy, $\mathbf{P}^k := \lim_{m \rightarrow \infty} (P_1^{k,m}, \dots, P_C^{k,m})$, where $\mathbf{P}^{k,0} = (0, \dots, 0)$ and for $m \geq 1$,

$$P_c^{k,m} = \sup \left\{ p \in [0, 1] : D_c^k(p, \mathbf{P}_{-c}^{k,m-1}) = S_c^k \right\}, \forall c \in C,$$

if the set is nonempty and $P_c^{k,m} = 0$ otherwise. Note that the iterative steps of defining the cutoffs correspond to the iterative steps of DA. Initially, the applicants who prefer college c most apply to c and c tentatively accepts applicants from among them in the descending order of score s_c up to its capacity. That is, for $c \in C$, $D_c^k(0, \dots, 0)$ is the measure of applicants to c and S_c^k is the capacity of c , so $P_c^{k,1}$ becomes the cutoff for c in step 1. More generally, in step m , a measure $D_c^k(P_c^{k,m-1}, \mathbf{P}_{-c}^{k,m-1})$ of applicants apply to c , and the same process determines the cutoff $P_c^{k,m}$ for college c .^{D.2} Due to the property of $D^k(\mathbf{p})$ observed above, $\mathbf{P}^{k,m} = (P_c^{k,m})_c$ is monotone non-decreasing, and the limit $\mathbf{P}^k$ is well defined. Importantly, the cutoffs at each step, and thus $\mathbf{P}^k$, are random since D^k is random.

Even though we are interested in the outcome of DA (i.e., the applicant-proposing deferred acceptance), it is useful to define the cutoffs that arise from CPDA (college-proposing deferred acceptance). Let the *CPDA cutoffs* be defined by $\mathbf{Q}^k := \lim_{m \rightarrow \infty} (Q_1^{k,m}, \dots, Q_C^{k,m})$,

^{D.2}The measure $D_c^k(P_c^{k,m-1}, \mathbf{P}_{-c}^{k,m-1})$ includes applicants retained from the previous round. The description we provide is a slight modification to the usual DA: applicants who have never been rejected by a college and have a score below $P_c^{k,m-1}$ do not apply to college c in round m . These applicants would have been rejected if they applied. Like the standard DA, the algorithm converges in at most Ck steps.

where $\mathbf{Q}^{k,0} = (1, \dots, 1)$ and for $m \geq 1$,

$$Q_c^{k,m} = \sup \left\{ p \in [0, 1] : D_c(p, \mathbf{Q}_{-c}^{k,m-1}) = S_c^k \right\}, \forall c \in C,$$

if the set is nonempty and $Q_c^{k,m} = 0$ otherwise. Similarly to before, we observe that $\mathbf{Q}^{k,m} := (Q_c^{k,m})_c$ are monotone non-increasing in m , so $\mathbf{Q}^k$ is well defined.

Finally, the standard lattice property of stable matchings and the extremality of DA and CPDA matchings imply that $\mathbf{P}^k \leq \mathbf{Q}^k$.^{D.3}

Next, suppose i unilaterally deviates to TT (ρ). We can define the resulting DA and CPDA cutoffs analogously, and denote them respectively by $\mathbf{P}_{(i)}^k$ and $\mathbf{Q}_{(i)}^k$ and observe $\mathbf{P}_{(i)}^k \leq \mathbf{Q}_{(i)}^k$. It is notationally convenient to define the cutoffs when no one deviates from σ^k by $\mathbf{P}_{(0)}^k := \mathbf{P}^k$ and $\mathbf{Q}_{(0)}^k := \mathbf{Q}^k$.

Our goal is to establish a desirable limit behavior of $(\mathbf{P}_{(i)}^k)_{i \in \mathbb{N}_0}$ as $k \rightarrow \infty$, where $\mathbb{N}_0 := \mathbb{N} \cup 0$. We accomplish this goal in Section D.1. To this end, however, we need to establish a few preliminary results.

We first establish a probabilistic bound for the distance between $\mathbf{D}_{(i)}^k$ and its expectation. For $i \in \mathbb{N}_0$ and $\mathbf{p} \in [0, 1]^C$, recall $\overline{\mathbf{D}}_{(i)}^k(\mathbf{p}) := \mathbb{E} [\mathbf{D}_{(i)}^k(\mathbf{p})]$, where the expectation is taken over the random draws of applicants' types (when F^k is constructed) and the randomness in the ROLs arising from $\sigma_{(i)}^k$ being (possibly) mixed. Recall that, throughout, we use $\|\cdot\|$ to denote the sup norm; i.e., for any $\mathbf{x}, \mathbf{x}' \in [0, 1]^{|C|}$, $\|\mathbf{x} - \mathbf{x}'\| := \sup_c |x_c - x'_c|$.

LEMMA 1. *Fix any strategy σ , any $\mathbf{p} \in [0, 1]^C$, and any $i \in \mathbb{N}_0$. Then, for any $\alpha > 0$,*

$$\Pr \left[\left\| \mathbf{D}_{(i)}^k(\mathbf{p}) - \overline{\mathbf{D}}_{(i)}^k(\mathbf{p}) \right\| > \alpha \right] \leq |C| \cdot e^{-2k\alpha^2}.$$

Proof. By McDiarmid's inequality (McDiarmid, 1989), for each $c \in C$,

$$\Pr \left\{ \left| D_{(i),c}^k(\mathbf{p}) - \overline{D}_{(i),c}^k(\mathbf{p}) \right| > \alpha \right\} \leq e^{-2k\alpha^2},$$

since for each $c \in C$, $|D_{(i),c}^k(p; R_1, \dots, R_k) - D_{(i),c}^k(p; R'_1, \dots, R'_k)| \leq 1/k$ whenever ROLs $(R_1, \dots, R_k)$ and $(R'_1, \dots, R'_k)$ differ only in one component (recall that demands depend on ROLs, although ROLs are usually suppressed in the notation).

It then follows that

$$\Pr \left[\left\| \mathbf{D}_{(i)}^k(\mathbf{p}) - \overline{\mathbf{D}}_{(i)}^k(\mathbf{p}) \right\| > \alpha \right]$$

^{D.3}A useful perspective is to view $\mathbf{P}^k$ and $\mathbf{Q}^k$ as the smallest and largest fixed points of a self map $\Phi : [0, 1]^C \rightarrow [0, 1]^C$ defined by $\Phi_c(\mathbf{p}) := \sup \{ p_c \in [0, 1] : D_c(p_c, \mathbf{p}_{-c}) = S_c^k \}$ if the set is nonempty and otherwise $\Phi_c(\mathbf{p}) := 0$. The monotonicity of Φ means that by Tarski's fixed point theorem, the fixed points of Φ form a complete lattice, admitting extremal points. Such extremal fixed points are obtained via the iterative steps we have defined.

$$\begin{aligned}
&= \Pr \left[\exists c \in \mathbf{C} \text{ s.t. } \left| D_{(i),c}^k(\mathbf{p}) - \overline{D}_{(i),c}^k(\mathbf{p}) \right| > \alpha \right] \\
&\leq \sum_{c \in \mathbf{C}} \Pr \left[\left| D_{(i),c}^k(\mathbf{p}) - \overline{D}_{(i),c}^k(\mathbf{p}) \right| > \alpha \right] \leq C \cdot e^{-2k\alpha^2}.
\end{aligned}$$

□

LEMMA 2. For any strategy σ and each $(k, i) \in \mathbb{N} \times \mathbb{N}_0$, the function $\overline{D}_{(i)}^k(\mathbf{p})$ is Lipschitz continuous with a constant L that is independent of (k, i) .

Proof. Let $\mathbf{p}$ and $\mathbf{p}'$ be two arbitrary cutoff vectors in $[0, 1]^C$. Define

$$\Theta_{\mathbf{p}, \mathbf{p}'} := \{(\mathbf{u}, \mathbf{s}) \in \Theta : \exists c \in \mathbf{C} \text{ such that } p_c < s_c < p'_c \text{ or } p'_c < s_c < p_c\}.$$

Since η is absolutely continuous with respect to Lebesgue measure, we have $\eta(\Theta_{\mathbf{p}, \mathbf{p}'}) \leq L\|\mathbf{p}' - \mathbf{p}\|$, where L is an upper bound for the density for all $\theta \in \Theta$. Then,

$$\begin{aligned}
&\left\| \overline{D}_{(i)}^k(\mathbf{p}') - \overline{D}_{(i)}^k(\mathbf{p}) \right\| \\
&= \sup_{c \in \mathbf{C}} \left| \mathbb{E}_{\theta, \sigma} \left[D_{(i),c}^k(\mathbf{p}') - D_{(i),c}^k(\mathbf{p}) \right] \right| \\
&= \sup_{c \in \mathbf{C}} \left| \mathbb{E}_{\theta} \left[\frac{1}{k} \sum_{j=1}^k \sum_{R \in \mathcal{R}} \mathbb{P}(\sigma_j(\theta_j) = R) \left(\begin{array}{c} \mathbb{I} \{c \in \arg \max_{\text{w.r.t. } R} \{c' \in \mathbf{C} : s_{j,c'} \geq p'_{c'}\}\} \\ \mathbb{I} \{c \in \arg \max_{\text{w.r.t. } R} \{c' \in \mathbf{C} : s_{j,c'} \geq p_{c'}\}\} \end{array} \right) \right] \right| \\
&\leq \sup_{c \in \mathbf{C}} \frac{1}{k} \sum_{j=1}^k \mathbb{E}_{\theta} \left[\sum_{R \in \mathcal{R}} \mathbb{P}(\sigma_j(\theta_j) = R) \left| \begin{array}{c} \mathbb{I} \{c \in \arg \max_{\text{w.r.t. } R} \{c' \in \mathbf{C} : s_{j,c'} \geq p'_{c'}\}\} \\ \mathbb{I} \{c \in \arg \max_{\text{w.r.t. } R} \{c' \in \mathbf{C} : s_{j,c'} \geq p_{c'}\}\} \end{array} \right| \right] \\
&\leq \frac{1}{k} \sum_{j=1}^k \mathbb{E}_{\theta} [\mathbb{I} \{\theta_j \in \Theta_{\mathbf{p}, \mathbf{p}'}\}] \\
&= \mathbb{E}_{\theta} [\mathbb{I} \{\theta_j \in \Theta_{\mathbf{p}, \mathbf{p}'}\}] = \eta(\Theta_{\mathbf{p}, \mathbf{p}'}) \leq L\|\mathbf{p}' - \mathbf{p}\|,
\end{aligned}$$

where the expectation $\mathbb{E}_{\theta, \sigma}$ is over applicant types and mixed strategies; $\mathbb{E}_{\theta}$ is an expectation over applicant types; and $\mathbb{P}(\sigma_j(\theta_j) = R)$ is the probability that applicant j of type θ_j submits R as prescribed by mixed strategy $\sigma_j(\theta_j)$ (with an abuse of notation, we denote i 's strategy by σ_i even though i deviates to truth-telling). The first inequality follows from Jensen's inequality and the second inequality holds since the two sets, $\{c' \in \mathbf{C} : s_{i,c'} \geq p'_{c'}\}$ and $\{c' \in \mathbf{C} : s_{i,c'} \geq p_{c'}\}$, are identical when $\theta_i \notin \Theta_{\mathbf{p}, \mathbf{p}'}$. □

LEMMA 3. There exists a subsequence of economies F^{k_ℓ} such that $\sup_{\mathbf{p}, i} \|\overline{D}_{(i)}^{k_\ell}(\mathbf{p}) - \overline{D}_{(0)}(\mathbf{p})\| \rightarrow 0$ as $\ell \rightarrow \infty$. Function $\overline{D}_0(\mathbf{p})$ is Lipschitz-continuous with the same constant L as Lipschitz-continuous functions $\overline{D}_{(i)}^{k_\ell}(\mathbf{p})$.

Proof. The sequence of functions $\{\bar{\mathbf{D}}^k(\mathbf{p})\}_{k=1}^\infty$ defined on a compact set $[0, 1]^C$ is uniformly bounded and uniformly equicontinuous (which follows from their Lipschitz property with a uniform constant L , as shown in Lemma 2). By the Arzela-Ascoli theorem, we can find a subsequence $\{\bar{\mathbf{D}}^{k_j}(\mathbf{p})\}_{j=1}^\infty$ which converges uniformly to Lipschitz-continuous function $\bar{\mathbf{D}}(\mathbf{p})$ with the same constant L , since the Lipschitz property is preserved in the limit.

Now consider any $i \neq 0$. For any $\mathbf{p} \in [0, 1]^C$ and $i \in \mathbb{N}$,

$$\left\| \bar{\mathbf{D}}_{(i)}^{k_j}(\mathbf{p}) - \bar{\mathbf{D}}^{k_j}(\mathbf{p}) \right\| = \left\| \mathbb{E} \left[\mathbf{D}_{(i)}^{k_j}(\mathbf{p}) - \mathbf{D}^{k_j}(\mathbf{p}) \right] \right\| \leq \mathbb{E} \left[\left\| \mathbf{D}_{(i)}^{k_j}(\mathbf{p}) - \mathbf{D}^{k_j}(\mathbf{p}) \right\| \right] \leq \frac{1}{k_j},$$

since changing the strategy from $\boldsymbol{\sigma}^{k_j}$ to $\boldsymbol{\sigma}_{(i)}^{k_j}$ can change the demand for any college at most by $1/k_j$.

Combining this result with the earlier observation, we conclude that there exists a subsequence in the sequence $\{\bar{\mathbf{D}}_{(i)}^k(\mathbf{p})\}_{k=1}^\infty$ that converges uniformly to the same $\bar{\mathbf{D}}(\mathbf{p})$ for all $i \in \mathbb{N}_0$. $\square$

Several of the results below are shown for a subsequence of economies described in Lemma 3 and its associated strategy profile. It is useful to introduce a specific name for such subsequences.

DEFINITION 3. A subsequence $\{F^{k_\ell}, \boldsymbol{\sigma}^{k_\ell}\}_\ell$ which induces a subsequence of Lipschitz-continuous demands $\{\mathbf{D}_{(i)}^{k_\ell}(\mathbf{p})\}_{k_\ell, 0 \leq i \leq N}$ with the same constant L is *expected-demand convergent* if there exist Lipschitz-continuous demands $\bar{\mathbf{D}}(\mathbf{p})$ with the same constant L such that $\sup_{\mathbf{p}, i} \|\bar{\mathbf{D}}_{(i)}^{k_\ell}(\mathbf{p}) - \bar{\mathbf{D}}(\mathbf{p})\| \rightarrow 0$ as $\ell \rightarrow \infty$.

LEMMA 4. Consider any expected-demand-convergent subsequence $\{F^{k_\ell}, \boldsymbol{\sigma}^{k_\ell}\}_\ell$. Then, for any $\epsilon > 0$,

$$\lim_{\ell \rightarrow \infty} \Pr \left\{ \sup_{i, \mathbf{p}} \|\mathbf{D}_{(i)}^{k_\ell}(\mathbf{p}) - \bar{\mathbf{D}}(\mathbf{p})\| > \epsilon \right\} = 0.$$

Proof. Since $\bar{\mathbf{D}}$ is Lipschitz-continuous, thus continuous, we can partition the space of $\mathbf{p}$'s into finite intervals of the form $Z(\boldsymbol{\kappa}) := \prod_c [p_{\kappa_c}, p_{\kappa_c+1}]$, where $\boldsymbol{\kappa} = (\kappa_c)_c \in \{0, \dots, n\}^C$, for some $n \in \mathbb{N}$,^{D.4} such that for each c ,

$$|\bar{D}_c(\mathbf{p}) - \bar{D}_c(\mathbf{p}')| < \frac{\epsilon}{2}$$

for all $\mathbf{p}, \mathbf{p}' \in Z(\boldsymbol{\kappa})$, $\forall \boldsymbol{\kappa} = (\kappa_c)_c$. There are n^C such intervals.

Consider any $\mathbf{p}$ and c . Let $\boldsymbol{\kappa}$ be the index of the interval such that $\mathbf{p} \in Z(\boldsymbol{\kappa})$. Let

$$\mathbf{p}'_{\boldsymbol{\kappa}} := (p_{\kappa_1+1}, \dots, p_{\kappa_{c-1}+1}, p_{\kappa_c}, p_{\kappa_{c+1}+1}, \dots, p_{\kappa_C+1}),$$

^{D.4}Naturally, we have $p_{\kappa_0} = \mathbf{0}$ and $p_{\kappa_n} = \mathbf{1}$.

$$\mathbf{p}_{\kappa}'' := (p_{\kappa_1}, \dots, p_{\kappa_{c-1}}, p_{\kappa_c+1}, p_{\kappa_{c+1}}, \dots, p_{\kappa_C}).$$

The demand for c is the highest at $\mathbf{p}_{\kappa}'$ and the lowest at $\mathbf{p}_{\kappa}''$ among the prices in $Z(\kappa)$. Then, ^{D.5}

$$\left| D_{(i),c}^{k_\ell}(\mathbf{p}) - \bar{D}_c(\mathbf{p}) \right| \leq \max \left\{ |D_{(i),c}^{k_\ell}(\mathbf{p}_{\kappa}') - \bar{D}_c(\mathbf{p}_{\kappa}')|, |D_{(i),c}^{k_\ell}(\mathbf{p}_{\kappa}'') - \bar{D}_c(\mathbf{p}_{\kappa}'')| \right\} + \frac{\epsilon}{2}. \quad (\text{D.1})$$

Suppose the event $\{\|\mathbf{D}_{(i)}^{k_\ell}(\mathbf{p}) - \bar{\mathbf{D}}(\mathbf{p})\| > \epsilon\}$ occurs for some $\mathbf{p}$ and i . Then, since $\|\mathbf{D}_{(i)}^{k_\ell}(\mathbf{p}) - \bar{\mathbf{D}}(\mathbf{p})\| = \sup_c |D_{(i),c}^{k_\ell}(\mathbf{p}) - \bar{D}_c(\mathbf{p})|$, there must exist c such that $|D_{(i),c}^{k_\ell}(\mathbf{p}) - \bar{D}_c(\mathbf{p})| \geq \epsilon$. From (D.1), there is $\mathbf{p}_{\kappa}^* \in \{\mathbf{p}_{\kappa}', \mathbf{p}_{\kappa}''\}$ is such that $|D_{(i),c}^{k_\ell}(\mathbf{p}_{\kappa}^*) - \bar{D}_c(\mathbf{p}_{\kappa}^*)| \geq \frac{\epsilon}{2}$. Since $\bar{D}_{(i),c}^{k_\ell}(\mathbf{p})$ converges to $\bar{D}_c(\mathbf{p})$ in sup norm by Lemma 3, there exists N' such that for all $\ell > N'$, $\sup_{i,\hat{\mathbf{p}}} |\bar{D}_{(i),c}^{k_\ell}(\hat{\mathbf{p}}) - \bar{D}_c(\hat{\mathbf{p}})| < \frac{\epsilon}{4}$. Consequently, for $\ell > N'$ and $\mathbf{p}_{\kappa}^*$, we must have

$$\left| D_{(i),c}^{k_\ell}(\mathbf{p}_{\kappa}^*) - \bar{D}_{(i),c}^{k_\ell}(\mathbf{p}_{\kappa}^*) \right| \geq \frac{\epsilon}{4}.$$

Combining the arguments so far, we conclude:

$$\begin{aligned} & \Pr \left\{ \sup_{\mathbf{p}, i} \|\mathbf{D}_{(i)}^{k_\ell}(\mathbf{p}) - \bar{\mathbf{D}}(\mathbf{p})\| > \epsilon \right\} \\ &= \Pr \left\{ \exists (\mathbf{p}, i) \text{ s.t. } \|\mathbf{D}_{(i)}^{k_\ell}(\mathbf{p}) - \bar{\mathbf{D}}(\mathbf{p})\| > \epsilon \right\} \\ &\leq \sum_{i=0}^{k_\ell} \Pr \left\{ \exists c \text{ and } \mathbf{p}_{\kappa}^* \text{ s.t. } |D_{(i),c}^{k_\ell}(\mathbf{p}_{\kappa}^*) - \bar{D}_{(i),c}^{k_\ell}(\mathbf{p}_{\kappa}^*)| \geq \frac{\epsilon}{4} \right\} \\ &\leq \sum_{i=0}^{k_\ell} \sum_{c=1}^C \sum_{\kappa \in \{0, \dots, n\}^C} \Pr \left\{ |D_{(i),c}^{k_\ell}(\mathbf{p}_{\kappa}^*) - \bar{D}_{(i),c}^{k_\ell}(\mathbf{p}_{\kappa}^*)| \geq \frac{\epsilon}{4} \right\} \\ &\leq n^C C \sum_{i=0}^{k_\ell} e^{-k_\ell \epsilon^2 / 8} \\ &= n^C C (k_\ell + 1) e^{-k_\ell \epsilon^2 / 8} \rightarrow 0 \text{ as } \ell \rightarrow \infty, \end{aligned}$$

where the last inequality follows from McDiarmid inequality (see Lemma 1). $\square$

^{D.5}This inequality follows from an argument used for the proof of the Glivenko-Cantelli theorem extended to a multidimensional case.

D.1 Asymptotics of Cutoffs for a Regular Strategy

We are now ready to establish, for a regular strategy, the uniform convergence in probability of cutoffs $\mathbf{P}_{(i)}^{k_\ell}$ to some deterministic cutoffs as $\ell \rightarrow \infty$ along any expected-demand-convergent subsequence. In fact, we show that the limit cutoffs are the cutoffs defined by the limit demand system (i.e., $\overline{\mathbf{D}}(\mathbf{p})$). Specifically, let $\overline{\mathbf{p}} = (\overline{p}_1, \dots, \overline{p}_C)$ be the DA cutoffs defined by $\overline{p}_c := \lim_{m \rightarrow \infty} \overline{p}_c^m$, where $\overline{\mathbf{p}}^0 = (0, \dots, 0)$ and for $m \geq 1$, and $\overline{\mathbf{p}}^m = (\overline{p}_c^m)_c$ is given by

$$\overline{p}_c^m = \sup \{p \in [0, 1] : \overline{D}_c(p, \overline{\mathbf{p}}_{-c}^{m-1}) = S_c\}, \forall c \in C,$$

if the set is nonempty, and $\overline{p}_c^m = 0$ otherwise.

Similarly, let $\overline{\mathbf{q}} = (\overline{q}_1, \dots, \overline{q}_C)$ be the CPDA (College-Proposing Deferred Acceptance) cutoffs defined by $\overline{q}_c := \lim_{m \rightarrow \infty} \overline{q}_c^m$ for each c , where $\overline{\mathbf{q}}^0 = (1, \dots, 1)$ and for $m \geq 1$, and $\overline{\mathbf{q}}^m = (\overline{q}_c^m)_c$ is given by

$$\overline{q}_c^m = \sup \{p \in [0, 1] : \overline{D}_c(p, \overline{\mathbf{q}}_{-c}^{m-1}) = S_c\}, \forall c \in C,$$

if the set is nonempty and $\overline{q}_c^m = 0$ otherwise. The interpretation is the same as the applicant-proposing DA cutoffs.

We note that $\overline{D}_c(p_c, \mathbf{p}_{-c})$ is non-increasing in p_c and non-decreasing in $\mathbf{p}_{-c}$. This is because this property, which holds for each realization of the k -random economy, is preserved when one takes expectation to obtain $\overline{\mathbf{D}}^k$ and takes a limit along a subsequence. It then follows that $\overline{\mathbf{p}}^m$ is a monotone non-decreasing sequence and $\overline{\mathbf{q}}^m$ is a monotone non-increasing sequence, and their limits are well defined. Moreover, $\overline{\mathbf{q}} \geq \overline{\mathbf{p}}$.

Since σ is γ -regular, ROL $\rho(\theta)$ is chosen by an applicant of type θ with probability at least γ . The full support of $E = [\eta, S]$ implies that there is a positive lower bound on the density of θ on the original limit economy. Thus, the resulting limit economy induced by σ is full support in terms of ordinal preferences and scores. Then, Theorem 1-(a) of [Azevedo and Leshno \(2016\)](#) guarantees that the induced limit economy has a unique stable matching, which in turn implies that $\overline{\mathbf{p}} = \overline{\mathbf{q}}$.

The proof of Proposition 1 is identical to the proof of Proposition 2 once we observe that $\sigma = (\sigma, \sigma, \dots)$ and that the whole sequence $\{F^k, \sigma^k\}_k$ in Proposition 1 is expected-demand-convergent. We thus present only the proof for Proposition 2.

Proof of Proposition 2. Consider an arbitrary expected-demand-convergence subsequence $\{F^{k_\ell}, \sigma^{k_\ell}\}_\ell$. Recall that there exists at least one such sequence (Lemma 3).

Fix $\epsilon > 0$. We will show that for all $\ell > N$, for some $N \in \mathbb{N}$,

$$\Pr \left\{ \sup_{0 \leq i \leq k_\ell} \|\mathbf{P}_{(i)}^{k_\ell} - \overline{\mathbf{p}}\| < \epsilon \right\} > 1 - \epsilon.$$

To begin, let M be such that for all $m \geq M$, $\max\{\|\overline{\mathbf{p}} - \overline{\mathbf{p}}^m\|, \|\overline{\mathbf{p}} - \overline{\mathbf{q}}^m\|\} < \epsilon/2$. Such

an M exists due to the convergence of $\bar{\mathbf{p}}^m$ and $\bar{\mathbf{q}}^m$ to $\bar{\mathbf{p}}$ (where the latter uses the fact that $\bar{\mathbf{q}} = \bar{\mathbf{p}}$). We next observe that for each c , for any $p, p' \in [0, 1]$

$$|\bar{D}_c(p', \mathbf{p}_{-c}) - \bar{D}_c(p, \mathbf{p}_{-c})| \geq \xi \gamma |p' - p|, \forall \mathbf{p}_{-c} \in [0, 1]^{C-1}. \quad (\text{D.2})$$

To obtain (D.2), first note that $|\bar{D}_c(p', \mathbf{p}_{-c}) - \bar{D}_c(p, \mathbf{p}_{-c})| \geq |\bar{D}_c(p', \mathbf{0}) - \bar{D}_c(p, \mathbf{0})|$, where $|\bar{D}_c(p', \mathbf{0}) - \bar{D}_c(p, \mathbf{0})|$ is the mass of applicants for whom c is top-ranked by $\boldsymbol{\sigma}$ and whose scores at c are between p and p' . Assuming, without loss of generality, that $p' > p$, we have

$$\begin{aligned} |\bar{D}_c(p', \mathbf{0}) - \bar{D}_c(p, \mathbf{0})| &\geq \gamma \int_{(u,s) \in \Theta: u_c > u_{c'} \forall c' \neq c, s_c \in [p, p']} \eta(\theta) d\theta \\ &\geq \gamma \frac{\xi C}{(\bar{u} - \underline{u})^C} (p' - p) \int_{u_c > u_{c'} \forall c' \neq c} 1 du_1 \dots du_C \\ &= \xi \gamma (p' - p). \end{aligned}$$

Recall that γ is the lower bound on the probability of truth-telling and that ξ determines the lower bound on density η .

Similarly, recall from Lemma 2 that one can find $L > 0$ such that $|\bar{D}_c(p_c, \mathbf{p}_{-c}) - \bar{D}_c(p_c, \mathbf{p}'_{-c})| \leq L \|\mathbf{p}'_{-c} - \mathbf{p}_{-c}\|$ for all c . Let $\lambda := \max \left\{ 1, \frac{L}{\gamma \xi}, \frac{1}{L} \right\}$ and $\nu = \frac{1}{2M\lambda^M} \epsilon$.

It follows from Lemma 4 and the convergence of $\mathbf{S}^{k_\ell} \rightarrow \mathbf{S}$ that, for any $\nu > 0$, there exists $N(\nu)$ such that for any $\ell > N(\nu)$, we have

$$\sup_{\mathbf{p}, i} \|\mathbf{D}_{(i)}^{k_\ell}(\mathbf{p}) - \bar{\mathbf{D}}(\mathbf{p})\| + \|\mathbf{S}^{k_\ell} - \mathbf{S}\| < \nu$$

with probability at least of $1 - \nu$. Below, we fix any such $\ell > N(\nu)$ and condition on the event $\mathcal{E} := \{\sup_{\mathbf{p}, i} \|\mathbf{D}_{(i)}^{k_\ell}(\mathbf{p}) - \bar{\mathbf{D}}(\mathbf{p})\| + \|\mathbf{S}^k - \mathbf{S}\| < \nu\}$.

We argue inductively that, for each step of DA $m = 1, \dots, M$, ^{D.6} $|P_{(i),c}^{k_\ell, m} - \bar{p}_c^m| \leq m\lambda^m \nu$, for each college c . Fix any college c . Consider any m , assuming that the result holds true up to step $m - 1$. There are two possibilities. Suppose first $P_{(i),c}^{k_\ell, m} > \bar{p}_c^m \geq 0$. Then,

$$\begin{aligned} 0 &= D_c^{k_\ell}(P_{(i),c}^{k_\ell, m}, \mathbf{P}_{(i),-c}^{k_\ell, m-1}) - S_c^k \\ &\leq \bar{D}_c(P_{(i),c}^{k_\ell, m}, \mathbf{P}_{(i),-c}^{k_\ell, m-1}) - S_c + \nu \\ &\leq \bar{D}_c(P_{(i),c}^{k_\ell, m}, \bar{\mathbf{p}}_{-c}^{m-1}) - S_c + \nu + L \left\| (P_{(i),c}^{k_\ell, m}, \mathbf{P}_{(i),-c}^{k_\ell, m-1}) - (P_{(i),c}^{k_\ell, m}, \bar{\mathbf{p}}_{-c}^{m-1}) \right\| \\ &\leq \bar{D}_c(P_{(i),c}^{k_\ell, m}, \bar{\mathbf{p}}_{-c}^{m-1}) - S_c + \nu + L(m-1)\lambda^{m-1}\nu \\ &\leq \bar{D}_c(P_{(i),c}^{k_\ell, m}, \bar{\mathbf{p}}_{-c}^{m-1}) - \bar{D}_c(\bar{p}_c^m, \bar{\mathbf{p}}_{-c}^{m-1}) + \nu + L(m-1)\lambda^{m-1}\nu \end{aligned}$$

^{D.6}Recall that M is defined so that for all $m \geq M$, $\max\{\|\bar{\mathbf{p}} - \bar{\mathbf{p}}^m\|, \|\bar{\mathbf{p}} - \bar{\mathbf{q}}^m\|\} < \epsilon/2$.

$$= \overline{D}_c(P_{(i),c}^{k_\ell,m}, \overline{\mathbf{p}}_{-c}^{m-1}) - \overline{D}_c(\overline{p}_c^m, \overline{\mathbf{p}}_{-c}^{m-1}) + (1 + L(m-1)\lambda^{m-1})\nu,$$

where the first equality follows from the definition of DA cutoff at step m and upon noting that $P_{(i),c}^{k_\ell,m} > 0$ (meaning that the set over which sup is taken is well defined and the condition is an equality at $P_{(i),c}^{k_\ell,m}$); the first inequality follows as we are conditioning on event $\mathcal{E}$; the second inequality follows from the Lipschitz bound of L for $\overline{D}$; the third inequality follows from the induction hypothesis that $|P_{(i),c'}^{k_\ell,m-1} - \overline{p}_{c'}^{m-1}| \leq (m-1)\lambda^{m-1}\nu$ for any $c' \in C$; ^{D.7} and the fourth inequality follows from the definition of cutoff for the limit economy at step m (which implies $\overline{D}_c(\overline{p}_c^m, \overline{\mathbf{p}}_{-c}^{m-1}) \leq S_c$).

Rewrite the string of inequalities and use (D.2) to obtain

$$(1 + L(m-1)\lambda^{m-1})\nu \geq \overline{D}(\overline{p}_c^m, \overline{\mathbf{p}}_{-c}^{m-1}) - \overline{D}(P_{(i),c}^{k_\ell,m}, \overline{\mathbf{p}}_{-c}^{m-1}) \geq \gamma\xi(P_{(i),c}^{k_\ell,m} - \overline{p}_c^m),$$

which in turn implies that

$$P_{(i),c}^{k_\ell,m} - \overline{p}_c^m \leq \frac{(1 + L(m-1)\lambda^{m-1})}{\gamma\xi}\nu = \frac{(1 - L\lambda^{m-1} + Lm\lambda^{m-1})}{\gamma\xi}\nu \leq m\lambda^m\nu. \quad (\text{D.3})$$

Recall that $\lambda = \max\left\{1, \frac{L}{\gamma\xi}, \frac{1}{L}\right\}$.

Suppose next $\overline{p}_c^m > P_{(i),c}^{k_\ell,m} \geq 0$. Then,

$$\begin{aligned} 0 &\geq D_c^{k_\ell}(P_{(i),c}^{k_\ell,m}, \mathbf{P}_{(i),-c}^{k_\ell,m}) - S_c^k \\ &\geq \overline{D}_c(P_{(i),c}^{k_\ell,m}, \mathbf{P}_{(i),-c}^{k_\ell,m}) - S_c - \nu \\ &\geq \overline{D}_c(P_{(i),c}^{k_\ell,m}, \overline{\mathbf{p}}_{-c}^m) - S_c - \nu - L(m-1)\lambda^{m-1}\nu \\ &= \overline{D}_c(P_{(i),c}^{k_\ell,m}, \overline{\mathbf{p}}_{-c}^m) - \overline{D}_c(\overline{p}_c^m, \overline{\mathbf{p}}_{-c}^m) - \nu - L(m-1)\lambda^{m-1}\nu \\ &= \overline{D}_c(P_{(i),c}^{k_\ell,m}, \overline{\mathbf{p}}_{-c}^m) - \overline{D}_c(\overline{p}_c^m, \overline{\mathbf{p}}_{-c}^m) - (1 + L(m-1)\lambda^{m-1})\nu. \end{aligned}$$

which follows analogously to the earlier string of inequalities except that the first line is an inequality because we need to allow for $P_{(i),c}^{k_\ell,m} = 0$ and the fourth line is an equality because $\overline{p}_c^m > 0$.

As above, we use (D.2) to obtain

$$(1 + L(m-1)\lambda^{m-1})\nu \geq \overline{D}(P_{(i),c}^{k_\ell,m}, \overline{\mathbf{p}}_{-c}^m) - \overline{D}(\overline{p}_c^m, \overline{\mathbf{p}}_{-c}^m) \geq \gamma\xi(\overline{p}_c^m - P_{(i),c}^{k_\ell,m}),$$

^{D.7}Note that $P_{(i),c'}^{k_\ell,0} = \overline{p}_{c'}^0 = 0$, and hence the inequality holds for $m = 1$.

which in turn implies that

$$\bar{p}_c^m - P_{(i),c}^{k_\ell,m} \leq \frac{(1 + L(m-1)\lambda^{m-1})\nu}{\gamma\xi} \leq m\lambda^m\nu. \quad (\text{D.4})$$

Combining (D.3) and (D.4), we have

$$|P_{(i),c}^{k_\ell,m} - \bar{p}_c^m| \leq m\lambda^m\nu. \quad (\text{D.5})$$

Since this result holds for all $m = 1, \dots, M$, we now conclude that, for each c ,

$$P_{(i),c}^{k_\ell} \geq P_{(i),c}^{k_\ell,M} \geq \bar{p}_c^M - M\lambda^M\nu \geq \bar{p}_c - M\lambda^M\nu - \frac{\epsilon}{2}, \quad (\text{D.6})$$

where the first inequality follows from the fact that $P_{(i),c}^{k_\ell}$ is the limit of a monotone non-decreasing sequence $(P_{(i),c}^{k_\ell,m})_m$ as $m \rightarrow \infty$, the second inequality follows from (D.5), and the third follows from the definition of M .

The exact same argument works for the CPDA process. Namely, for each c , $|Q_{(i),c}^{k_\ell,m} - \bar{q}_c^m| \leq m\lambda^m\nu$ for $m = 1, \dots, M$ so that we have

$$Q_{(i),c}^{k_\ell} \leq Q_{(i),c}^{k_\ell,M} \leq \bar{q}_c^M + M\lambda^M\nu \leq \bar{q}_c + M\lambda^M\nu + \frac{\epsilon}{2}. \quad (\text{D.7})$$

Combining (D.6) and (D.7), recalling $\bar{p}_c = \bar{q}_c$ and $Q_{(i),c}^{k_\ell} \geq P_{(i),c}^{k_\ell}$, $\forall c$, we obtain

$$\sup_{0 \leq i \leq k_\ell} \|\mathbf{P}_{(i)}^{k_\ell} - \bar{\mathbf{p}}\| \leq M\lambda^M\nu + \frac{\epsilon}{2},$$

for all $0 \leq i \leq k_\ell$. As $\nu = \frac{1}{2M\lambda^M}\epsilon$, then, for all $\ell > N(\nu)$, where $N(\nu) \in \mathbb{N}$ is defined above, we have

$$\sup_{0 \leq i \leq k_\ell} \|\mathbf{P}_{(i)}^{k_\ell} - \bar{\mathbf{p}}\| \leq \epsilon,$$

in the event $\mathcal{E}$, which occurs with probability at least $1 - \frac{1}{2M\lambda^M}\epsilon \geq 1 - \epsilon$. $\square$

E Proofs of Theorems

E.1 Proof of Theorem 1

Proof of Theorem 1. Let $\mathbf{p} = \bar{\mathbf{p}}(\rho)$, where $\bar{\mathbf{p}}(\rho)$ is the unique market-clearing cutoff for the limit demand system induced by TT. In this proof, we refer to $\bar{\mathbf{p}}(\rho)$ simply as $\bar{\mathbf{p}}$.

Recall that

$$\Theta^\delta(\bar{\mathbf{p}}) := \{(\mathbf{u}, \mathbf{s}) \in \Theta \mid \exists j \in \mathbf{C} \text{ s.t. } |s_j - \bar{p}_j| \leq \delta\}$$

is the set of types whose score for some college is δ -close to its market-clearing cutoff in

the limit demand system.

For each type $\theta = (\mathbf{u}, \mathbf{s})$, there exists at least one SRS strategy against $\bar{\mathbf{p}}$ that violates WTT (see footnote 10); denote this strategy by $\hat{R}(\theta)$. The applicants with types $\theta \in \Theta^\delta(\bar{\mathbf{p}})$ play $\rho(\theta)$ and the applicants with types $\theta \notin \Theta^\delta(\bar{\mathbf{p}})$ randomize between $\rho(\theta)$ (with probability γ) and $\hat{R}(\theta)$ (with probability $1 - \gamma$).

Fix any $\epsilon > 0$. Take any $\epsilon' > 0$ such that $\epsilon'(\bar{u} - \underline{u}) < \epsilon$. By Proposition 1, there exists $K \in \mathbb{N}$ such that for all $k > K$, $\Pr\{\|\mathbf{P}_{(i)}^k - \bar{\mathbf{p}}\| < \delta\} \geq 1 - \epsilon'$, where $\mathbf{P}_{(i)}^k$ is the vector of cutoffs associated with the matching in F^k under the prescribed strategy with at most one applicant deviating to TT. Let $\mathcal{E}^k$ denote the event where $\|\mathbf{P}_{(i)}^k - \bar{\mathbf{p}}\| < \delta$ holds. We now show that the prescribed strategy profile forms an interim ϵ -Bayesian Nash equilibrium for each k -random economy for $k > K$.

First, for any type $\theta \in \Theta^\delta(\bar{\mathbf{p}})$, the prescribed strategy, $\rho(\theta)$, is trivially optimal given the strategy-proofness of DA. Consider an applicant with any type $\theta \notin \Theta^\delta(\bar{\mathbf{p}})$, and suppose that all other applicants employ the prescribed strategy. Now condition on event $\mathcal{E}^k$. Recall that the set of feasible colleges is the same for type $\theta \notin \Theta^\delta(\bar{\mathbf{p}})$ whether the cutoffs are $\hat{\mathbf{P}}_{(i)}^k$ or $\bar{\mathbf{p}}$, provided that $\|\hat{\mathbf{P}}_{(i)}^k - \bar{\mathbf{p}}\| < \delta$. Hence, given event $\mathcal{E}^k$, strategy $\hat{R}(\theta)$ is a best response and the prescribed mixed strategy attains the maximum payoff for type $\theta \notin \Theta^\delta(\bar{\mathbf{p}})$.

Of course, the event $\mathcal{E}^k$ may not occur, but that probability is no greater than ϵ' for $k > K$, and the maximum payoff loss in that case from failing to play her best response is $\bar{u} - \underline{u}$. Hence, the payoff loss she incurs by playing the prescribed mixed strategy is at most

$$\epsilon'(\bar{u} - \underline{u}) < \epsilon.$$

This proves that the strategy profile forms a robust equilibrium. $\square$

E.2 Proof of Theorem 2

Proof of Theorem 2. Fix any γ -regular robust equilibrium strategy profile σ , for any arbitrary $\gamma \in (0, 1]$. Suppose to the contrary that σ is not asymptotically stable. Then, by definition, there exists $\varepsilon > 0$ and a subsequence of finite economies $\{F^{k_j}\}_j$ such that, for all k_j ,

$$\Pr\left(\text{The fraction of applicants playing SRS against } \mathbf{P}^{k_j} \text{ is at least } 1 - \varepsilon\right) < 1 - \varepsilon, \quad (\text{E.8})$$

where the applicants play σ^{k_j} , a k_j -truncation of σ .

By Lemma 4, there exists a subsubsequence $\{\mathbf{D}^{k_{j_\ell}}\}_\ell$ that converges to $\bar{\mathbf{D}}$ uniformly in probability. By Proposition 2, $\mathbf{P}_{(i)}^{k_{j_\ell}}$ converges to $\bar{\mathbf{p}}$ uniformly over i in probability, where $\bar{\mathbf{p}}$ is the deterministic cutoffs defined by the limit of demands.

Define a set of applicants

$$\hat{\Theta} := \{(\mathbf{u}, \mathbf{s}) : |u_c - u_{c'}| > \delta \text{ for all } c \neq c'\} \cap \{(\mathbf{u}, \mathbf{s}) : |s_c - \bar{p}_c| > \delta \text{ for all } c\}.$$

These are the applicants whose payoffs from distinct colleges differ at least by δ and whose score at each college c differs from its limit economy cutoff $\bar{p}_c$ at least by δ .

Take δ to be small enough s.t. $\eta(\hat{\Theta}) > (1 - \varepsilon)^{1/3}$. This can be done since η is absolutely continuous.

By WLLN, we know that $\eta^{k_{j\ell}}(\hat{\Theta})$ converges to $\eta(\hat{\Theta})$ in probability, and therefore there exists L_1 such that for all $\ell > L_1$ we have

$$\Pr\left(\eta^{k_{j\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2}\right) \geq (1 - \varepsilon)^{1/2}. \quad (\text{E.9})$$

Consider the event

$$A^{k_{j\ell}} := \left\{ \sup_{0 \leq i \leq k_{j\ell}} \|\mathbf{P}_{(i)}^{k_{j\ell}} - \bar{\mathbf{p}}\| < \delta \right\}.$$

Since $\mathbf{P}_{(i)}^{k_{j\ell}} \xrightarrow{p} \bar{\mathbf{p}}$ uniformly over $i \in \mathbb{N}_0$, there exists L_2 such that, for all $\ell > L_2$, we have

$$\Pr\left(A^{k_{j\ell}}\right) \geq \max\left\{(1 - \varepsilon)^{1/6}, 1 - (1 - \varepsilon)^{1/2} \left[(1 - \varepsilon)^{1/3} - (1 - \varepsilon)^{1/2}\right]\right\}. \quad (\text{E.10})$$

Since σ forms a robust equilibrium, there exists L_3 such that for all $\ell > L_3$, $\sigma^{k_{j\ell}}$ is a $\delta \left[(1 - \varepsilon)^{1/6} - (1 - \varepsilon)^{1/3}\right]$ -BNE for economy $F^{k_{j\ell}}$.

By WLLN, there exists $\hat{L} \in \mathbb{N}$ such that $\hat{L}$ i.i.d. Bernoulli random variables with parameter $p = (1 - \varepsilon)^{1/3}$ have a sample mean greater than $(1 - \varepsilon)^{1/2}$ with probability no less than $(1 - \varepsilon)^{1/3}$. Next, let L_4 be such that $\ell > L_4$ implies $(1 - \varepsilon)^{1/2} k_{j\ell} > \hat{L}$.

Now let's fix an arbitrary $\ell > \max\{L_1, L_2, L_3, L_4\}$. We wish to show that in economy $F^{k_{j\ell}}$,

$$\Pr\left(\text{The fraction of applicants playing SRS against } \mathbf{P}^{k_{j\ell}} \text{ is no less than } 1 - \varepsilon\right) \geq 1 - \varepsilon,$$

which would contradict (E.8) and complete the proof.

We first prove that in economy $F^{k_{j\ell}}$, an applicant with $\theta \in \hat{\Theta}$ plays SRS against $\bar{\mathbf{p}}$ with probability no less than $(1 - \varepsilon)^{1/3}$. To see this, suppose to the contrary that there exists some applicant i and some type $\theta \in \hat{\Theta}$ such that

$$\Pr(\sigma_i(\theta) \text{ plays SRS against } \bar{\mathbf{p}}) < (1 - \varepsilon)^{1/3}.$$

Suppose now the applicant i deviates to TT. By doing so, she will do weakly better in all circumstances (since TT is a dominant strategy) and strictly so by at least δ (since $\theta \in \hat{\Theta}$) conditional on the deviation changing her assignment. Her assignment would change (at least) whenever she was not playing SRS against $\bar{\mathbf{p}}$ under $\sigma_i(\theta)$ and event $A^{k_{j\ell}}$ occurs. This is because in event $A^{k_{j\ell}}$, the strategy $\sigma_i(\theta)$ is SRS against $\mathbf{P}_{(0)}^{k_{j\ell}}$ if and only if $\sigma_i(\theta)$ is

SRS against $\bar{\mathbf{p}}$ for type $\theta \in \hat{\Theta}$, and deviating to truthful reporting would produce a stable assignment against $\bar{\mathbf{p}}$ for such a type. In sum, applicant i with type $\theta \in \hat{\Theta}$ would gain from deviation by at least

$$\begin{aligned} & \delta \cdot \Pr \left(\sigma_i(\theta) \text{ is not SRS against } \bar{\mathbf{p}} \text{ and event } A^{k_{j_\ell}} \text{ occurs} \right) \\ & \geq \delta \left[\Pr \left(A^{k_{j_\ell}} \right) - \Pr \left(\sigma_i(\theta) \text{ plays SRS against } \bar{\mathbf{p}} \right) \right] \\ & \geq \delta \left[(1 - \varepsilon)^{1/6} - (1 - \varepsilon)^{1/3} \right], \end{aligned}$$

The above inequalities contradict the construction of L_3 .^{E.8} Therefore, in economy $F^{k_{j_\ell}}$, for each applicant $i = 1, \dots, k_{j_\ell}$ and each $\theta \in \hat{\Theta}$, we have

$$\begin{aligned} & \Pr \left(\sigma_i(\theta) \text{ plays SRS against } \bar{\mathbf{p}} \mid \eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \right) \\ & = \Pr \left(\sigma_i(\theta) \text{ plays SRS against } \bar{\mathbf{p}} \right) \geq (1 - \varepsilon)^{1/3}, \end{aligned} \tag{E.11}$$

where the first equality holds because applicant i 's choice of a mixed strategy is independent of random draws of the applicants' types.^{E.9}

It then follows that

$$\begin{aligned} & \Pr \left(\begin{array}{l} \text{The fraction of applicants with } \theta \in \hat{\Theta} \\ \text{playing SRS against } \bar{\mathbf{p}} \text{ is no less than } (1 - \varepsilon)^{1/2} \end{array} \mid \eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \right) \\ & \geq \Pr \left(\begin{array}{l} \eta^{k_{j_\ell}}(\hat{\Theta}) \cdot k_{j_\ell} \text{ i.i.d. Bernoulli random variables with} \\ p = (1 - \varepsilon)^{1/3} \text{ have a sample mean no less than } (1 - \varepsilon)^{1/2} \end{array} \mid \eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \right) \\ & \geq \Pr \left(\begin{array}{l} \hat{L} \text{ i.i.d. Bernoulli random variables with } p = (1 - \varepsilon)^{1/3} \\ \text{have a sample mean no less than } (1 - \varepsilon)^{1/2} \end{array} \right) \\ & \geq (1 - \varepsilon)^{1/3}, \end{aligned} \tag{E.12}$$

where the first inequality follows from (E.11) and the fact that $\sigma_i(\theta)$'s are independent across applicants, and the second inequality holds since $\ell > L_4$ and since, by the definition of L_4 , for any such ℓ , $\eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2}$ implies $\eta^{k_{j_\ell}}(\hat{\Theta}) \cdot k_{j_\ell} > \hat{L}$.

Comparing the finite economy random cutoffs $\mathbf{P}^{k_{j_\ell}}$ with the deterministic limit cutoffs

^{E.8}Recall that L_3 was defined so that $\ell > L_3$ means that the strategy profile $\sigma^{k_{j_\ell}}$ is a $\delta \left[(1 - \varepsilon)^{1/6} - (1 - \varepsilon)^{1/3} \right]$ -BNE for the economy $F^{k_{j_\ell}}$.

^{E.9}In other words, how a fixed type plays in equilibrium does not depend on how many of them are drawn.

$\bar{\mathbf{p}}$ yields:

$$\begin{aligned}
& \Pr \left(\begin{array}{l} \text{The fraction of applicants with } \theta \in \hat{\Theta} \\ \text{playing SRS against } \mathbf{P}^{k_{j_\ell}} \text{ is no less than } (1 - \varepsilon)^{1/2} \end{array} \middle| \eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \right) \\
& \geq \Pr \left(\begin{array}{l} \text{The fraction of applicants with } \theta \in \hat{\Theta} \\ \text{playing SRS against } \bar{\mathbf{p}} \text{ is no less than } (1 - \varepsilon)^{1/2} \\ \text{and event } A^{k_{j_\ell}} \text{ occurs} \end{array} \middle| \eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \right) \\
& \geq \Pr \left(\begin{array}{l} \text{The fraction of applicants with } \theta \in \hat{\Theta} \\ \text{playing SRS against } \bar{\mathbf{p}} \text{ is no less than } (1 - \varepsilon)^{1/2} \end{array} \middle| \eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \right) \\
& \quad - \Pr \left(A^{k_{j_\ell}} \text{ does not occur} \middle| \eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \right) \\
& \geq (1 - \varepsilon)^{1/3} - \frac{1 - \Pr(A^{k_{j_\ell}})}{\Pr(\eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2})} \\
& \geq (1 - \varepsilon)^{1/3} - \frac{(1 - \varepsilon)^{1/2} [(1 - \varepsilon)^{1/3} - (1 - \varepsilon)^{1/2}]}{(1 - \varepsilon)^{1/2}} \\
& = (1 - \varepsilon)^{1/2},
\end{aligned} \tag{E.13}$$

where the first inequality follows since in event $A^{k_{j_\ell}}$, the strategy $\sigma_i(\theta)$ is SRS against $\mathbf{P}^{k_{j_\ell}}$ if and only if $\sigma_i(\theta)$ is SRS against $\bar{\mathbf{p}}$ for type $\theta \in \hat{\Theta}$; the third inequality follows from (E.12); and the fourth inequality follows from (E.10).

We finally have in economy $F^{k_{j_\ell}}$

$$\begin{aligned}
& \Pr \left(\text{The fraction of applicants playing SRS against } \mathbf{P}^{k_{j_\ell}} \text{ is no less than } 1 - \varepsilon \right) \\
& \geq \Pr \left(\begin{array}{l} \text{At least a fraction } (1 - \varepsilon)^{1/2} \text{ of applicants with } \theta \in \hat{\Theta} \text{ play SRS against } \mathbf{P}^{k_{j_\ell}} \\ \text{and } \eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \end{array} \right) \\
& = \Pr \left(\eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \right) \\
& \quad \times \Pr \left(\begin{array}{l} \text{at least a fraction } (1 - \varepsilon)^{1/2} \text{ of applicants} \\ \text{with } \theta \in \hat{\Theta} \text{ play SRS against } \mathbf{P}^{k_{j_\ell}} \end{array} \middle| \eta^{k_{j_\ell}}(\hat{\Theta}) \geq (1 - \varepsilon)^{1/2} \right) \\
& \geq (1 - \varepsilon)^{1/2} \cdot (1 - \varepsilon)^{1/2} \\
& = 1 - \varepsilon,
\end{aligned}$$

where the second inequality follows from the construction of L_1 (see (E.9)) and from (E.13).

Therefore, we have obtained a contradiction to (E.8). $\square$

F Implications for Market Design Research

Our theoretical results, along with existing the field evidence, suggest that non-truthful behavior may be widespread but it rarely leads to payoff consequences. This observation has implications for empirical market design. In simulated data, we illustrate these implications for two exercises common in the literature: the estimation of applicant preferences and the analysis of a counterfactual policy.

F.1 Estimating Applicant Preferences

There are two typical identifying assumptions in the literature for the estimation of applicant preferences.

The first is WTT (Hällsten, 2010; Kirkebøen, 2012). Recall that applicant i is WTT if her submitted ROL ranks her most-preferred colleges according to her true preferences, while every unranked college is less desirable to i than any ranked college. Let $\succ_i$ denote the inferred preference relation of i . Consider an applicant whose submitted ROL is c_3 - c_1 , while there are four colleges available, $\{c_1, c_2, c_3, c_4\}$. WTT infers that $c_3 \succ_i c_1 \succ_i c_2, c_4$.

The second assumption is stability (Akyol and Krishna, 2017; Bucarey, 2018; Combe, Tercieux, and Terrier, 2016; Fack, Grenet, and He, 2019). An outcome is stable if every applicant is matched with her most-preferred college among the feasible ones. Suppose that the aforementioned applicant has c_2 and c_3 feasible and thus is matched with c_3 . Stability infers $c_3 \succ_i c_2$. In contrast to WTT, stability does not make any inference about infeasible colleges, c_1 and c_4 .

According to our Theorem 1, in a robust equilibrium, applicant i may not be WTT; hence incorrect WTT inference. For example, she may rank more desirable but infeasible c_4 arbitrarily. Stability makes inference only about feasible colleges and, according to our Theorem 2, is satisfied asymptotically. From WTT to stability, we gain robustness to untruthfulness but utilize less information. Therefore, the estimation based on stability will be less efficient than WTT, yet less likely to be biased because it uses fewer possibly incorrectly inferred preference relations.

F.1.1 Monte Carlo Simulations

We evaluate the performance of WTT and stability in simulated data that resembles the typical college admissions studied in the previous sections. The environment has 12 colleges and 1800 applicants. Colleges rank applicants by an ex ante known score. Applicant preferences follow a conditional logit model that depends on the distance to each college and applicant type, $T_i \in \{0, 1\}$. T_i is equal to one for two-thirds of the applicants whose

score is below the median, and we call them disadvantaged. A disadvantaged applicant value an odd-indexed college more than a $T_i = 0$ applicant, *ceteris paribus*.

We consider three data generating processes (DGPs). The first is the *Truth-Telling* (TT): every applicant truthfully ranks *all* colleges. The second is *Payoff Irrelevant Mistakes* (PIM): a fraction of applicants skip colleges with which they are never matched in the simulations. We thus expect that stability is satisfied in both TT and PIM. The last DGP is *Payoff Relevant Mistakes* (PRM): in addition to skipping never-matched colleges, applicants may skip some colleges with which they have a low match probability, leading to some payoff-relevant mistakes or violations of stability. According to our theory, these mistakes, although rare, can happen in a robust equilibrium of a finite economy. The three DGPs, each of which has 150 simulation samples, are summarized in Table F.1.

Table F.1: ROLs and Mistakes in Monte Carlo Simulations

	Data Generating Processes (DGPs) w/ Different Applicant Strategies		
	Truth-Telling (TT)	Payoff Irrelevant Mistakes (PIM)	Payoff Relevant Mistakes (PRM)
Average length of submitted ROLs	12	7.34	6.58
WTT: <i>Weak Truth-Telling</i> (%) ^a	100	50	44
Matched w/ favorite feasible college (%) ^b	100	100	97

Notes: Each entry in the table is an average over the 150 simulation samples for a given DGP. In each sample, there are 1800 applicants and 12 colleges with a total of 1500 seats. ^aAn applicant is WTT if she truthfully ranks her top K_i ($1 \leq K_i \leq 12$) preferred colleges, where K_i is the observed number of colleges ranked by i . Omitted colleges are always less preferred than any ranked college. ^bA college is feasible to an applicant, if the applicant's score is above the college's ex-post admission cutoff.

An applicant's utility function has four unknown parameters, $\beta_1, \dots, \beta_4$. To estimate the parameters from the data, we apply a rank-ordered logit model when assuming WTT and a conditional logit model when assuming stability to estimate four parameters. Appendix G provides more details.

Bias-variance tradeoff. We illustrate several patterns in the estimation in Figure F.1 for one of the parameters, $\beta_1 = 0.3$. When applicants report truthfully, WTT and stability are both consistent but WTT is more efficient (panel a). However, WTT leads to a biased estimator whenever some applicants are not truthful (panels b–c). In contrast, stability performs well when there are no or just a few payoff-relevant mistakes. These results illustrate a bias-variance tradeoff: from WTT to stability, the variance of the estimator increases while the bias decreases whenever it exists.

Mis-Estimated Preferences. A direct consequence of an inconsistent estimator is the mis-estimation of applicant preferences. As an example, let us consider colleges 10 and 11. For a disadvantaged applicant ($T_i = 1$) with an equal distance to these two colleges, the true probability that she prefers college 11 to college 10 is 0.91 (the straight, dashed line

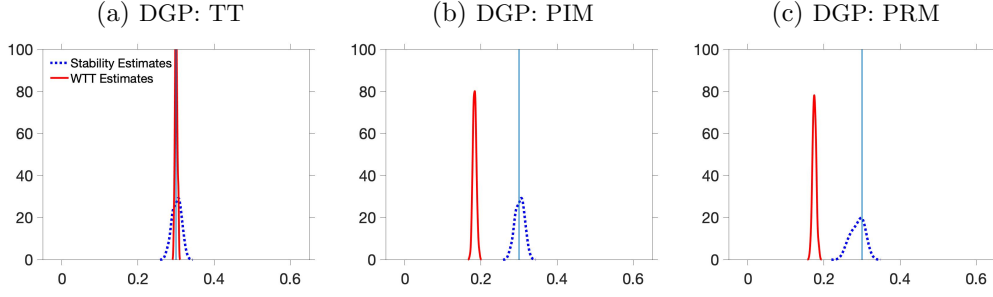


Figure F.1: Distribution of Estimates based on Weak Truth-Telling or Stability ($\beta_1 = 0.3$)

Notes: The figures focus on the estimates of one parameter ($\beta_1 = 0.3$) from two approaches, weakly truth-telling (WTT, the solid line) and stability (the dotted line). Each panel uses the 150 simulation samples given a DGP and reports an estimated density of the estimates based on a normal kernel function. See Table F.1 for more details on the three DGPs.

in Figure F.2). Using the logit formula, we calculate the same probability based on the two sets of estimates, and Figure F.2 presents the average across the 150 samples given an estimation approach and a DGP. Clearly, WTT produces significant biases in PIM and PRM, while stability only leads to a small bias in PRM.

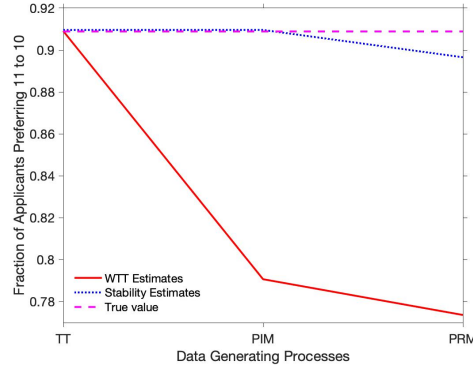


Figure F.2: True and Estimated Probabilities That an Applicant Prefers College 11 to College 10

Notes: The figure presents the probability that a disadvantaged applicant ($T_i = 1$), with an equal distance to both colleges, prefers college 11 to college 10. The true value is 0.91 (the straight, dashed line). With the logit formula, we calculate the probability based on the WTT-based estimates, and the solid line presents the average over the 150 simulation samples in each DGP. Similarly, the dotted line describes those from the stability-based estimates.

F.2 Counterfactual Analysis

Making policy recommendations based on counterfactual analysis is one of the main objectives of market design research, for which our theoretical results also provide some implications.

The literature has two types of approaches to counterfactual analysis. The first is based on *submitted ROLs*. See, for example, the analysis of National Resident Matching Program by Roth and Peranson (1999) and kindergarten allocation in Estonia by Veski, Biró, Pöder, and Lauri (2017). It is assumed that submitted ROLs under the existing policy are true ordinal preferences and that an applicant will submit the same ROL under the counterfactual policy. Our Theorem 1 implies that this assumption is not plausible in a robust equilibrium.

In the second type of approaches, the researcher uses estimated preferences and let every applicant submit a truthful ROL under the counterfactual policy. Estimated preferences can be obtained based on WTT or stability as discussed in Section F.1. Assuming truth-telling in the counterfactual is justified by Corollary 2, as any regular robust equilibrium leads to an asymptotically stable matching that is well approximated by the stable matching from truthful reporting.

As an illustration, we use the Monte Carlo simulations in Section F.1 and consider a counterfactual policy in which applicants with $T_i = 1$ are given priorities over those with $T_i = 0$, while applicants of the same type are still ranked according to their scores. The mechanism is still DA in which everyone can rank all colleges.

F.2.1 Performance in Monte Carlo Simulations

Recall that we have 150 samples in the simulations for each DGP (TT, PIM, or PRM). Additionally, for each DGP, we generate the true outcome under the counterfactual as a benchmark. That is, we assume applicants potentially make mistakes under the counterfactual policy as they do under the current policy.

We focus on the three approaches to counterfactual analysis: submitted ROLs, the WTT-based estimation, and the stability-based estimation. We calculate how each approach perform in terms of predicting the new policy’s effects on outcomes and on welfare.

Mis-predicted Cutoffs. An informative statistic of an outcome is college cutoffs. Figure F.3 shows, given each DGP, how the three approaches mis-predict cutoffs under the counterfactual policy. For each college, indexed from 1 to 12, we calculate the average difference between the predicted cutoffs and the true cutoffs across the 150 simulation samples.

In panel (a), the DGP is TT, and thus the submitted ROLs coincide with true ordinal preferences. Consequently, the predicted cutoffs from the submitted-ROLs approach are the true ones. The other two approaches also lead to almost the same cutoffs.

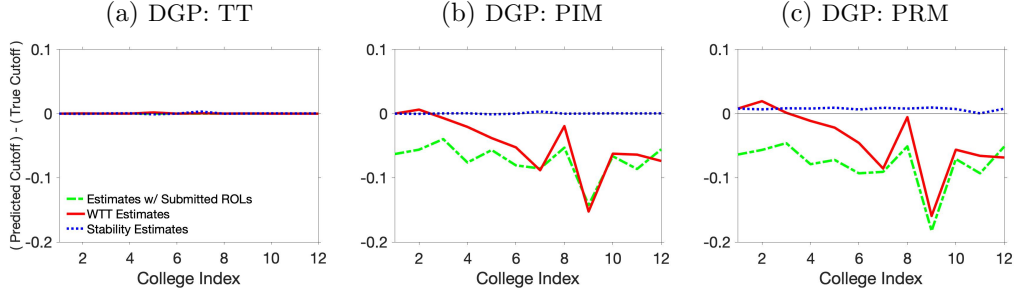


Figure F.3: Comparison of the Three Approaches: Biases in Predicted Cutoffs

Notes: In a given DGP, each panel presents how the predicted cutoffs from each approach differ from the true ones that are simulated based on the actual behavior (i.e., the true preferences with possible mistakes). Given a DGP, we simulate the colleges' cutoffs following each approach and calculate the mean deviation from the true ones.

In panel (b), which corresponds to DGP PIM, only the stability-based estimation is consistent, and indeed it has the smallest mis-predictions relative to the other two. As applicants tend to omit popular colleges, which have higher indices in our setting, from their submitted ROLs in this DGP, the approaches based on WTT and submitted ROLs systematically underestimate the demand for these colleges and thus their cutoffs.

When the DGP contain payoff-relevant mistakes (PRM, in panel c), none of the approaches is unbiased. However, the stability-based estimates seem to have a negligible mis-prediction compared to the other two.

Mis-predicted Matches. Panel (a) of Figure F.4 further shows how each of the three approaches mis-predicts individual outcomes for applicants with $T_i = 1$. Recall that the counterfactual policy is intended to help those applicants. The stability-based approach incorrectly predicts the match for 4.5 percent of applicants on average whenever stability is satisfied (in DGPs TT and PIM). Even in PRM, its mis-prediction rate is merely 7.7 percent. The WTT-based approach has a lower mis-prediction rate in DGP TT but underperforms relative to stability in the other two DGPs. The submitted-ROLs approach has the highest mis-prediction rates in all DGPs except TT. Among the applicants with $T_i = 0$, Figure G.6 in Appendix G shows that the comparison of the approaches follows the same pattern.

Mis-predicted Welfare Effects. We now investigate the welfare effects on the $T_i = 1$ applicants of the counterfactual policy. Given a simulation sample and a DGP, we compare the outcomes of each applicant under the two policies. If the applicant is matched with a “more-preferred” college according to the true/estimated preferences, she is a *winner*; she

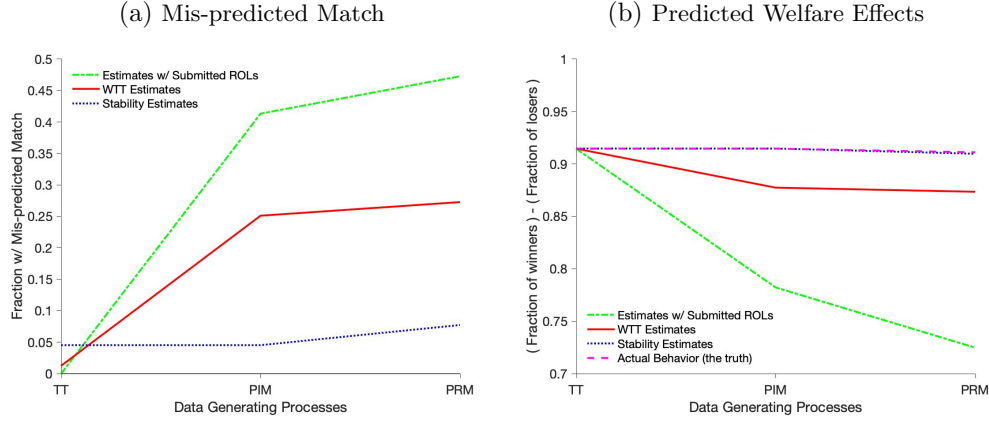


Figure F.4: Three Approaches to Counterfactual Analysis: Disadvantaged Applicants $T_i = 1$

Notes: The figure shows the averages among $T_i = 1$ applicants across the 150 samples in each DGP. On average, there are 599 such applicants in a sample. Given a DGP, we simulate an outcome under the counterfactual policy and compare it to the truth from the actual behavior (i.e., the true preferences with possible mistakes). Panel (a) shows the average mis-prediction rates. Panel (b) shows the predicted welfare effects by each approach. It is measured by the difference between the fractions of winners and losers. See Table G.3 in Appendix G for more details.

is a *loser* if she is matched with a “less-preferred” one.^{F.10}

Panel (b) of Figure F.4 shows the difference between the fractions of winners and losers, averaged across the 150 samples.^{F.11} Among the $T_i = 1$ applicants, the stability-based estimates are almost identical to the truth, even in the DGP with payoff-relevant mistakes (PRM). In contrast, the other two approaches’ predictions are close to the true value only in DGP TT; both tend to be biased toward a zero effect when applicants make mistakes (DGPs PIM and PRM). The reason for the bias is clear. Under WTT, the preferences for popular colleges are underestimated. Meanwhile, the submitted-ROLs approach ignores the likely changes in ROLs under the new policy. In particular, disadvantaged applicants find previously out-of-reach colleges now within reach, so they may include these colleges in their ROLs.

^{F.10}Because each approach to counterfactual analysis estimates applicant preferences in a unique way, an applicant’s utility associated with a college can differ across the approaches. Therefore, the measured welfare effects of the counterfactual policy may differ even when an applicant is matched with the same college.

^{F.11}The outcome does not change for 9–28 percent of applicants. See Table G.3 for more detailed summary statistics.

G Monte Carlo Simulations

Complementing Section F, this appendix provides additional details on the Monte Carlo simulations that we perform to assess the implications of our theoretical results. Section G.1 specifies the environment, Section G.2 describes the data generating processes, Section G.3 presents the estimation and the results, and, finally, Section G.4 presents some additional results on the counterfactual analysis.

G.1 Simulated Environment

We consider a finite economy in which $k = 1,800$ applicants compete for admission to $C = 12$ colleges. The vector of college capacities is specified as follows:

$$\{S_c\}_{c=1}^{12} = \{150, 75, 150, 150, 75, 150, 150, 75, 150, 150, 75, 150\}.$$

Setting the total capacity of colleges (1,500 seats) to be strictly smaller than the number of applicants (1,800) ensures that each college has a strictly positive cutoff in equilibrium.

The economy is located in an area within a circle of radius 1. The applicants are uniformly distributed across the area, and the colleges are evenly located on a circle of radius 1/2 around the centroid. The Cartesian distance between applicant i and college c is denoted by $d_{i,c}$.

Applicants are matched with colleges through a serial dictatorship, a special case of DA. Applicants are asked to submit an ROL of colleges, and there is no limit on the number of choices to be ranked. Without loss of generality, colleges have a priority structure such that all colleges rank applicant i ahead of i' if $i' < i$. One may consider the order being determined by certain test scores.

To represent applicant preferences over colleges, we adopt a parsimonious random utility model without an outside option. As is traditional and more convenient in empirical analysis, we now let the applicant utility functions take any value on the real line; we continue to use u as a notation for utility functions. That is, applicant i 's utility from being matched with college c is specified as follows:

$$u_{i,c} = \beta_1 \cdot c + \beta_2 \cdot d_{i,c} + \beta_3 \cdot T_i \cdot A_c + \beta_4 \cdot Small_c + \epsilon_{i,c}, \forall i \text{ and } c, \quad (\text{G.14})$$

where $\beta_1 \cdot c$ is college c 's baseline quality; $d_{i,c}$ is the distance from applicant i 's location to college c ; $T_i = 1$ or 0 is applicant i 's type (e.g., disadvantaged or not); $A_c = 1$ or 0 is college c 's type (e.g., known for resources for disadvantaged applicants); $Small_c = 1$ if college c is small, 0 otherwise; and $\epsilon_{i,c}$ is a type-I extreme value and i.i.d. across i and c .

The type of college c , A_c , is 1 if c is an odd number; otherwise, $A_c = 0$. The type of applicant i , T_i , is 1 with probability 2/3 among the lower-ranked applicants ($i \leq 900$); $T_i = 0$ for all $i > 900$. This way, we may consider those with $T_i = 1$ as the disadvantaged.

The coefficients of interest are $(\beta_1, \beta_2, \beta_3, \beta_4)$ which are fixed at $(0.3, -1, 2, 0)$ in the

simulations. By this specification, colleges with larger indices are of higher quality, and $Small_c$ does not affect applicant preference. The purpose of estimation is to recover these coefficients and therefore the distribution of preferences.

G.2 Data Generating Processes

Each simulation sample contains an independent preference profile obtained by randomly drawing $\{d_{i,c}, \epsilon_{i,c}\}_c$ and T_i for all i from the distributions specified above. In all samples, applicant scores, college capacities, and college types (A_c) are kept constant.

We first simulate the joint distribution of the 12 colleges' cutoffs by letting every applicant submit an ROL ranking all colleges truthfully. After running the serial dictatorship, we calculate the cutoffs in each simulation sample. Figure G.5 shows the marginal distribution of each college's cutoff from the 1000 samples. Note that colleges with smaller capacities tend to have higher cutoffs. For example, college 11, with 75 seats, often has the highest cutoff, although college 12, with 150 seats, has the highest baseline quality.

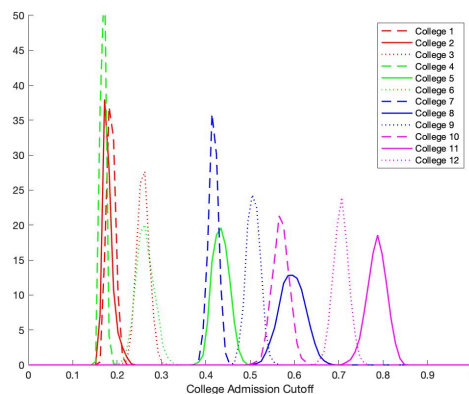


Figure G.5: Simulated Distribution of Cutoffs when Everyone is Truth-telling

Notes: Assuming everyone is truth-telling, we calculate the cutoffs of all colleges in each simulation sample. The figure shows the marginal distribution of each college's cutoff, in terms of percentile rank (between 0 (lowest) and 1 (highest)). Each curve is an estimated density based on a normal kernel function. A solid line indicates a small college with 75, instead of 150, seats. The simulation samples for cutoffs use independent draws of $\{d_{i,c}, \epsilon_{i,c}\}_c$ and T_i .

To generate data on applicant behaviors and outcomes, we simulate another 150 samples with new independent draws of $\{d_{i,c}, \epsilon_{i,c}\}_c$ and T_i for all i . These samples are used for the estimation and counterfactual analysis, and, in each of them, we consider three types of data generating processes (DGPs) with different applicant strategies.

- (i) **TT (Truth-Telling):** Every applicant submits an ROL of 12 colleges according to

her true preferences. Because everyone finds every college acceptable, this is TT as defined in our theoretical model.^{G.12}

- (ii) **PIM (Payoff Irrelevant Mistakes)**: A fraction of applicants omit from their ROLs some of the colleges with which they are never matched according to the simulated distribution of cutoffs. For a given applicant, an omitted college may have a high (expected) cutoff and thus be “out of reach;” alternatively, an omitted college may have a low cutoff, but the applicant is always accepted by one of her more-preferred colleges. There are 55 percent of applicants who omit at least one college. As applicants with $T_i = 1$ have lower scores, they are more likely to omit than those with $T_i = 0$: 61 percent of $T_i = 1$ drop at least one college, compared to 51 percent of $T_i = 0$. Among applicants who are never matched with any college, we randomly choose some colleges for them to include in their ROLs.
- (iii) **PRM (Payoff Relevant Mistakes)**: Taking the data generated under PIM, we let more applicants to omit never-matched colleges and also let some of them make payoff-relevant mistakes. That is, some applicants omit some of the colleges with which they have a chance of being matched lower than 30 percent according to the simulated distribution of cutoffs. Recall that the joint distribution of cutoffs is only simulated once under the assumption that everyone is truth-telling. On average, 60 percent of applicants drop at least one college.

To summarize, for each of the 150 samples, we simulate the matching game 3 times: TT (Truth-Telling), PIM (payoff-irrelevant mistakes), and PRM (payoff-relevant mistakes). See Table F.1 in the main text for summary statistics.

G.3 Estimation and Results

With the simulated data, the random utility model described by equation (G.14) is estimated under two different identifying assumptions.

We first re-write the random utility model (equation G.14) as follows:

$$\begin{aligned} u_{i,c} &= \beta_1 \cdot c + \beta_2 \cdot d_{i,c} + \beta_3 \cdot T_i \cdot A_c + \beta_4 \cdot Small_c + \epsilon_{i,c} \\ &\equiv V_{i,c} + \epsilon_{i,c}, \forall i = 1, \dots, k \text{ and } c = 1, \dots, C; \end{aligned}$$

we also define $\mathbf{X}_i = (\{d_{i,c}, A_c, Small_c\}_c, T_i)$ to denote the observable applicant characteristics and college attributes; and $\boldsymbol{\beta}$ is the vector of coefficients, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$.

The key for each estimation approach is to characterize the choice probability of each ROL or each college, where the uncertainty originates from $\epsilon_{i,c}$, because the researcher

^{G.12}This is equivalent to the definition of *strict truth-telling* in Fack, Grenet, and He (2019) when there are no unacceptable colleges.

does not observe its realization. In contrast, we do observe the realization of $\mathbf{X}_i$, submitted ROLs, and outcomes.

G.3.1 Weak Truth-Telling (WTT)

Naturally, one may start by a truth-telling assumption such as TT in which every applicant truthfully ranks every college in her ROL. The fact that applicants rarely rank all available colleges motivates a weaker version of truth-telling. Weak truth-telling, or WTT, can be considered as a truncated version of TT, entails two assumptions: (a) the observed number of choices ranked in any ROL is exogenous to applicant preferences and (b) every applicant ranks her top preferred colleges according to her preferences, although she may not rank all colleges. Although WTT is weaker than TT, it is still susceptible to untruthful ROLs from our robustness perspective: the robust equilibrium constructed in Theorem 1 fails WTT.

The submitted ROLs specify a rank-ordered logit model that can be estimated by Maximum Likelihood Estimation (MLE). We define this as the WTT-based estimator.

The probability of applicant i submitting $R = r^1 - r^2 - \dots - r^{|R|} \in \mathcal{R}$ is:

$$\begin{aligned} & \Pr(\sigma_i(\mathbf{u}_i, \mathbf{s}_i) = R \mid \mathbf{X}_i; \beta) \\ &= \Pr\left(u_{i,r^1} > \dots > u_{i,r^{|R|}} > u_{i,c}, \forall c \notin \{r^1, \dots, r^{|R|}\} \mid \mathbf{X}_i; \beta; |\sigma_i(\mathbf{u}_i, \mathbf{s}_i)| = |R|\right) \\ & \times \Pr(|\sigma_i(\mathbf{u}_i, \mathbf{s}_i)| = |R| \mid \mathbf{X}_i; \beta). \end{aligned}$$

Under the assumptions that $|\sigma_i(\mathbf{u}_i, \mathbf{s}_i)|$ is orthogonal to $u_{i,c}$ for all c and that $\epsilon_{i,c}$ is a type-I extreme value, we can focus on the choice probability conditional on $|\sigma_i(\mathbf{u}_i, \mathbf{s}_i)|$ and obtain:

$$\begin{aligned} & \Pr(\sigma_i(\mathbf{u}_i, \mathbf{s}_i) = R \mid \mathbf{X}_i; \beta; |\sigma_i(\mathbf{u}_i, \mathbf{s}_i)| = |R|) \\ &= \Pr\left(u_{i,r^1} > \dots > u_{i,r^{|R|}} > u_{i,c}, \forall c \notin \{r^1, \dots, r^{|R|}\} \mid \mathbf{X}_i; \beta; |\sigma_i(\mathbf{u}_i, \mathbf{s}_i)| = |R|\right) \\ &= \prod_{c \in \{r^1, \dots, r^{|R|}\}} \left(\frac{\exp(V_{i,c})}{\sum_{c' \not\prec_R c} \exp(V_{i,c'})} \right) \end{aligned}$$

where $c' \not\prec_R c$ indicates that c' is not ranked before c in R , which includes c itself and the colleges not ranked in R .

With a location normalization (e.g., $V_{i,1} = 0$), the model can be estimated by MLE with the following log-likelihood function:

$$\ln L_{WTT}(\beta \mid \mathbf{X}, \{|\sigma_i(\mathbf{u}_i, \mathbf{s}_i)|\}_i)$$

$$= \sum_{i=1}^k \sum_{c \text{ ranked in } \sigma_i(\mathbf{u}_i, \mathbf{s}_i)} V_{i,c} - \sum_{i=1}^k \sum_{c \text{ ranked in } \sigma_i(\mathbf{u}_i, \mathbf{s}_i)} \ln \left(\sum_{c' \notin \sigma_i(\mathbf{u}_i, \mathbf{s}_i)} \exp(V_{i,c'}) \right).$$

The WTT-based estimator, $\hat{\beta}^{WTT}$, is the solution to $\max_{\beta} \ln L_{WTT}(\beta \mid \mathbf{X}, \{|\sigma_i(\mathbf{u}_i, \mathbf{s}_i)|\}_i)$.

G.3.2 Stability

The assumption of stability implies that every applicant is assigned her favorite feasible college given the ex-post cutoffs. The random utility model can be estimated by MLE based on a conditional logit model where each applicant's choice set is restricted to the ex-post feasible colleges and where the matched college is the favorite among all her feasible colleges. If applicants play a regular robust equilibrium, stability is satisfied asymptotically according to Theorem 2. We define this estimator as the stability-based estimator.

Suppose that the matching is μ , which leads to a vector of cutoffs $\mathbf{P}$. With information on how colleges rank applicants, we can find a set of colleges that are ex-post feasible to i , $\mathcal{C}(s_i, \mathbf{P})$.

The conditions specified by the stability of μ imply the likelihood of applicant i matching with c^* in $\mathcal{C}(s_i, \mathbf{P})$:

$$\Pr \left(c^* = \mu(i) = \arg \max_{c \in \mathcal{C}(s_i, \mathbf{P})} u_{i,c} \mid \mathbf{X}_i, \mathcal{C}(s_i, \mathbf{P}); \beta \right).$$

Given the parametric assumptions on utility functions, the corresponding (conditional) log-likelihood function is:

$$\ln L_{ST}(\beta \mid \mathbf{X}, \mathcal{C}(s_i, \mathbf{P})) = \sum_{i=1}^k V_{i, \mu(i)} - \sum_{i=1}^k \ln \left(\sum_{c' \in \mathcal{C}(s_i, \mathbf{P})} \exp(V_{i,c'}) \right).$$

The stability-based estimator, $\hat{\beta}^{ST}$, is the solution to $\max_{\beta} \ln L_{ST}(\beta \mid \mathbf{X}, \mathcal{C}(s_i, \mathbf{P}))$.

A key assumption of this approach is that the feasible set $\mathcal{C}(s_i, \mathbf{P})$ is exogenous to i . It is satisfied when the mechanism is the serial dictatorship.

G.3.3 Estimation Results

Table G.2 provides summary statistics on the estimates from the WTT and stability approaches.

G.4 Counterfactual Analysis

We now provide some details on the counterfactual analysis. Recall that we consider the following counterfactual policy: applicants with $T_i = 1$ are given priority over those with

Table G.2: Estimation Using Different Approaches: Monte Carlo Results

DGPs	Identifying Assumption	Quality ($\beta_1 = 0.3$)		Distance ($\beta_2 = -1$)		Interaction ($\beta_3 = 2$)		Small college ($\beta_4 = 0$)	
		mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
<i>A. Both approaches are consistent.</i>									
TT	WTT	0.30	0.00	2.00	0.03	-1.00	0.03	0.00	0.02
	Stability	0.30	0.01	2.01	0.12	-1.00	0.09	0.00	0.07
<i>B. Only stability is consistent.</i>									
PIM	WTT	0.18	0.00	1.22	0.04	-0.64	0.04	-0.07	0.02
	Stability	0.30	0.01	2.01	0.12	-1.00	0.09	0.00	0.07
PRM	WTT	0.17	0.00	1.12	0.04	-0.60	0.04	-0.06	0.02
	Stability	0.29	0.02	1.92	0.21	-0.97	0.09	-0.02	0.10

Notes: This table presents estimates (mean and standard deviation across 150 samples) of the random utility model described in equation (G.14). The true values are $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.3, -1, 2, 0)$. It shows results in the three data generating processes (DGPs) with two identifying assumptions, WTT and stability.

$T_i = 0$, while within each type, applicants are still ranked according to their indices. That is, given $T_i = T_{i'}$, i is ranked higher than i' by all colleges if and only if $i > i'$. The matching mechanism is still the serial dictatorship in which everyone can rank all colleges.

The effects of the counterfactual policy are evaluated by the following four approaches.

- (i) **Actual behavior (the truth):** We use the true coefficients in utility functions to simulate counterfactual outcomes. They will be used as a benchmark against which alternative approaches will be evaluated. In keeping with our DGPs above, the “actual behavior” ranges from TT to untruthful reporting (see Section G.2). Specifically, DGP TT requires everyone to submit a truthful 12-college ROL; in DGP PIM, some applicants omit their never-matched colleges; and in DGP PRM, some applicants omit some colleges with which they have a low chance of being matched.
- (ii) **Submitted ROLs:** One assumes that the submitted ROLs under the existing policy are true ordinal preferences and that applicants will submit the same ROLs even when the existing policy is replaced by the counterfactual.
- (iii) **WTT:** One assumes that the submitted ROLs represent top preferred colleges in true preference order, and therefore applicant preferences can be estimated from the data with WTT as the identifying assumption. Under the counterfactual policy, we simulate applicant preferences based on the estimates and let applicants submit truthful 12-college ROLs.
- (iv) **Stability:** We estimate applicant preferences from the data with stability as the identifying assumption. Under the counterfactual policy, we simulate applicant preferences based on the estimates and let applicants submit truthful 12-college ROLs.

Note that we assume truthful reporting in the counterfactual in the last two approaches. This is necessary because none of these approaches estimates how applicants choose ROLs,

while we have to specify applicant behavior in counterfactual analysis. This assumption of truthful reporting in the counterfactual analysis is justified by Corollary 2.^{G.13}

When simulating counterfactual outcomes, we use the same 150 simulated samples for estimation. In particular, we use the same simulated $\{\epsilon_{i,c}\}_c$ when constructing preference profiles after preference estimation. By holding constant $\{\epsilon_{i,c}\}_c$, we isolate the effects of different estimators.

To summarize, for each of the 150 simulation samples, we conduct 12 different counterfactual analyses: 3 (DGPs: TT, PIM, and PRM) $\times$ 4 (actual behavior and 3 counterfactual approaches—submitted ROLs, WTT, and stability).

G.4.1 Performance of the Approaches in Counterfactual Analysis

Taking the counterfactual outcomes based on the actual behavior as our benchmark, we evaluate the three approaches from two perspectives: predicting the policy’s effects on outcomes and on welfare.

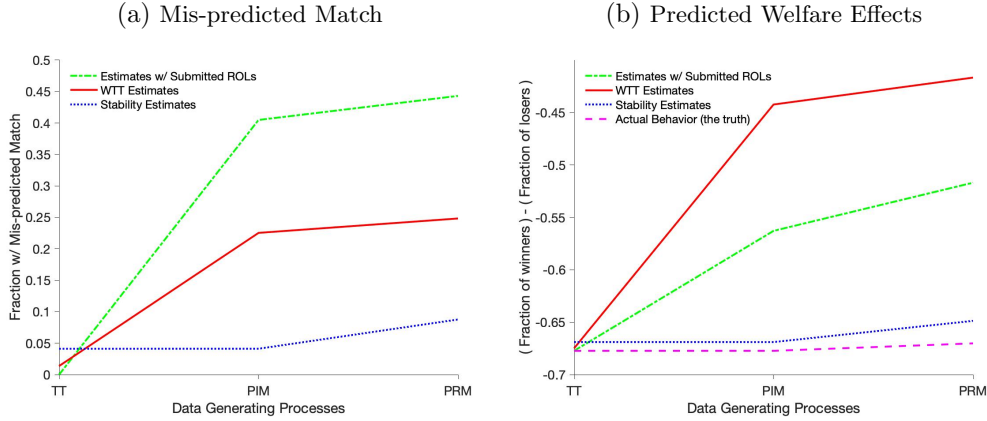


Figure G.6: Three Approaches to Counterfactual Analysis: Applicants $T_i = 0$

Notes: This figure shows the averages among $T_i = 0$ applicants across the 150 samples in each DGP. On average, there are 1201 such applicants in a sample. Given a DGP, we simulate an outcome under the counterfactual policy and compare it to the truth from the actual behavior (i.e., the true preferences with possible mistakes). Panel (a) shows the average mis-prediction rates. Panel (b) shows the predicted welfare effects by each approach. It is measured by the difference between the fractions of winners and losers. See Table G.3 for more details.

Complementing Figure F.4 for applicants with $T_i = 1$ in the main text, Figure G.6 shows the mis-predicted match and predicted welfare effects for applicants with $T_i = 0$. The general patterns are the same as in Figure F.4: WTT and submitted ROLs produce

^{G.13}Corollary 2 rests on the uniqueness of stable matching in $E = [\eta, S]$, guaranteed by the full support assumption on η . While the current priority structure violates full support, serial dictatorship produces a unique stable outcome, and thus validates the corollary for the current environment.

biased predictions whenever some applicants are not truthful, while stability performs well in all DGPs.

Table G.3: Welfare Effects of the Counterfactual Policy (percentage points)

Approaches to counterfactual		Worse off		Better off		Indifferent	
		mean	s.d.	mean	s.d.	mean	s.d.
<i>Panel A: Applicants with $T_i = 1$</i>							
DGP: TT	Submitted ROLs	0	0	91	1	9	1
	WTT	0	0	91	1	9	1
	Stability	0	0	91	1	9	1
	Actual Behavior (the Truth)	0	0	91	1	9	1
DGP: PIM	Submitted ROLs	0	0	78	2	22	2
	WTT	0	0	88	1	12	1
	Stability	0	0	91	1	9	1
	Actual Behavior (the Truth)	0	0	91	1	9	1
DGP: PRM	Submitted ROLs	0	0	72	2	28	2
	WTT	0	0	87	1	13	1
	Stability	0	0	91	1	9	1
	Actual Behavior (the Truth)	0	0	91	1	9	1
<i>Panel B: Applicants with $T_i = 0$</i>							
DGP: TT	Submitted ROLs	68	2	0	0	32	2
	WTT	68	2	0	0	32	2
	Stability	67	2	1	0	32	2
	Actual Behavior (the Truth)	68	2	0	0	32	2
DGP: PIM	Submitted ROLs	56	2	0	0	44	2
	WTT	53	2	9	1	37	2
	Stability	67	2	1	0	32	2
	Actual Behavior (the Truth)	68	2	0	0	32	2
DGP: PRM	Submitted ROLs	52	2	1	0	47	2
	WTT	52	2	10	1	38	2
	Stability	66	3	1	1	32	2
	Actual Behavior (the Truth)	67	2	0	0	32	2

Notes: This table presents the estimated effects of the counterfactual policy (giving $T_i = 1$ applicants priority in admission) on applicants with $T_i = 1$ (Panel A) and those with $T_i = 0$ (Panel B). On average, there are 599 applicants with $T_i = 1$ (standard deviation 14) and 1201 applicants with $T_i = 0$ (standard deviation 14) in each simulation sample. The table shows results in the three data generating processes (DGPs) with four approaches. The one using submitted ROLs assumes that submitted ROLs represent applicant true ordinal preferences; WTT assumes that every applicant truthfully ranks her top K_i ($1 < K_i \leq 12$) preferred colleges (K_i is observed); and stability implies that every applicant is matched with her favorite feasible college, given the ex-post cutoffs. The truth is simulated with the possible mistakes in each DGP. The welfare change of each applicant is calculated in the following way: we first simulate the counterfactual match and investigate if a given applicant is better off, worse off, or indifferent by comparing the two matches according to estimated/assumed/true ordinal preferences. In each simulation sample, we calculate the percentage of different welfare change; the table then reports the mean and standard deviation of the percentages across the 150 simulation samples.

Moreover, Table G.3 (Panel A for applicants with $T_i = 1$ and Panel B for those with $T_i = 0$) presents detailed statistics on the fractions of applicants being worse off, better off,

and indifferent based on different approaches.

References

- ABDULKADIROĞLU, A., P. A. PATHAK, AND A. E. ROTH (2009): “Strategy-proofness versus Efficiency in Matching with Indifferences: Redesigning the NYC High School Match,” *American Economic Review*, 99(5), 1954–1978.
- ABDULKADIROĞLU, A., AND T. SONMEZ (2003): “School Choice: A Mechanism Design Approach,” *American Economic Review*, 93, 729–747.
- ABDULKADIROĞLU, A., N. AGARWAL, AND P. A. PATHAK (2017): “The Welfare Effects of Coordinated Assignment: Evidence from the New York City High School Match,” *American Economic Review*, 107(12), 3635–89.
- AGARWAL, N. (2015): “An empirical model of the medical match,” *American Economic Review*, 105(7), 1939–1978.
- AKYOL, Ş. P., AND K. KRISHNA (2017): “Preferences, Selection, and Value Added: A Structural Approach,” *European Economic Review*, 91, 89–117.
- ARTEMOV, G., Y.-K. CHE, AND Y. HE (2020): “Strategic Mistakes: Implications for Market Design Research,” mimeo.
- AZEVEDO, E. M., AND E. BUDISH (2018): “Strategy-proofness in the Large,” *The Review of Economic Studies*, 86(1), 81–116.
- AZEVEDO, E. M., AND J. D. LESHNO (2016): “A supply and demand framework for two-sided matching markets,” *Journal of Political Economy*, 124, 1235–1268.
- BUCAREY, A. (2018): “Who Pays for Free College? Crowding Out on Campus,” mimeo.
- CHE, Y.-K., D. HAHM, AND Y. HE (2022): “Leveraging Uncertainties to Infer Preferences: Robust Analysis of School Choice,” *Unpublished manuscript*.
- CHE, Y.-K., AND F. KOJIMA (2010): “Asymptotic Equivalence of Probabilistic Serial and Random Priority Mechanisms,” *Econometrica*, 78(5), 1625–1672.
- CHE, Y.-K., AND O. TERCIEUX (2019): “Efficiency and Stability in Large Matching Markets,” *Journal of Political Economy*, 127(5).
- CHEN, L., AND J. S. PEREYRA (2019): “Self-selection in school choice,” *Games and Economic Behavior*, 117(C), 59–81.
- CHEN, Y., AND T. SÖNMEZ (2002): “Improving Efficiency of On-campus Housing: An Experimental Study,” *American Economic Review*, 92, 1669–1686.

- CHIAPPORI, P.-A., AND B. SALANIÉ (2016): “The Econometrics of Matching Models,” *Journal of Economic Literature*, 54(3), 832–861.
- COMBE, J., O. TERCIEUX, AND C. TERRIER (2016): “The Design of Teacher Assignment: Theory and Evidence,” in *ACM Conference on Economics and Computation*.
- DEB, J., AND E. KALAI (2015): “Stability in Large Bayesian Games with Heterogeneous Players,” *Journal of Economic Theory*, 157, 1041–1055.
- DREYFUSS, B., O. HEFFETZ, AND M. RABIN (2019): “Expectations-Based Loss Aversion May Help Explain Seemingly Dominated Choices in Strategy-Proof Mechanisms,” SSRN Working Paper 3381244.
- DUBINS, L. E., AND D. A. FREEDMAN (1981): “Machiavelli and the Gale-Shapley algorithm,” *American Mathematical Monthly*, 88, 485–494.
- FACK, G., J. GRENET, AND Y. HE (2019): “Beyond Truth-Telling: Preference Estimation with Centralized School Choice and College Admissions,” *American Economics Review*, 109(4), 1486–1529.
- FOX, J. T. (2009): “Structural Empirical Work Using Matching Models,” *New Palgrave Dictionary of Economics. Online edition*.
- FOX, J. T., AND P. BAJARI (2013): “Measuring the Efficiency of an FCC Spectrum Auction,” *American Economic Journal: Microeconomics*, 5(1), 100–146.
- FUDENBERG, D., AND J. TIROLE (1991): *Game Theory*. MIT Press, Cambridge, Massachusetts.
- GALE, D., AND L. S. SHAPLEY (1962): “College Admissions and the Stability of Marriage,” *American Mathematical Monthly*, 69, 9–15.
- HAERINGER, G., AND F. KLIJN (2009): “Constrained School Choice,” *Journal of Economic Theory*, 144, 1921–1947.
- HÄLLSTEN, M. (2010): “The Structure of Educational Decision Making and Consequences for Inequality: A Swedish Test Case,” *American Journal of Sociology*, 116(3), 806–54.
- HASSIDIM, A., A. ROMM, AND R. I. SHORRER (2020): “The limits of incentives in economic matching procedures,” *Management Science*.
- JUDD, K. L. (1985): “The law of large numbers with a continuum of IID random variables,” *Journal of Economic Theory*, 35(1), 19 – 25.
- KALAI, E. (2004): “Large Robust Games,” *Econometrica*, 72, 1631–1665.

- KIRKEBØEN, L. J. (2012): “Preferences for Lifetime Earnings, Earnings Risk and Monpecuniary Attributes in Choice of Higher Education,” *Statistics Norway Discussion Papers* No. 725.
- LI, S. (2017): “Obviously strategy-proof mechanisms,” *American Economic Review*, 107(11), 3257–87.
- LIU, Q., AND M. PYCIA (2016): “Ordinal Efficiency, Fairness, and Incentives in Large Markets,” SSRN Working Paper 1872713.
- MATEJKA, F., AND A. MCKAY (2015): “Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model,” *American Economic Review*, 105(1), 272–298.
- MCDIARMID, C. (1989): “On the method of bounded differences,” *Surveys in Combinatorics*, pp. 148–188.
- MCKELVEY, R., AND T. PALFREY (1995): “Quantal Response Equilibria for Normal Form Games,” *Games and Economic Behavior*, 10, 6–38.
- PYCIA, M. (2019): “Evaluating with Statistics: Which Outcome Measures Differentiate Among Matching Mechanisms?,” University of Zurich.
- REES-JONES, A. (2017): “Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match,” *Games and Economic Behavior*, 108, 317–330.
- ROTH, A. E. (1982): “The Economics of Matching: Stability and Incentives,” *Mathematics of Operations Research*, 7, 617–628.
- ROTH, A. E. (1991): “A natural experiment in the organization of entry-level labor markets: regional markets for new physicians and surgeons in the United Kingdom,” *The American economic review*, pp. 415–440.
- ROTH, A. E., AND E. PERANSON (1999): “The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design,” *American Economic Review*, 89(4), 748–780.
- SHORRER, R. I., AND S. SÓVÁGÓ (2018): “Obvious Mistakes in a Strategically Simple College-Admissions Environment,” SSRN Working Paper 2993538.
- SIMS, C. A. (2003): “Implications of Rational Inattention,” *Journal of Monetary Economics*, 50(3), 665–690.
- SOTOMAYOR, M. (2008): “The stability of the equilibrium outcomes in the admission games induced by stable matching rules,” *International Journal of Game Theory*, 36(3), 621–640.

VESKI, A., P. BIRÓ, K. PÖDER, AND T. LAURI (2017): “Efficiency and Fair Access in Kindergarten Allocation Policy Design,” *The Journal of Mechanism and Institution Design*, 2(1), 57–104.