Strategic ‘Mistakes’: Implications for Market Design Research*

Georgy Artemov† Yeon-Koo Che‡ Yinghua He§

December 19, 2017

Abstract

Using a rich data set on Australian college admissions, we show that even in strategically straightforward situations, a non-negligible fraction of applicants adopt strategies that are unambiguously dominated; however, the majority of these ‘mistakes’ are payoff irrelevant. We then propose a new equilibrium solution concept that allows for mistakes. Applying it to a strategy-proof mechanism in which colleges strictly rank applicants, we show that equilibrium strategies need not be truth-telling, but that every equilibrium outcome is asymptotically stable. Our Monte Carlo simulations illustrate the differences between the empirical methods based on truth-telling or outcome stability, revealing that the latter is more robust to potential mistakes. Taken together, our results suggest that strategy-proof mechanisms perform reasonably well in real life, although applicants’ mistakes should be carefully taken into account in empirical analysis.

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

Keywords: Strategic mistakes, payoff relevance of mistakes, robust equilibria, truthful-reporting strategy, stable-response strategy, stable matching, preference estimation, demand estimation, counterfactual analysis.

1 Introduction

Strategy-proofness—or making it a dominant strategy to reveal one’s own preferences truthfully—is an important desideratum in market design. It makes straightforward for a participant to act in

*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 the Asia Pacific Industrial Organization Conference, Columbia, “Econometrics Meets Theory” Conference at NYU, NBER Market Design Group Meeting, Workshop “Matching in Practice” (Brussels, 2017), and Paris School of Economics for their comments.

†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; Toulouse School of Economics, France. Email: yinghua.he@gmail.com.

one's best interest, thus minimizing the scope for making mistakes. It also equalizes the playing field, as even an unsophisticated participant is protected from strategic behavior of others. Further, strategy-proofness aids empirical research by making participants' choices easy to interpret.

However, this view has been challenged by a growing number of authors who find that strategic mistakes are not uncommon, even in strategy-proof environments. Laboratory experiments have shown that a significant fraction of subjects do not report their preferences truthfully in strategy-proof mechanisms such as the applicant-proposing deferred acceptance (DA) and the top-trading cycles mechanisms (see, e.g., Chen and Sönmez, 2002). More alarmingly, similar problems occur in high-stakes real-world contexts. In a study of admissions to Israeli graduate programs in psychology (which use DA), Hassidim, Romm, and Shorrer (2016) find that approximately 19% of applicants either did not list a scholarship position for a program or ranked a non-scholarship position higher than the corresponding scholarship position. Since a scholarship position is unambiguously preferred to a non-scholarship position of the same program, such behavior constitutes a dominated strategy. In a similar vein, Shorrer and Sóvágó (2017) find that a large fraction of the applicants employ a dominated strategy in the Hungarian college-admissions process, which uses a strategically simple mechanism. Also using data generated by DA, Rees-Jones (2017) reports that 17% of the 579 surveyed US medical seniors indicate misrepresenting their preferences in the National Resident Matching Program, and Chen and Pereyra (2015) document similar evidence for high school choice in Mexico City.

These findings raise questions on the mechanisms widely used in practice and their empirical assessment. At the same time, the mere presence of "mistakes" is not enough to draw conclusions on the matter. If mistakes occur only when they make little difference to the outcome, then the full rationality hypothesis can be a reasonable assumption for analyzing a mechanism. One would thus require a deeper understanding of the nature of mistakes and which circumstances lead to those mistakes.

To this end, we first study a data set from the Victorian Tertiary Admissions Centre (VTAC), a central clearinghouse that organizes tertiary education admissions in Victoria, Australia. VTAC uses a mechanism that resembles a serial dictatorship with the serial order given by a nationwide test score, the Equivalent National Tertiary Entrance Rank (ENTER). Every applicant is to be matched with one of the tertiary "courses". A course roughly corresponds to a major at a given college in other countries. One important feature of the mechanism is that applicants can submit a rank-order list (ROL) ranking up to 12 courses, which makes the mechanism non-strategy-proof (Haeringer and Klijn, 2009); however, we can still identify certain unambiguously dominated strategies. Similar to Hassidim, Romm, and Shorrer (2016), our study exploits one specific feature of the system. An applicant can apply for a given college-major pair as either (i) a "full-fee" course (or FF course) which charges full tuition; or (ii) a Commonwealth-supported course (to which we will refer as "reduced-fee" or RF course), which subsidizes approximately 60% of tuition; or both, whenever available. For any given college-major pair, the RF course clearly dominates the FF course; hence ranking FF but not RF course of the same college-major pair in an

ROL that does not fill up the 12 slots—henceforth called a **skip**—is an unambiguously dominated strategy.[1]

In the sample year of 2007, we find that 1,009 applicants skipped. This number represents 3.6% of the 27,922 applicants who ranked fewer than 12 courses or 35% of the total 2,915 applicants who listed at least one FF course that is also offered as an RF course. These figures are consistent with those documented in Hassidim, Romm, and Shorrer (2016) and can be viewed as non-negligible.

However, the vast majority of these mistakes were not payoff relevant. For an applicant who skipped at least one RF course, the skip is payoff relevant if, holding the ROLs submitted by other applicants constant, adding the omitted RF course into the applicant's ROL would lead to a different assignment outcome for her. This procedure identifies payoff relevant mistakes by 14–201 applicants out of the $1,009$ who skipped at least one RF course, with the exact number depending on how these applicants would have ranked the skipped courses in their ROLs (e.g., at the top of the ROLs or just ahead of the FF courses). These applicants represent between 1.39 percent and 19.92 percent of all applicants who skip. When one uses all applicants as base, the payoff-relevant mistakes comprise only 0.05–0.72 percent. It should be emphasized that these statistics are calculated for the mistakes that we identify by comparing listed FF and omitted RF courses. It is possible that applicants make other mistakes that are not identified without further assumptions or estimations.

Our rich micro data set is well suited to investigating who made mistakes, whether the mistakes are payoff relevant and what circumstances led to them. We find that applicants' academic ability (measured independently of ENTER) is negatively correlated with skips, suggesting that misunderstanding the mechanism may play a role in making mistakes. However, even controlling for academic ability, ENTER is also negatively correlated with mistakes. This finding suggests that applicants omit courses to which they are unlikely to be admitted; specifically, a lower ENTER applicant finds more courses to be out-of-reach and skip them. Furthermore, we find no evidence that omitting courses is a conscious attempt to game the system to receive a better match.

In contrast, the individual characteristics correlated with *payoff-relevant* mistakes are very different. They are no longer correlated with academic ability, suggesting that lower-ability applicants do not make more payoff-relevant mistakes. Moreover, the probability of making payoff-relevant mistakes is positively, rather than negatively, correlated with ENTER, which may reflect the fact that, when one's score rises, more courses become feasible, and a mistake is more likely to be declared payoff-relevant.

A unique feature of the Victorian system allows us to observe further differences between skips and payoff-relevant mistakes. The applicants are permitted to modify their submitted ROLs over time. We find that while the number of applicants who skip *increases* over time, the number of applicants who make payoff-relevant mistakes *decreases*. We also exploit the fact that we

[1] As will be discussed in detail, ranking an FF course ahead of the RF version of the same college-major pair need not be a dominated strategy in our empirical setting. For an applicant who fills up 12 slots, we cannot identify a skip as a dominated strategy without information about the applicant's preferences because any ROL that respects the true preference order among the courses included in the ROL is not dominated (Haeringer and Klijn, 2009).

observe ROLs submitted before and after the applicants receive their ENTER. We study applicants' response to a "shock": the difference between the realized ENTER and the ENTER forecasted based on one's academic ability measure. A larger positive shock, despite making an applicant eligible for a larger set of courses, leads to a *reduction* in payoff-relevant mistakes. In contrast, a shock has no effect on skips.

To the extent that mistakes do occur, it is important to understand the implications of mistakes for market design research. Of particular interest is how mistakes—some of them payoff relevant—affect our ability to empirically recover the underlying preferences of participants and to perform counterfactual analyses of new hypothetical market designs. To study these questions, we first develop a theoretical model of applicants' behavior in a large matching market operated by a DA mechanism (of which serial dictatorship is a special case). Colleges rank applicants by some score, and every applicant knows her own score before submitting applications. In keeping with the empirical findings, we focus on an equilibrium concept—called *robust equilibrium*—which allows applicants to make mistakes as long as they become virtually payoff irrelevant as the market size grows arbitrarily large.

We show that submitting ROLs that differ from true preferences, conditional on applying at all, is robust equilibrium behavior for all except a vanishing fraction of applicants. In such an equilibrium, an applicant skips the colleges which she is unlikely to be assigned, either because they are "out of reach" (i.e., their cutoff scores are higher than hers) or because she feels "she can do better" (i.e., she has an admittable score for another college she prefers to them). Such behavior is supported as robust equilibrium behavior since, as the market grows large, the sub-optimality of playing such a strategy disappears for all but a vanishing fraction of applicants. If applicants behave according to our robustness concept, this result implies that the observed ROLs need not reflect the applicants' true preference orders. This finding calls into question the empirical methods for preference estimation and counterfactual analysis based on the assumption that applicants submit true preferences as their ROLs in a DA mechanism.

We next show that in *any* robust equilibrium, as the market grows large, almost all applicants are assigned to the most preferred feasible college. A college is feasible to an applicant if she can be accepted by that college by submitting some ROL, e.g., top ranking that college, while holding others' submitted ROLs constant. That is, any robust equilibrium leads to an asymptotically stable matching outcome. This result implies that stability is a valid identification restriction in a sufficiently large market. While truthful reporting implies a stable matching, a stable matching need not involve truthful ROLs. Hence, stability is a weaker restriction. The two theoretical results provide the sense in which the empirical methods based on truthful reporting are vulnerable to the types of mistakes documented in the first part of the paper and, at the same time, the sense in which the empirical methods based on a weaker stability notion are relatively robust to them. We further show that, while applicants' equilibrium *behavior* generally differs from truth-telling, the resulting *outcome* in a large market is close to one that would emerge if all applicants reported their preferences truthfully, which justifies the vast theoretical literature making the truth-telling assumption in strategy-proof environments.

To gain quantitative insights, we perform a Monte Carlo simulation of college admissions in which applicant's preferences follow a multinomial logit model. We assume a serial dictatorship mechanism with a pre-specified serial order. Even though this mechanism is strategy-proof, in keeping with our empirical findings, we entertain alternative scenarios that vary in both the extent to and frequency with which applicants make mistakes. Specifically, the assumed behavior ranges from truthful reporting (i.e., no mistakes) to behavior exhibiting varying degrees of payoff-irrelevant skips to behavior exhibiting varying degrees of payoff-relevant mistakes. Under these alternative scenarios, we structurally estimate applicant preferences using truthful reporting and stability as two alternative identifying assumptions. In addition, to account for a certain degree of payoff-relevant mistakes, we further propose a robust approach based on stability.

The estimation results highlight the bias-variance trade-off: the estimator based on truthful reporting uses more information on revealed preferences of applicants and has a much lower variance than the alternatives. However, the truthful-reporting assumption introduces downward biases in the estimators of college quality, especially among popular or small colleges. This is because this assumption incorrectly claims every omitted college to be inferior to the ones ranked, while popular and small colleges are often out-of-reach to many applicants and thus very likely to be skipped. In contrast, the estimator based on stability is immune to all payoff-irrelevant skips, because it places no preference restrictions on skipped, out-of-reach colleges. Even when there are some payoff-relevant mistakes, the biases in the estimator based on stability are small, and those from the robust approach are even smaller.

When biased estimates are used in counterfactual analysis, the effect of the policy being analyzed may be mis-predicted. We use a hypothetical affirmative action policy that prioritizes disadvantaged applicants to quantify this effect. When simulating the true counterfactual outcomes as the benchmark, we allow applicants to skip or to make payoff relevant mistakes. In terms of predicting the counterfactual matching outcome, estimates from truthful reporting perform worse than those from stability when applicants make mistakes. When we evaluate the welfare effects of the policy, the truthful-reporting assumption under-estimates the benefits to disadvantaged applicants as well as the harm to others; stability, however, predicts effects close to the true values. The robust approach further improves upon the stability assumption. In addition, we evaluate another common approach to counterfactual analysis in the market design research: holding submitted ROLs constant across two policies. We show that this approach, by failing to account for the likely change in ROLs submitted by applicants, produces an even larger bias than the truthful-reporting assumption.

Other Related Literature. Participants' mistakes in strategy-proof mechanisms are the focus of a recent and fast-growing literature. Li (2017) proposes an explanation different from ours: participants do not comprehend every detail of the mechanism and may not play a dominant strategy when its dominance is not "obvious."[2] Arguably, this is less of a concern when participants can be assured that they would not gain from misreporting their preferences, as is often done in many real-world strategy-proof mechanisms. In addition, mechanisms that are commonly used in practice are not obviously strategy-proof and cannot be transformed into ones satisfying obvious strategy-proofness (Ashlagi and Gonczarowski, 2016; Pycia and Troyan, 2016).

Azevedo and Budish (2015) propose the notion of "strategy-proofness in the large". Similarly to our notion of robust equilibrium, participants are allowed to make mistakes of ϵ size that vanish as a market grows large. However, the focus of the two papers is entirely different. While they ask what conditions on mechanisms would guarantee approximate strategy-proofness, we ask what approximate solution concept can explain the mistakes, or the non-truth-telling behaviors, that we observe in practice.

In our model, we assume that colleges rank applicants strictly based on scores that are known by applicants before applying. This is the feature of Victorian college admissions mechanism, which is shared by numerous real-life centralized matching markets, including college admissions in Chile, Hungary, Ireland, Norway, Spain, Sweden, Taiwan, Tunisia, and Turkey, as well as school choice in Finland, Ghana, Romania, Singapore, and Turkey (see Table 1 in Fack, Grenet, and He (2017) as well as the references below).

In such matching markets, the truthful-reporting assumption has been utilized to estimate applicants' preferences. With data from Ontario, Canada, which employs a decentralized system with applications being relayed by a platform, Drewes and Michael (2006) assume that applicants rank programs truthfully when submitting the non-binding ROLs. With college admissions data from Sweden's centralized system, Hällsten (2010) adopts a rank-ordered logit model for preference estimation under a version of the truthful-reporting assumption. Similarly, with data from the centralized college admissions in Norway, Kirkebøen (2012) also imposes a version of the truthful-reporting assumption but sometimes excludes from an applicant's choice set every college program for which the applicant does not meet the formal requirements or is below its previous-year cutoff.

Holding the submitted ROLs constant across two policies is also a common approach to counterfactual analysis in market design research, especially when the existing mechanism is strategy-proof and applicants are assumed to submit a truthful, *complete* ROLs. For instance, Roth and Peranson (1999) use data from the National Resident Matching Program and simulate matching outcomes under alternative market designs. Combe, Tercieux, and Terrier (2016) and Veski, Biró, Poder, and Lauri (2016) adopt the same approach in their counterfactual analysis for centralized teacher assignment in France and kindergarten allocation in Estonia, respectively. Our paper suggests that such an approach may entail bias, the magnitude of which will depend on how much applicants' behavior will change under the counterfactual policy.[3]

[2] A strategy is "obviously dominant" if its worst payoff is no worse than the best payoff of a deviant strategy, where the best and the worst are defined over a set of strategies that a participant can distinguish.

[3] In our simulations (Section 4), we evaluate an affirmative action policy that prioritizes some applicants. In contrast, Roth and Peranson (1999) and Combe, Tercieux, and Terrier (2016) evaluate alternative mechanisms without changing priority structure. One may argue that the resulting bias may be smaller in these two papers

There is also an important and well-researched setting where "colleges" do not rank applicants strictly before the application and break ties with a lottery. An example is the school choice program in New York City (Abdulkadiroglu, Pathak, and Roth, 2009; Abdulkadiroglu, Agarwal, and Pathak, Forthcoming; Che and Tercieux, 2015a). This setting is conceptually different and is not addressed in the present paper. Indeed, a key ingredient of our proofs is that, for a given applicant, the probability of admission to some colleges converges to zero as economy grows. When applicants are ranked by colleges according to a post-application lottery, the probability of being admitted to a given college would be bounded away from zero; hence our results would not hold.

Strategic mistakes have also been empirically studied in the literature on non-strategy-proof mechanisms, especially the Boston immediate-acceptance mechanism, a common mechanism used in school choice. The specific school systems considered in the literature include those in Barcelona (Calsamiglia, Fu, and Güell, 2014), Boston, MA (Abdulkadiroglu, Pathak, Roth, and Sonmez, 2006), Beijing (He, 2017), Cambridge, MA (Agarwal and Somaini, Forthcoming), New Haven, CT (Kapor, Neilson, and Zimmerman, 2016), Seoul (Hwang, 2017), and Wake County, NC (Dur, Hammond, and Morrill, Forthcoming). Their empirical approaches to identifying mistakes are different from ours, because without detailed information on participants' true preferences, it is almost impossible to identify those who play a dominated strategy under a non-strategy-proof mechanism.

The rest of the paper is organized as follows. In Section 2, we study the frequency and nature of strategic mistakes from the VTAC college admissions data. In Section 3, we explore the theoretical implications of the findings for market design research, both empirical and theoretical. In Section 4, we report the Monte Carlo simulations performed on the alternative methods. Section 5 concludes.

2 Strategic Mistakes in Australian College Admissions

2.1 Institutional Details and Data

We use the data for the year 2007 from the Victorian Tertiary Admission Centre (VTAC), which is a centralized clearinghouse for admissions to tertiary courses in Victoria. Applicants are required to rank tertiary courses by which they want to be considered. VTAC also collects academic and demographic information about applicants.

The unit of admission in Victoria, a *course*, is a combination of (i) a tertiary institution; (ii) a field of study that the applicant wants to pursue; and (iii) a tuition payment. A tertiary institution may be either a university (including programs not granting bachelor degrees) or a technical school. A field of study is roughly equivalent to a *major* in US universities. In 2007, the full fees are approximately AUD17,000 (USD13,000) per year. The reduced fees are set by the government and have a median about AUD7,000 (USD5,500) per year. The government offers student loans, which are subsidized for students in RF courses. The normal duration of a university course is

compared to the bias we report.

three years. Apart from tuition payments, there is no difference between FF and RF courses. Furthermore, both FF and RF courses share the same course description (see Figure A.1 for an example), so there is no information friction associated with finding the RF course corresponding to an FF course. In total, there were 1899 tertiary programs in 2007. 881 of them offered both the FF and RF options, among which 97 percent were university programs as described above. There were also approximately 800 RF-only programs and 200 FF-only programs.

Applicants are required to submit their applications in the form of a rank-order list (ROL), along with other information at the end of September. In mid-December, applicants receive their Equivalent National Tertiary Entrance Ranks (ENTER), as a number between zero and 99.95 in 0.05 increments, to which we will refer as **score** throughout the paper. For applicant i, $Score_i$ is i's rank and shows a percentage of applicants with scores below $Score_i$. For most applicants, score is the sole determinant of the admission. As ROLs are initially submitted before score is known, applicants have an opportunity to revise their ROL after the release of score. Offers are extended to the applicants in January-February. Applicants have approximately two weeks to accept by enrolling in the course they are offered.

Once applications are finalized, courses rank applicants and transmit their admission offers to VTAC. Using an applicant's ROL, VTAC picks the highest-ranked course that has admitted the applicant, *one of each type* (FF/RF), and transmits the offer(s) to the applicant. That is, if an applicant's ROL contains both RF and FF courses, this applicant may receive two offers, one of each type. This feature means that ranking an FF course ahead of the corresponding RF course is not a dominated strategy as the list of RF courses is treated as separate from the list of FF courses.

When ranking applicants, courses follow a pre-specified, published set of rules. For the largest category of applicants, admission is based almost exclusively on their scores. We focus on these applicants and refer to them as "V16" applicants, following the code assigned to them by VTAC. They are the current high school students who follow the standard Victorian curriculum. As applicants are admitted based on their scores, the admission decision of a course for V16 applicants can be expressed as a "cutoff": the lowest score sufficient for admission to the course.[4]

Applicants can rank up to 12 courses. The applicants who exhaust the length of their ROLs may be forced to omit some courses that they find desirable; hence we focus only on those who list fewer than 12 courses. These applicants comprise 75 percent of the V16 applicants.

Out of the 27,922 V16 applicants who list fewer than 12 courses (below we refer to the collection of these applicants as the "full sample"), 24,625 applicants ranked at least one program that offers both RF and FF courses. Among them, 2,915 applicants ranked at least one FF course that has a corresponding RF version. These 2,915 applicants constitute the "FF subsample".[5]

[4] We define the cutoff of a course as the median of the highest 5% scores among all rejected applicants and the lowest 5% scores among all accepted applicants. See Appendix A for details on course selection, non-V16 applicants, and the definition of a cutoff.

[5] In the following, when each of the samples is used for regression analysis, we sometimes drop some observations that have missing values in at least one of the control variables. This explains why the number of observations in

The first two columns of Table 1 show that the average length of the submitted ROLs in the full sample is 6.61, including 6.20 RF courses. This implies that applicants mainly apply to RF courses. Even in the FF subsample, the average number of FF courses in the submitted ROLs is only 2.33, while an average ROL contains 7.67 choices.

Table 1: ROLs, Skips, and Mistakes among V16 Applicants Listing Fewer than 12 Courses

	Full sample	FF subsample	Skips	Payoff-relevant mistakes	
				Upper bound	Lower bound
Total number of applicants	27,922	2,915	1,009	201	14
Percentage of the full sample	100.00	10.44	3.61	0.72	0.05
Percentage of the FF subsample		100.00	34.6	6.90	0.48
Percentage of the "Skips" subsample			100.00	19.92	1.39
Average length of submitted ROLs	6.61	7.67	7.47	7.65	6.57
Average number of RF courses in submitted ROLs	6.20	5.34	4.72	4.55	0.93
Average number of FF courses in submitted ROLs	0.41	2.33	2.75	3.10	5.64

Notes: ''Full sample" refers to all V16 applicants who list fewer than 12 courses in the 2007 college admissions. "FF subsample" are the applicants from the full sample who list at least one full-fee course which has a corresponding reduced-fee version. "Skips" refers to the applicants who list a full-fee course but do not list the corresponding reduced-fee course at least once. "Payoff-relevant mistakes" refers to the applicants who would have received a different assignment if they had not skipped a reduced-fee course.

2.2 Skips and Payoff-Relevant Mistakes

If an applicant with less than 12 courses in her ROL lists an FF course but does not list the corresponding RF course, we say that the applicant "skips". Even if an applicant skips an RF course, the skip may not affect the applicants' assignment for two reasons. First, the applicant's score may be below the course's cutoff, and the course is not feasible for the applicant. Second, the applicant may have been assigned to a more desirable course than the one skipped. When, holding other applicants' ROLs constant, correcting the skip leads to a change in the applicant's assignment, we say that the applicant makes a "payoff-relevant mistake". We will demonstrate that most of the skips applicants make are not payoff-relevant mistakes.

As we do not know where in the applicant's ROL a skipped course should be, we report the lower and upper bounds of payoff-relevant mistakes. To calculate the lower bound, we assume that the skipped RF course would have been ranked below all the RF courses in the applicant's ROL. To calculate the upper bound, we assume that the skipped RF course would have been ranked first in the applicant's ROL.[6]

In Table 1, we report the number of applicants who make at least one mistake of skipping a course and the number of applicants for whom skipping a course becomes a payoff-relevant mistake at least once, showing both the upper and lower bounds. The table suggests that among applicants who list at least one FF course, 35 percent skip. Among the applicants who skip, the skip is not a payoff-relevant mistake for at least 80 percent of them. Those with a payoff-relevant mistake are between 0.05 and 0.72 percent of all applicants in the full sample.

the following tables is usually smaller than the original size of the full sample or the FF subsample.

[6] See Appendix A for details on calculating the bounds.

2.2.1 Correlation between Applicant's Scores and Skips

Suppose that an applicant expects that course c's cutoff will be above the applicant's score; hence, the applicant does not expect to be assigned to course c even if c is listed in the applicant's ROL. We call such a course "out of reach" for the applicant.

Our leading hypothesis, denoted by H_1, is that (i) applicants may omit out-of-reach courses and (ii) there is no systematic pattern in omitting out-of-reach courses. Specifically, part (ii) means that the likelihood of omitting such a course is independent of the rank of this course in the true preferences of the applicant.

Hypothesis H_1 implies a negative correlation between the probability that an applicant skips a course and the applicant's score. Indeed, the lower the applicant's score, the larger the set of out-of-reach courses. Moreover, as an RF course is expected to have a higher cutoff than the corresponding FF course, the set of out-of-reach courses will contain some RF courses without their FF counterparts. Hence, when an applicant omits courses that are out of reach, it is more likely that this applicant will be identified as making a skip.

To investigate the relationship between score and skip, we run the following regression:

$$Skip_i \times 100 = \alpha + \beta Score_i + Controls_i + \epsilon_i, \tag{1}$$

where $Skip_i = 1$ if applicant i has made at least one skip and is zero otherwise,[7] and $Controls_i$ includes i's demographics. We expect β to be negative.

There are several alternative explanations for the negative correlation between $Score_i$ and $Skip_i$:

H_2: Applicants' scores are correlated with cognitive abilities. Those with higher cognitive abilities are able to comprehend the mechanism better, thereby reducing the probability of skipping. To test H_2, we will use another measure of applicant ability, different from $Score_i$, as a control variable.

H_3: Skipping a course is an instance of an applicant's misguided attempt to gain a better assignment from the mechanism. Specifically, applicants drop out-of-reach courses from the top of their ROL so that their within-reach courses are ranked higher. The important difference between H_3 and H_1 is where the skipped courses lie in an applicant's ROL. H_3 implies that such courses are concentrated at the top of ROL, while H_1 does not impose any such restrictions, as out-of-reach courses can be anywhere in the ROL.

We report the estimation results of equation (1) in Table 2. All odd-numbered columns show the results from the full sample, while all even-numbered columns focus on the FF subsample. The control variables include school fixed effects and application demographics, e.g., the applicants' gender, median income (in logarithm) in the postal code in which the applicant resides, citizenship status, region born, and language spoken at home. The regressions also include eleven dummy

[7] We use $Skip_i \times 100$ to make the results more readable. If we use $Skip_i$ instead, every estimate will be one percent of those reported here.

Table 2: Probability of Skipping a Reduced-fee Course

	Full sample (1)	FF subsample (2)	Full sample (3)	FF subsample (4)	Full sample (5)	FF subsample (6)
Score	-0.06***	-0.71***	-0.04***	-0.55***	-0.04***	-0.56***
	(0.01)	(0.06)	(0.01)	(0.08)	(0.01)	(0.07)
GAT			-0.05***	-0.35***	-0.04***	-0.33***
			(0.01)	(0.12)	(0.01)	(0.10)
(School fees >AUD11,000) × *Score*					-0.03***	-0.05**
					(0.01)	(0.02)
Other controls	Yes	Yes	Yes	Yes	Yes	Yes
# of Applicants	26,325	2,766	26,325	2,766	26,325	2,766
R^2	0.37	0.29	0.37	0.30	0.36	0.17

Notes: The dependent variable in every regression is *Skip* × 100; *Skip* = 1 if at least one course is skipped and is 0 otherwise. Columns (1), (3), and (5) use the full sample and columns (2), (4), and (6) use the FF subsample, excluding the applicants with missing values in the control variables. Other control variables include gender, postal-code-level median income (in logarithm), citizenship status, region born, language spoken at home, high school fixed effects, and dummy variables for the number of full-fee courses. "School fees > AUD11,000" is an indicator that the applicant attends a school that charges more than AUD11,000 (approximately USD8,000) in fees. Standard errors clustered at high school level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

variables that correspond to the number of FF courses listed by an applicant,[8] because the number of FF courses listed may have a "mechanical" effect on the number of skips. If an FF course is not listed, then no RF course can be skipped, by definition. Hence, the more FF courses that are listed, the larger the opportunities for skipping are.

Columns (1) and (2) are baseline regressions showing the negative relation between skips and scores. To account for H_2, we include as an explanatory variable the General Achievement Test (GAT). Although GAT and score (ENTER) are both correlated with an applicant's ability, GAT is not correlated with an applicant's admission probabilities as this test is not used in admission decisions. Furthermore, GAT is likely to be a better measure of an applicant's ability to *understand the mechanism* used by VTAC, because it is a test of general knowledge and skills in written communication, mathematics, science and technology, humanities, arts and social sciences, similar in content to the SAT/ACT in the U.S. GAT is designed to avoid testing the specific content of classes that applicants may take in high school. In contrast, the score is an aggregate of grades achieved by an applicant from her high school classes. These classes may differ significantly among applicants with the same score. The score is the most similar to a grade point average in the U.S. system.[9] The generic and standardized nature of GAT likely makes it a better measure of the comprehension of the mechanism compared to the score.

When both GAT and the score are included in the regression, the coefficient on the score shows how much more likely, in percentage terms, an applicant with the same GAT but a different admission probability (captured by the score) is to make a skip. The results are reported in columns (3) and (4) of Table 2. The coefficient on GAT is negative and significant, suggesting that H_2 is valid and that cognitive abilities may play a role in the explanation of skips. Even after

[8] Our results are robust to the exclusion of these dummies.

[9] The assessment for each subject is standardized across schools, similarly to Advanced Placement exams in the U.S. The assessments are then aggregated using different weights into an applicant's aggregate points. Using the aggregate points, a rank of each applicant is derived. We refer to this rank as a score.

controlling for cognitive ability, $Score_i$ continues to be negatively correlated with $Skip_i$, which is consistent with H_1.

The specifications in columns (5) and (6) in Table 2 control for high school fees. Victoria has a significant private school system. These schools have both well-resourced career advising services and disproportionate numbers of applicants with higher scores. Thus, we include an interaction of the score and an indicator that the applicant attends a school that charges more than AUD11,000 (approximately USD8,000) in fees.[10] Applicants from expensive private schools are less likely to skip an RF course than those with the same score from public schools.

To address H_3, we exploit the prediction of the location of skipped courses in an applicant's ROL. H_3 predicts that courses will be skipped from the top of an ROL, while H_1 does not impose any such restriction. Consider two applicants, i and j, who are identical except that i skips and j does not. H_3 generates the following testable prediction (relative to H_1).

(i) Suppose that H_3 holds. As i drops high-cutoff courses from the top of i's ROL and keeps the bottom of the list the same, we expect that the cutoffs of top-ranked courses in i's ROL will be lower than the cutoffs of top-ranked courses in j's ROL. The cutoffs for bottom-ranked courses will be the same for both i and j.

(ii) Suppose that H_3 does not hold and that i skips courses from anywhere in i' ROL. Then, the cutoffs of both top- and bottom-ranked courses in i's list will be lower than those in j's list.

Based on these predictions, we use the following regression to test H_3:

$$\textit{Cutoffs top-ranked courses}_i - \textit{Cutoffs bottom-ranked courses}_i = \gamma + \delta Skip_i + \zeta Score_i + Controls_i + \epsilon_i. \tag{2}$$

We expect the coefficient δ to be negative if H_3 holds. The results are presented in Table 3. We use three different definitions for *Cutoffs top-ranked courses*$_i$ and *Cutoffs bottom-ranked courses*$_i$. In specifications (1)-(3), we take the difference between the cutoffs of the top-ranked and the bottom-ranked RF courses for an individual applicant; in specifications (4)-(6), we take the difference between the average of the two highest-ranked and the average of two lowest-ranked RF courses; and in (7)-(9), we do the same with three RF courses.[11] In addition, as robustness checks, we vary the sample and the control variables included in the regressions.

Table 3 shows that the coefficient on skip is insignificant in nearly all the regressions; when it is significant, it has a positive sign (column 3). Therefore, there is no evidence that applicants eliminate high-cutoff courses from the top of their ROL, and thus, there is no support for H_3 that applicants attempt manipulations.

[10] We cannot include the dummy variable "School fees $>$ AUD$11,000$" alone because of the inclusion of high school fixed effects in all regressions.

[11] Recall that the relative position of RF and FF courses in an ROL has no payoff consequences in this environment. We only focus on the RF courses in an ROL because a skip involves an omitted RF course and because the majority of the courses listed in an ROL are RF. Among applicants in the FF subsample, an ROL on average includes 5.34 RF courses and 2.33 FF courses (cf. Table 1).

Table 3: Correlation between Skip and the Difference between Cutoffs of Top- and Bottom-ranked Courses

Dependent variable	Diff. b/t Top & Bottom			Diff. b/t Top 2 & Bottom 2			Diff. b/t Top 3 & Bottom 3		
Sample	FF Subsample		Full	FF Subsample		Full	FF Subsample		Full
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Skip	-1.43	-0.34	1.41*	-0.73	-0.19	0.68	-0.09	0.17	-0.06
	(1.11)	(1.13)	(0.83)	(1.03)	(1.03)	(0.79)	(0.68)	(0.68)	(0.42)
Score	0.12**	0.10**	0.14***	0.10**	0.09**	0.11***	0.02	0.02	0.01***
	(0.05)	(0.05)	(0.01)	(0.04)	(0.04)	(0.01)	(0.02)	(0.02)	(0.00)
GAT	-0.06	-0.06	-0.01	-0.03	-0.02	0.01	0.00	0.00	0.00
	(0.06)	(0.06)	(0.02)	(0.05)	(0.05)	(0.02)	(0.02)	(0.02)	(0.01)
ROL length		1.18***	0.97***		1.05***	0.96***		0.56***	0.35***
		(0.22)	(0.06)		(0.27)	(0.06)		(0.20)	(0.04)
Other Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
# of Applicants	2,517	2,517	25,168	1,990	1,990	21,542	1,212	1,212	15,596
R^2	0.21	0.22	0.06	0.25	0.26	0.07	0.37	0.38	0.04

Notes: The dependent variable of all regressions is the difference between the cutoffs of top- and bottom-ranked RF courses but varies in the number of courses we consider. Columns (1)—(3) use the top- and the bottom-ranked courses; columns (4)—(6) use the top two and the bottom two; and columns (7)—(9) use the top three and the bottom three courses. Columns (1), (2), (4), (5), (7), and (8) use the FF subsample and columns (3), (6) and (9) use the full sample, excluding the applicants with missing values in the control variables. ROL length refers to the number of RF courses listed in ROL. Other controls include gender, postal-code-level income (in logarithm), citizenship status, region born, language spoken at home, high school fixed effects. Standard errors clustered at high school level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.2.2 Payoff-relevant Mistakes

Recall that hypothesis H_1 implies that applicants skip out-of-reach courses. Thus, omitting a within-reach course is an unexplained mistake. That is, unlike skips, which we expect to vary systematically with score, H_1 does not imply any particular patterns in payoff-relevant mistakes. In this section, we investigate the characteristics of those who make payoff-relevant mistakes. Due to the sample size, we use the upper bound definition of payoff relevance: for an applicant who skips an RF course, the mistake is payoff-relevant if adding the RF course at the top of the applicant's ROL would change the applicant's assignment.

We run the following regression:

$$\textit{Payoff-relevant Mistake}_i \times 100 = \theta + \iota \textit{Score}_i + \textit{Controls}_i + \epsilon_i, \tag{3}$$

where $\textit{Payoff-relevant Mistake}_i = 1$ if i's skip is payoff-relevant, 0 otherwise.

In Table 4, we present the results. Note that the two variables that have been significantly negative in the regressions of skips, GAT and interaction of score and school fees, no longer have robust significance. Only the interaction is significantly negative (at a 10% level) in one regression (column 4). Combining the regression results of equations (1) and (3), which are reported in Tables 2 and 4 respectively, it appears that while higher-ability applicants as measured by GAT are more successful in avoiding skips, lower-ability applicants do not make more payoff-relevant mistakes. Another notable observation is that payoff-relevant mistakes are positively correlated with score, while skips are negatively correlated with score. The positive correlation

Table 4: Probability of Making Payoff-relevant Mistakes

			Sub-sample including applicants who			
	Full sample		FF subsample		Skip	
	(1)	(2)	(3)	(4)	(5)	(6)
Score	0.01***	0.01***	0.16***	0.19***	0.66***	0.61***
	(0.00)	(0.00)	(0.04)	(0.04)	(0.15)	(0.15)
GAT	0.00	0.00	0.00	0.01	0.31	0.28
	(0.01)	(0.01)	(0.06)	(0.06)	(0.19)	(0.19)
(School fees >AUD11,000) × *Score*		0.01		-0.11*		0.22
		(0.01)		(0.06)		(0.27)
Other controls	Yes	Yes	Yes	Yes	Yes	Yes
# of Applicants	26,325	26,325	2,766	2,766	947	947
R^2	0.14	0.14	0.25	0.25	0.48	0.48

Notes: The dependent variable, $\textit{Payoff-relevant Mistake}_i \times 100$, is equal to 100 if an applicant makes at least one payoff-relevant mistake and is 0 otherwise. Columns (1) and (2) are based on the full sample and columns (3) and (4) are based on the FF sample of applicants, excluding the applicants with missing values in the control variables. Columns (5) and (6) include only those applicants from the full sample who make at least one skip. Other control variables include gender, postal-code-level income (in logarithm), citizenship status, region born, language spoken at home, high school fixed effects and dummy variables for the number of full-fee courses. Standard errors clustered at high school level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

may be explained "mechanically": skipping an RF course is more likely to be payoff-relevant for an applicant with a high score.[12]

2.2.3 Changes in ROL Over Time

To further test our hypothesis that skips are the outcomes of omitting out-of-reach courses, we use an unusual feature of the Victorian centralized mechanism: a requirement that applicants submit their "preliminary" ROL several months before the deadline of their final ROL and before applicants learn their scores. If not changed, a preliminary ROL becomes final and is used for admission. As a small effort is needed to change an ROL, we treat the preliminary ROL as the best estimate of the final ROL that an applicant would submit given the information that the applicant has at the time.

We investigate how the number of skips and payoff-relevant mistakes changes over time using the following two regressions:

$$\Delta(\#\textit{Skips}_i) = \tau^s + \textit{Demeaned\_Controls}_i + \epsilon_i, \quad (4)$$

$$\Delta(\#\textit{Payoff-relevant Mistakes}_i) = \tau^m + \textit{Demeaned\_Controls}_i + \epsilon_i, \quad (5)$$

where $\Delta(\#\textit{Skips}_i)$ represents the difference between the numbers of skips in the final and the preliminary ROLs and $\Delta(\#\textit{Payoff-relevant Mistakes}_i)$ is the analogous difference for payoff-relevant mistakes. The control variables are demeaned, and therefore, the constants τ^s and τ^m measure the average change over time in our sample. Equivalently, we calculate the average changes (τ^s

[12] A higher score may make a skipped RF course feasible. When we put such a course back at the top of an ROL, it will be counted as a payoff-relevant mistake.

and τ^m) by setting all the control variables at their sample mean.

Columns (1)—(4) of Table 5 show the results.[13] As the number of listed FF courses may have a mechanical effect on skips and mistakes, we add the change in the number of FF courses as a control variable in columns (3) and (4). In all specifications for payoff-relevant mistakes (columns 2 and 4), the constant is negative and significant, implying that the number of payoff-relevant mistakes decreases over time on average in the sample. In contrast, the effect of the revision of ROL on skips (columns 1 and 3) is significantly positive. Both results are consistent with applicants responding to new information on the set of within-reach colleges and re-optimizing their ROLs. Alternative explanations, such as learning about the mechanism, will not predict the increase in the number of skips.

Table 5: Skips and Payoff-relevant Mistakes: Changes over Time

	Changes over time				Effects of a shock to score			
	#Skips	#Mistakes	#Skips	#Mistakes	#Skips	#Mistakes	#Skips	#Mistakes
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	1.05***	-0.12**	0.72***	-0.19***	1.04***	-0.13**	0.71***	-0.20***
	(0.14)	(0.05)	(0.11)	(0.05)	(0.14)	(0.05)	(0.11)	(0.05)
Shock to *Score*/100					-2.33	-1.97***	-0.49	-1.61**
					(2.28)	(0.72)	(1.72)	(0.64)
Change in # FF courses			43.24***	8.69***			43.24***	8.68***
			(2.75)	(1.26)			(2.75)	(1.26)
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of Applicants	26,325	26,325	26,325	26,325	26,325	26,325	26,325	26,325
R^2	0.02	0.04	0.13	0.42	0.02	0.04	0.14	0.41

Notes: “Mistake” means a payoff-relevant mistake. “FF courses” means full-fee courses. The dependent variable in regressions (1), (3), (5) and (7) is the difference in the number of skips between the final ROL and the preliminary ROL. The dependent variable in regression (2), (4), (6) and (8) is the difference in the number of payoff-relevant mistakes between the final ROL and the preliminary (November) ROL. “Shock to *Score*” is the difference between an applicant's realized and expected scores. All regressions are based on the full sample of applicants with non-missing values of control variables. All control variables are demeaned, while other (demeaned) control variables include gender, postal-code-level income (in logarithm), citizenship status, region born, language spoken at home, high school fixed effects. Standard errors clustered at high school level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Furthermore, to capture an applicant's response to an unexpectedly high or low score, we add an additional explanatory variable, shock to score, into equations (4) and (5). It is defined as the difference between an applicants' realized and expected scores, where the expected score is calculated using GAT (see Appendix A.3 for the definition of expected score).

We report the results in columns (5)—(8) of Table 5. Mechanically, if applicants do not change their ROL to improve its optimality, a positive shock will increase the probability that a skip becomes a payoff-relevant mistake. The results show that the number of payoff-relevant mistakes *decreases* with a positive shock (columns 6 and 8), implying that applicants eliminate skips when they realize that they are more likely to become payoff relevant. At the same time, a shock has no significant effect on the number of skips (columns 5 and 7).

[13] These results are robust to the inclusion of GAT as an additional control variable.

3 Theoretical Implications of Strategic Mistakes

The preceding analysis suggests that applicants tend to make mistakes when they are unlikely to affect the outcome. In this section, we explore the implications of these findings for the empirical methods that are commonly employed in market design research. Specifically, we consider a large matching market operated by the Gale and Shapley's deferred acceptance algorithm (Gale and Shapley, 1962), and adopt an equilibrium concept that permits participants to make mistakes as long as they become virtually payoff irrelevant as the market size grows arbitrarily large.[14]

3.1 Primitives

We begin with Azevedo and Leshno (2016) (denoted as AL) as our modeling benchmark. A (generic) economy consists of a finite set of *courses*, or *colleges*, $C = \{c_1, ..., c_C\}$ and a set of *applicants*. Each applicant has a type $\theta = (u, s)$, where $u = (u_1, ..., u_C) \in [\underline{u}, \overline{u}]^C$ is a vector of von-Neumann Morgenstern utilities of attending colleges for some $\underline{u} \leq 0 < \overline{u}$, and $s = (s_1, ..., s_C) \in [0, 1]^C$ is a vector of scores representing the colleges' preferences or applicants' priorities at colleges, with an applicant with a higher score having a higher priority at a college. A vector u induces ordinal preferences over colleges, denoted by an ROL of "acceptable" colleges, $\rho(u)$, of length $0 \leq \ell \leq C$.[15] Assuming that an applicant has an outside option of zero payoff, the model allows for the possibility that applicants may find some colleges unacceptable. Let $\Theta = [\underline{u}, \overline{u}]^C \times [0, 1]^C$ denote the set of applicant types. One special case is the **serial dictatorship** in which colleges' preferences for applicants are given by a single score. Australian tertiary admissions can be seen as a case of serial dictatorship. We shall incorporate this case by an additional restriction that the scores of each applicant satisfy $s_1 = ... = s_C$.

A *continuum economy* consists of the same finite set of colleges and a unit mass of applicants with type $\theta \in \Theta$ and is given by $E = [\eta, S]$, where η is a probability measure over Θ representing the distribution of the applicant population over types, and S represents the masses of seats $S = (S_1, ..., S_C)$ available at the colleges, where $S_j > 0$ and $\sum_{j=1}^{C} S_1 < 1$. We assume that η admits continuous density that is positive in the interior of its support (i.e., full support). In the case of serial dictatorship, this assumption holds with a reduced dimensionality of support; applicants' scores are one-dimensional numbers in $[0, 1]$. The atomlessness ensures that indifferences either in applicant preferences or in college preference arises only for a measure 0 set of applicants.[16] The

[14] There is a growing number of studies on large matching markets, including Pittel (1989), Immorlica and Mahdian (2005), Kojima and Pathak (2009), Lee (2014), Lee and Yariv (2014), Ashlagi, Kanoria, and Leshno (forthcoming), Che and Tercieux (2015a), Che and Tercieux (2015b), Abdulkadiroglu, Che, and Yasuda (2015), Che and Kojima (2010), Liu and Pycia (2011), Azevedo and Leshno (2016), Azevedo and Hatfield (2012) and Che, Kim, and Kojima (2013). The current work differs largely from these papers because of the solution concept that we adopt and the issue of focus here.

[15] In the case of a tie, $\rho(u)$ produces a ranking by breaking the tie in some arbitrary (but exogenous) way. Since we shall assume that the distribution of the types is atomless, the tie-breaking becomes immaterial.

[16] At the same time, atomlessness rules out the environments where some applicants are ranked the same at some schools and lotteries are used to break ties, such as NYC's high school admissions (Abdulkadiroglu, Agarwal, and

full-support assumption means that both applicants' and colleges' preferences are rich (except for the case of serial dictatorship). A *matching* is defined as a mapping $\mu : C \cup \Theta \to 2^{\Theta} \cup (C \cup \Theta)$ satisfying the usual two-sidedness and consistency requirements as well as "open on the right" as defined in AL (see p. 1241). A *stable matching* is also defined in the usual way satisfying individual rationality and no-blocking.[17]

According to AL, a stable matching is characterized via *market-clearing cutoffs*, $P = (P_1, ..., P_C) \in [0,1]^C$, satisfying the demand-supply condition $D_c(P) \leq S_c$ with equality in case of $P_c > 0$ for each $c \in C$, where the demand $D_c(P)$ for college c is given by the measure of applicants whose favorite college among all feasible ones (i.e., with cutoffs less than his scores) is c. Specifically, given the market-clearing cutoffs P, the associated stable matching assigns those who demand c at P to college c. Given the full-support assumption, Theorem 1-i of AL guarantees a unique market-clearing price P^* and a unique stable matching μ^*. Given the continuous density assumption, $D(\cdot)$ is C^1 and $\partial D(P^*)$ is invertible.

With the continuum economy E serving as a benchmark, we are interested in a sequence of finite random economies approximating E in the limit. Specifically, let $F^k = [\eta^k, S^k]$ be a *k-random economy* that consists of k applicants each with type θ drawn independently according to η, and the vector $S^k = [k \cdot S]/k$ of capacity per applicant, where $[x]$ is the vector of integers nearest to x (with a rounding down in case of a tie); η^k is a random measure. A matching is defined in the usual way.

Consider a sequence of k-random economies $\{F^k\}$. An applicant-proposing DA is used to assign applicants to colleges. We assume that colleges are acting passively and report their preferences and capacities truthfully.[18] We are interested in characterizing an equilibrium behavior of the applicants in the DA. In one sense, this task is trivial: since DA is strategy-proof, it is a weakly dominant strategy for each player to rank order all acceptable colleges (i.e., with payoff $u_c \geq 0$) according to true preference order. We call such a strategy the **truthful-reporting strategy** (TRS).[19] We are, however, interested in a more robust solution concept allowing for any approximately-optimal behavior. We assume that each applicant observes her own type θ but not the types of other applicants; as usual, all applicants understand as common knowledge the structure of the game. Given this structure, the DA induces a Bayesian game in which the strategy of each applicant specifies a distribution over ROLs of length no greater than C as a function of her type θ.

In any game (either the limit or the k-random economy), applicant i's Bayesian strategy is a measurable function $\sigma_i : \Theta \to \Delta(\mathcal{R})$, where $\mathcal{R}$ is the set of all possible ROLs an applicant may

Pathak, Forthcoming).

[17] Individual rationality requires that no participant (an applicant or a college) is assigned a partner that is not acceptable. No blocking means that no applicant-college pair exists such that the applicant prefers the college over her assignment and the college has either a positive measure of vacant positions or admits a positive measure of applicants whom the college ranks below that applicant.

[18] In the context of VTAC, the common college preferences make this an ex-post equilibrium strategy (see Che and Koh, 2016).

[19] By this definition, TRS does not allow applicants to rank unacceptable colleges, which reduces the multiplicity of equilibria and thus gives TRS a better chance to be the unique equilibrium. However, applying to unacceptable options can happen in real-life matching markets, as documented in He (2017).

submit. Note that the strategies can be asymmetric; i.e., we do not restrict attention to symmetric equilibria. In any economy (either continuum or finite), a profile of applicant strategies induces cutoffs $P \in [0,1]^C$ and a distribution of types of applicants assigned to alternative colleges, which we refer to as *an outcome*. We say a college c is *feasible* to an applicant if her score at c is no less than P_c, and we say that an applicant *demands* college c if c is feasible and she ranks c in her ROL ahead of any other feasible colleges.

We are interested in the following solution concept:

Definition 1. *For a sequence $\{F^k\}$ of k-random economies, the associated sequence $\{(\sigma^k_{1\le i\le k})\}_k$ of strategy profiles is said to be a* ***robust equilibrium*** *if, for any $\epsilon > 0$, there exists $K \in \mathbb{N}$ such that for $k > K$, $\{(\sigma^k_{1\le i\le k})\}_k$ is an interim ϵ-Bayes Nash equilibrium—namely, for i, σ^k_i gives applicant i of each type θ a payoff within ϵ of the highest possible (i.e., supremum) payoff she can get by using any strategy when all the others employ σ^k_{-i}.*

This solution concept is arguably more sensible than the exact Bayesian Nash equilibrium if, for a variety of reasons, market participants may not play their best response exactly, but they do approximately in the sense of not making mistakes of significant payoff consequences, in a sufficiently large economy.

Indeed, without relaxing a solution concept in some way, one cannot explain the kind of departure from the dominant strategy observed in the preceding section. To see why, recall that in the continuum economy, any stable matching, and hence the outcome of DA mechanism, gives rise to cutoffs that are degenerate. The applicants then face no uncertainty about the set of feasible colleges. Hence, one can easily construct a dominated strategy that can do just as well as a dominant strategy—namely submitting the truthful rank-ordering of colleges. For instance, an applicant may list only one college, the most preferred feasible one, and do just as well as reporting her truthful ROL. However, such a strategy will not be optimal for any finite economy, no matter how large it is. In a finite economy, the cutoffs are not degenerate, so any strategy departing from the truthful-reporting strategy will result in payoff loss with a positive probability. Hence, to explain a behavior that departs from a dominant strategy, one must relax exact optimality on the applicants' behavior. At the 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 by allowing applicants to make some mistakes but requiring that the payoff losses from the mistakes disappear as the market grows arbitrarily large.

3.2 Analysis of Robust Equilibria

Robust equilibria require that applicants should make no mistakes with any real payoff consequences as the market gets large. Does this requirement mean that a large fraction of applicants must report truthfully in a strategy-proof mechanism? We show below that this need not be the case. Specifically, we will construct strategies that are not TRS, yet do not entail significant payoff loss for almost all types of applicants in a large economy.

To begin, recall P^* (the unique market-clearing cutoffs for the limit continuum economy). We define a **stable-response strategy** (SRS) as *any* strategy that demands the most preferred feasible college for an applicant given P^* (i.e., she ranks that college ahead of all other feasible colleges). The set of SRSs is typically large. For example, suppose that $C = \{1, 2, 3\}$, and colleges 2 and 3 are feasible, and an applicant prefers 2 to 3. Then, 7 ROLs—1-2-3, 2-3-1, 2-1-3, 1-2, 2-1, 2-3, 2—constitute her SRSs out of 10 possible ROLs she can choose from. Formally, if an applicant has $\ell \leq C$ feasible colleges, then the number of SRSs is $\sum_{a \leq \ell-1, b \leq C-\ell} \binom{a+b+1}{b} a!b!$. For each type $\theta = (u, s)$ with $\rho(u) \neq \emptyset$ (i.e., with at least one acceptable college), there exists at least one SRS that is untruthful.[20]

For the next result, we construct such a strategy. To begin, let $\hat{r} : \mathcal{R} \times [0, 1]^C \to \mathcal{R}$ be a transformation function that maps a preference order $\rho \in \mathcal{R}$ and a score vector s to an ROL with the properties that (i) $\hat{r}(\rho, s) \neq \rho$ for all $\rho \neq \emptyset$ (i.e., untruthful) and $\hat{r}(\emptyset, s) = \emptyset$; and (ii) $\hat{r}(\rho, s)$ ranks the most preferred feasible college ahead of all other feasible colleges for each $\rho \neq \emptyset$ (where feasibility is defined given P^*). The existence of such a strategy is established above. We then define an SRS $\hat{R} : \Theta \to \mathcal{R}$, given by $\hat{R}(u, s) := \hat{r}(\rho(u), s)$, for all $\theta = (u, s)$.[21] Let

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

be the set of types whose score for some college is δ-close to its market-clearing cutoff for the continuum economy.

Theorem 1. *Fix any arbitrarily small $(\delta, \gamma) \in (0, 1)^2$. The following sequence of strategy profiles is a robust equilibrium: in each k-random economy,*

- *all applicants with types $\theta \in \Theta^\delta$ play TRS and*
- *all applicants with types $\theta \notin \Theta^\delta$ randomize between TRS (with probability γ) and untruthful SRS $\hat{R}(\theta)$ (with probability $1 - \gamma$).*

The intuition for Theorem 1 rests on the observation that the uncertainty about cutoffs, and hence the payoff risk of playing non-TRS strategies, vanishes in a sufficiently large economy. Specifically, the sequence of strategy profiles that we construct satisfies two properties: (a) it prescribes a large fraction of participants to deviate from TRS with a large enough probability, and yet, (b) it gives rise to cutoffs P^* in the limit continuum economy, namely the cutoffs that would prevail if *all* applicants played TRS. That these two properties can be satisfied simultaneously is not trivial and requires some care, since the cutoffs may change as applicants deviate from TRS. Indeed, the feature that all applicants play TRS with some small probability γ is designed to ensure

[20] If an applicant's most preferred college is infeasible, she can drop that college. If her favorite college is feasible, then she can drop a less-preferred acceptable college or add an unacceptable college at the bottom of the list, whichever exists.

[21] Note that this SRS is constructed via the transformation function $\hat{r}$. In principle, an SRS can be defined without such a transformation function, although this particular construction simplifies the proof below.

that the same unique stable matching obtains under the prescribed strategies. Given these facts, the well-known limit theorem, due to Glivenko and Cantelli, implies that the (random) cutoffs for any large economy generated by the i.i.d. sample of applicants are sufficiently concentrated around P^* under the prescribed strategies, so that applicants whose scores are δ away from P^* will suffer very little payoff loss from playing any SRS that deviates (possibly significantly) from TRS. Indeed, our construction ensures that these are precisely the applicants who play an SRS and deviate from TRS. Since (δ, γ) is arbitrary, the following striking conclusion emerges.

Corollary 1. *There exists a robust equilibrium in which every applicant, except for those with no acceptable colleges, submits an untruthful ROL with probability arbitrarily close to one.*

To the extent that a robust equilibrium is a reasonable solution concept, the result implies that we should not be surprised to observe a non-negligible fraction of market participants making "mistakes"—more precisely, playing dominated strategies—even in a strategy-proof environment. This result also calls into question any empirical method relying on TRS—any particular strategy for that matter—as an identifying restriction.

If strategic mistakes of the types observed in the preceding section undermine the prediction of TRS, do they also undermine the stability of the outcome? This is an important question on two accounts. If mistakes jeopardize stability in a significant way, then this situation may call into question the rationale for DA to the extent that mistakes do occur. If stability remains largely intact despite the presence of mistakes, then they do not raise a fundamental concern.

Aside from the stability prediction, the question is also important from the perspective of empirical approaches. Stability has been an important identification assumption invoked by researchers for preference estimation in a number of contexts, e.g., in decentralized two-sided matching (for surveys, see, Fox, 2009; Chiappori and Salanié, 2016) as well as centralized matching with or without transfers (e.g., Fox and Bajari, 2013; Agarwal, 2015). Our second theorem shows that strategic mistakes captured by robust equilibrium leaves the stability property of DA largely unscathed. To this end, we begin by defining a notion of stability in a large market.

Definition 2. *For a sequence $\{F^k\}$ of k-random economies with DA matching, the associated sequence $\{(\sigma^k_{1\leq i\leq k})\}_k$ of strategy profiles is said to be **asymptotically stable** if the fraction of applicants assigned their most preferred feasible colleges (given the equilibrium cutoffs) converges in probability to 1 as $k \to \infty$.*[22]

We call a sequence $\{(\sigma^k_{1\leq i\leq k})\}_k$ of strategy profiles **regular** if there exists some $\gamma > 0$ such that the proportion of applicants playing TRS is at least $\gamma > 0$. We now state the main theorem:

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

[22] More formally, we require that for any $\epsilon > 0$ there exists $K \in \mathbb{N}$ such that in any random-k 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.

A key argument for this theorem is to show that in *any* regular robust equilibrium the uncertainty about colleges' DA cutoffs vanishes as the market becomes large. Asymptotic stability then follows immediately from this, since the applicants must *virtually* know what the true cutoffs are in the limit as the market grows large; hence all applicants (more precisely, all except for a vanishing proportion) should play their stable responses relative to the *true* cutoffs; otherwise, their mistakes entail significant payoff losses even in the limit, which contradicts the robustness requirement of the solution concept.

The argument is nontrivial since the cutoffs in the k-random economy depend on the equilibrium strategies and since our robust equilibrium concept imposes very little structure on these strategies. To prove the argument, we fix an arbitrary sequence of regular robust equilibrium strategy profiles $\{(\sigma^k_{1\leq i\leq k})\}_k$ and study the sequence of (random) demand vectors $\{D^k(P)\}$ for any fixed cutoffs P induced by these strategies.[23] Although very little can be deduced about these strategies, the fact that an individual applicant's influence on the (aggregate) demand vector is vanishing in the limit leads to a version of law of large numbers: namely, $D^k(P)$ converges pointwise to its expectation $\bar{D}^k(P) := \mathbb{E}[D^k(P)]$ at the fixed P in probability as $k \to \infty$ (see McDiarmid, 1989). Further, a subsequence $\{\bar{D}^{k_\ell}(\cdot)\}$ of the expected demand functions converges uniformly to some continuous function $\bar{D}(\cdot)$ (the Arzela-Ascoli theorem). Combining these two results and a further argument in the spirit of the Glivenko-Cantelli theorem show that, along a sub-subsequence of k-random economies, the actual (random) demand $D^{k_{\ell_j}}(\cdot)$ converges uniformly to $\bar{D}(\cdot)$ in probability. Finally, the *regularity* of the strategies implies that $\partial\bar{D}(\cdot)$ is invertible, which in turn implies that the cutoffs of k-random economies converge in probability to some degenerate cutoffs $\bar{P}$ along that sub-subsequence as $k \to \infty$.[24]

While Theorem 2 already provides some justification for *stability* as an identification assumption for a sufficiently large market, a question arises as to whether the robustness concept would predict the same outcome as would emerge had all applicants reported their preferences truthfully. Our answer is in the affirmative:[25]

Corollary 2. *Fix any regular robust equilibrium. The associated sequence of outcomes converges in probability to the unique stable matching outcome of the continuum economy $E = [\eta, S]$. That is, the limit outcome would be the same as if all applicants reported their preferences truthfully.*

[23] A vector $D^k(P) = (D^k_{c_1}(P), ..., D^k_{c_C}(P))$ describes the fractions of applicants who "demand" alternative colleges at vector P of cutoffs given their ROL strategies in the k-random economy. More precisely, a component $D^k_c(P)$ of the vector is the fraction of applicants in economy F^k for whom c is the favorite feasible college according to their chosen ROLs and the cutoffs P. Note that the demand vector is a random variable since the applicant types are random and they may play mixed strategies.

[24] It is enough to show this convergence occurs along a sub-subsequence of k-random economies: if asymptotic stability were violated, then a nonvanishing proportion of applications must play non-SRS strategies on a subsequence of k-random economies, and the preceding convergence argument can be used to show that these strategies entail significant payoff losses to applicants for a sub-subsequence of these economies (for which the cutoffs converge to degenerate cutoffs). See the precise argument in Appendix B.

[25] This 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 $\epsilon-$BNE (which is weaker than BNE) converges to an exact BNE as the economy grows large.

Although Theorem 1 questions truth-telling as a *behavioral prediction*, Corollary 2 supports truth-telling as a means for predicting an *outcome*. In this sense, the corollary validates the vast theoretical literature on strategy-proof mechanisms that assume truth-telling. This result also suggests that when one evaluates the *outcome* of a counterfactual scenario involving a strategy-proof mechanism, one can simply assume that applicants report their preferences truthfully in that scenario, as we will do in our Monte Carlo simulation.

3.3 Discussion

In our setting, colleges strictly rank applicants by some score, and applicants know their scores before playing the college admissions game. These features ensure that the uncertainty applicants may face about feasibility of alternative colleges vanishes as the market grows large. By imposing this condition, we thus exclude school choice problems in which applicants are ranked by a lottery after they submit their ROLs (see Pathak, 2011, for a survey). The theorems above suggest that estimation techniques developed for these settings that use truth-telling as an identifying assumption, such as Abdulkadiroglu, Agarwal, and Pathak (Forthcoming), should not be applied in settings where colleges rank applicants strictly and applicants can predict their ranking. The settings that satisfy both conditions are common. These settings are typical in tertiary and selective school admissions and may be used in assignments to public high schools (where "score" may refer to an exam score or to a continuously measured distance from residence to school).

Furthermore, when evaluating a counterfactual policy, a number of papers use the ROLs submitted by the applicants under the original policy (Roth and Peranson, 1999; Combe, Tercieux, and Terrier, 2016; Veski, Biró, Poder, and Lauri, 2016). These works rely, explicitly or implicitly, on the assumption that if the mechanism is strategy-proof, applicants submit their true preferences across different policy environments. The theorems above indicate that this assumption is not well justified; indeed, we demonstrate in the next section that such an assumption can lead to significant biases evaluating welfare effects.

The comparison between stability and truthful-reporting strategies is also considered in Fack, Grenet, and He (2017). The authors consider exact Bayesian Nash equilibrium and argue that the truthful-reporting strategy may not be the unique equilibrium and that it may not be an equilibrium at all if there is an application cost. However, their approach does not explain the mistakes observed in our data. Given that there is an uncertainty in cutoffs, ranking an FF course and skipping the corresponding RF course is suboptimal. Skipping cannot be justified by an application cost either because the cost of ranking an RF course is plausibly zero in this scenario. In contrast, our Theorem 1 provides a natural explanation for such mistakes.

A more significant difference between the two papers can be found in Theorem 2. Fack, Grenet, and He (2017) justify stability by showing *a* sequence of Bayesian Nash equilibria that is asymptotically stable in the limit as the economy gets large. Their result, however, does not rule out asymptotically unstable equilibria even in a large market. Going beyond mere existence, our Theorem 2 shows that all regular robust equilibria are asymptotically stable, thus further justifying the use of stability for identification. The proof is more challenging than in Fack, Grenet, and He (2017) and uses different techniques.

Our Theorem 2 has a similar flavor to the main result of Deb and Kalai (2015). Theorem 2 of Deb and Kalai implies that all participants enjoy approximately their full information optimal payoff (holding the actions of the other participants fixed) from any approximate Bayesian Nash equilibrium.[26] Despite the resemblance, their theorem is not applicable in our setting. Specifically, a crucial condition needed for their result is LC2: the effect that any participant can unilaterally have on an opponent's payoff is uniformly bounded and decreases with the number of participants in the game. This condition does not hold in our setting since even in an arbitrarily large economy, an applicant may be displaced from a college because of a single change in a submitted ROL by some other applicant. Indeed, instead of imposing continuity directly on the payoff function of the applicants (which is not well justified in our setting), our result exploits the continuity exhibited by the aggregate demand functions generated by randomly sampling individuals from the same distribution.[27]

4 Analysis with Monte Carlo Simulations

This section provides details on the Monte Carlo simulations that we perform to assess the implications of our theoretical results. Section 4.1 specifies the model, Section 4.2 describes the data generating processes, Section 4.3 presents the estimation and testing procedures, Section 4.4 discusses the estimation results, and, finally, Section 4.5 presents the counterfactual analyses.

4.1 Model Specification

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 as in Figure C.2 (Appendix C) which plots one simulation sample. The colleges (represented by big red dots) are evenly located

[26] This result would lead to the same conclusion as Theorem 2, since most of the applicants would be matched with their favorite feasible college.

[27] Similar to Deb and Kalai, McDiarmids' inequality also plays a role in our argument, but its use, as well as the overall proof strategy, is quite different from theirs. In our case, the strategic interaction occurs via "cutoffs" of colleges playing the role of market-clearing prices. It is crucial for the uncertainty in the cutoffs to disappear in a large market (so that any non-SRS strategy could lead to a discrete payoff loss). This requires the (random) aggregate demand functions to converge in probability to a degenerate continuous function. McDiarmid's inequality (McDiarmid, 1989), along with Arzela-Ascoli and Glivenko-Cantelli theorems, proves useful for this step.

on a circle of radius 1/2 around the centroid, and the applicants (represented by small blue dots) are uniformly distributed across the area. The Cartesian distance between applicant i and college c is denoted by $d_{i,c}$.

Applicants are matched with colleges through a serial dictatorship. 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, as in the case in Victoria, Australia. Moreover, the order is common knowledge at the time of submitting ROL.[28]

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.[29] That is, applicant i's utility from being matched with college c is specified as follows:

$$u_{i,c} = \beta_1 \cdot c + \beta_2 \cdot d_{i,c} + \beta_3 \cdot T_i \cdot A_c + \beta_4 \cdot Small_c + \epsilon_{i,c}, \forall i \text{ and } c, \tag{6}$$

where $\beta_1 \cdot c$ is college c's "baseline" quality; $d_{i,c}$ is the distance from applicant i's residence 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, implying that the variance of utility shocks is normalized.

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 a 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.

4.2 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

[28] Our theoretical model in Section 3 considers applicant scores to be private information. That is, every applicant knows her own score but not those of others, and therefore no one knows for sure the exact rank she has at a college. We obtain similar simulation results if we allow scores to be private information.

[29] In the theoretical discussion, we restrict the utility functions to be in $[\underline{u}, \overline{u}]$. One can apply a monotonic transformation to make the utility functions take values on the real line.

the cutoffs in each simulation sample. Figure C.3 in Appendix C 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.

To generate data on applicant behaviors and matching outcomes for preference estimation, we simulate another 200 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) **TRS (Truthful-Reporting Strategy)**: Every applicant submits an ROL of 12 colleges according to her true preferences. Because everyone finds every college acceptable, this is TRS as defined in our theoretical model (Section 3).[30]

(ii) **IRR (Payoff Irrelevant Skips)**: A fraction of applicants skip colleges with which they are never matched according to the simulated distribution of cutoffs. For a given applicant, a skipped college can have a high (expected) cutoff and thus be "out of reach;" the college may also have a low cutoff, but the applicant is always accepted by one of her more-preferred colleges. To specify the fraction of skippers, we first randomly choose approximately 21.4 percent of the applicants to be never-skippers who always rank all colleges truthfully. All other applicants are potential skippers. Among them, we consider three skipping scenarios. In **IRR1**, approximately one-third of the potential skippers skip all the "never-matched" colleges; **IRR2** adds another one-third; and **IRR3** makes all of them skip. Applicants with $T_i = 1$ are more likely to skip than those with $T_i = 0$, as their scores tend to be lower: 95 percent of $T_i = 1$ are potential skippers, compared to 70 percent of $T_i = 0$ (see Tables C.4 and C.5 in the appendix, respectively). Applicants who are never matched may skip all colleges; we randomly choose a college for such applicants, so that they submit one-college ROLs.

(iii) **REL (Payoff Relevant Mistakes)**: **In addition to IRR3**, i.e., given all the potential skippers have skipped the never-matched colleges, we now let them make payoff-relevant mistakes. That is, applicants skip some of the colleges with which they have some chance of being matched according to the simulated distribution of cutoffs. Recall that the joint distribution of cutoffs is only simulated once under the assumption that everyone plays TRS. In each of the four DGPs, **REL1-4**, we specify a threshold admission probability, and the potential skippers omit the colleges at which they have an admission probability lower than the threshold. For REL1-4, the thresholds are 7.5, 15, 22.5, and 30 percent, respectively.

To summarize, for each of the 200 samples, we simulate the matching game 8 times: 1 (TRS or truthful-reporting strategy) + 3 (IRR, or payoff-irrelevant skips) + 4 (REL, or payoff-relevant mistakes).

[30] This is equivalent to the definition of *strict truth-telling* in Fack, Grenet, and He (2017), when there are no unacceptable colleges.

It should be emphasized that the cutoff distribution, which an applicant uses to determine which colleges will be skipped, is not re-simulated in any of these DGPs: we always use the same distribution generated by the 1,000 simulations with applicants reporting truthfully. The distribution of cutoffs does not change with payoff-irrelevant mistakes, so it is of no consequence in the IRR DGPs. For the REL DGPs, it is advantageous to use the same cutoff distribution across all four of them rather than a re-simulated "equilibrium" cutoff distribution, to ensure consistency across DGPs.[31] The most natural candidate for a cutoff distribution to be used across four REL DGPs is the distribution based on truthful reporting, as in the TRS and IRR DGPs. Indeed, for an applicant to calculate an equilibrium cutoff distribution correctly, we need to assume that an applicant correctly predicts not only the distribution of preferences but also the joint distribution of preferences and mistakes. This is a demanding assumption, especially because changes in cutoff distribution need not be monotonic with mistakes.

Table 6 shows how applicants skip in the simulations. The reported percentages are averaged over the 200 samples. Recall that an applicant does not make any (ex-post) payoff-relevant mistake if she is matched with her favorite feasible college, as defined in Section 2. The percentage of applicants who make payoff-relevant mistakes is presented in the second row of Table 6 and ranges from 2% to 10% of the total population. Because every college is acceptable, any instance where an applicant does not rank a college is a skip, also as defined in Section 2. Across the DGPs, 25–79% of applicants make a skip; among this population of skippers, the fraction of applications making payoff-relevant mistakes in the REL simulations ranges from 2.5% to 12.7%.

4.3 Identifying Assumptions and Estimation

With the simulated data at hand, the random utility model described by equation (6) is estimated under three different identifying assumptions:

(i) **WTT** (Weak Truth-Telling). Naturally, one may start by a truth-telling assumption such as TRS. However, in the absence of outside options, TRS implies that every applicant ranks all available colleges. The fact that applicants rarely rank all available colleges motivates a weaker version of truth-telling, following the literature. WTT, which can be considered as a truncated version of TRS, 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 TRS, it is equally susceptible to "mistakes" from our robustness perspective: the robust equilibrium constructed in Theorem 1 also fails WTT. The

[31] An inconsistency may arise for the following reason. If cutoff distributions differ across DGPs, then an applicant who skips a college in, say, REL1 (where the probability threshold for skipping a college is 7.5%), may not skip that same college in REL2. Although the probability threshold increases to 15% in REL2, the applicant's admission probability at that college can increase to above 15% because a college cutoff may decrease in REL2. Furthermore, if that college is sometimes the applicant's feasible college, an estimation based on stability may improve from REL1 to REL2. Hence, with re-simulated cutoff distributions, there would be no natural order among REL1-REL4.

Table 6: Skips and Mistakes in Monte Carlo Simulations (Percentage Points)

	Scenarios: Data Generating Processes w/ Different Applicant Strategies							
	Truthful-Reporting Strategy	Payoff Irrelevant Skips			Payoff Relevant Mistakes			
	TRS	IRR1	IRR2	IRR3	REL1	REL2	REL3	REL4
WTT: *Weak Truth-Telling*[a]	100	85	69	53	52	52	52	52
Matched w/ favorite feasible college[b]	100	100	100	100	98	95	93	90
Skippers[c]	0	25	53	79	79	79	79	79
By number of skips:								
Skipping 11 colleges	0	17	37	55	67	70	73	74
Skipping 10 colleges	0	6	11	17	10	8	6	4
Skipping 9 colleges	0	2	4	6	1	1	0	0
Skipping 8 colleges	0	0	0	0	0	0	0	0
Skipping 7 colleges	0	0	0	0	0	0	0	0
Skipping 6 colleges	0	0	0	0	0	0	0	0
Skipping 5 colleges	0	0	0	0	0	0	0	0
Skipping 4 colleges	0	0	0	0	0	0	0	0
Skipping 3 colleges	0	0	0	0	0	0	0	0
Skipping 2 colleges	0	0	0	0	0	0	0	0
Skipping 1 college	0	0	0	0	0	0	0	0
TRS: *Truthful-Reporting Strategy*[d]	100	75	47	21	21	21	21	21
Reject WTT: Hausman Test[e]	5	10	60	100	97	95	91	87

Notes: This table presents the configurations of the eight data generating processes (DGPs). Each entry is a percentage averaged over the 200 simulation samples. In every sample, there are 1800 applicants competing for admissions to 12 colleges that have a total of 1500 seats. Tables C.4 and C.5 further show the breakdown by T_i. [a]An applicant is "*weakly truth-telling*" 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. [b]A college is feasible to an applicant, if the applicant's index (score) is higher than the college's ex-post admission cutoff. If an applicant is matched with her favorite feasible college, she cannot form a blocking pair with any college. [c]Given that every college is acceptable to all applicants and is potentially over-demanded, an applicant is a skipper if she does not rank all colleges. [d]An applicant adopts the "*truthful-reporting strategy*" if she truthfully ranks all available colleges. [e]In each DGP, this row reports the percentage of samples that the WTT (weakly truth-telling) assumption is rejected at 5% level in favor of the stability assumption. The test is based on the Durbin-Wu-Hausman test and discussed in detail in Section 4.3.

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

(ii) **Stability**. The assumption of stability implies that applicants are assigned their favorite feasible colleges 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, then stability holds in a large economy, according to Theorem 2. We define this estimator as the "stability" estimator.

(iii) **Robustness**. When there are payoff-relevant mistakes, some applicants may not be matched with their favorite feasible college. As a remedy, we propose a new approach called "robustness." We construct a hypothetical set of feasible colleges for each applicant by inflating the cutoffs of all but her matched college. An applicant's matched college is more likely to be her favorite in the hypothetical set of feasible colleges, because the set now contains fewer colleges. The estimation is similar to the stability estimator, except for the modified feasible sets. We call this the "robust" estimator.

The formulation of the likelihood functions of the three approaches are discussed in detail in

Appendix C, but with Table 6 we can evaluate the validity of WTT and stability already. Across the eight DGPs, the fraction of skippers increases from zero (in TRS) to 79 percent in IRR3 and remains at the same level in REL1-4. The WTT assumption is exactly satisfied only in TRS, and the fraction of applicants who are weakly truth-telling decreases from 85 percent in IRR1 to 53 percent in IRR3, stabilizing near 52 percent in all REL DGPs. In contrast, stability is always satisfied in TRS and IRR1-3, while the fraction of applicants that can form a blocking pair with some college increases from 2 percent in REL1 to 10 percent in REL4.

To test WTT against stability, we construct a Durbin-Wu-Hausman-type test statistic from the estimates of the WTT and stability approaches, following Fack, Grenet, and He (2017). Under the null hypothesis, both WTT and stability are satisfied, while under the alternative only stability holds. When all applicants, except the lowest-ranked 17 percent, are matched, WTT implies stability, but not the reverse. Therefore, if WTT is satisfied, the estimator based on WTT is consistent and efficient, while the stability estimator is consistent but inefficient.

The last row of Table 6 shows both the size and the power of this test. When the null is true (e.g., when DGP is TRS), the null is rejected at the desired rate, 5 percent. When the null is not true (in IRR1–3), the null is rejected with a 10–100 percent rate. Notice that when 47 percent of applicants (as in IRR3) are violating the WTT assumption, we already have a 100 percent rejection rate of the null hypothesis. In REL1-4, both WTT and stability are violated, while the latter is violated to a lesser extent. The test is no longer valid, although it still rejects the null at high rates.

4.4 Estimation Results

We now compare the performance of the three estimators based on three different identifying assumptions. Our main comparison is along two dimensions. One is the bias in the estimates of β_1. This coefficient measures the average quality difference between a pair of colleges $(c-1, c)$, because the utility function (equation 6) includes the term $\beta_1 \cdot c$. The other dimension is the comparison between the estimated and the true ordinal preferences. In particular, we calculate the estimated preference ordering of colleges 10 and 11.

4.4.1 Bias-Variance Tradeoff

Figure 1 plots the distributions of each estimator of β_1. Appendix C, especially Table C.6, provides more details. A consistent estimator should have mean 0.3. Recall that all DGPs use the same 200 simulated preference profiles and that what differs across DGPs is how applicants play the game.

It is evident in the figure that the WTT estimator always has a smaller variance than the other two. Intuitively, this is because WTT leads to more information being used for estimation.

Figure 1a presents the best-case scenario for the WTT estimator. That is, WTT is exactly satisfied and we use the maximum possible information (i.e., the complete ordinal preferences). As expected, the WTT estimator is consistent, as are those based on stability or robustness.

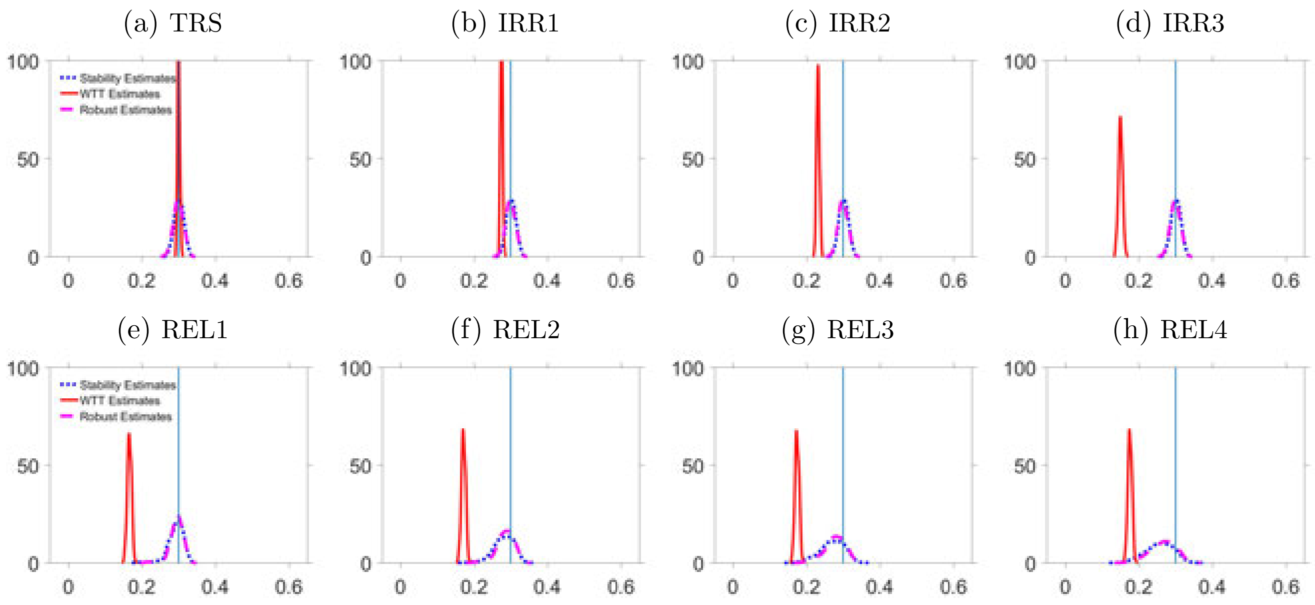


Figure 1: Estimates based on Weak Truth-Telling, Stability, or Robustness ($\beta_1 = 0.3$)

Notes: The figures focus on the estimates of the quality coefficient (β_1) from three approaches, weakly truth-telling (WTT, the red solid line), stability (the blue dotted line), and robustness (the purple dashed line). The distributions of the estimates across the 200 simulation samples are reported. A consistent estimator should have mean 0.3. Each subfigure uses the 200 estimates from the 200 simulation samples given a DGP and reports an estimated density based on a normal kernel function. Note that TRS as a DGP means that every applicant truthfully ranks all colleges; IRR1-3 only include payoff irrelevant skips, while REL1-4 have payoff relevant mistakes. See Table 6 for more details on the eight DGPs.

The results from the data containing payoff-irrelevant skips are summarized in Figure 1b–d. As expected, the stability estimates (the blue dotted line) and the robust estimates (the purple dashed line) are invariant to payoff-irrelevant skips and stay the same as that those in the TRS DGP. In contrast, the WTT estimates (the red solid line) are sensitive to the fraction of skippers. Even when 25 percent of applicants are skippers and 15 percent are violating the WTT assumption (IRR1), the estimates based on WTT from the 200 samples have mean 0.27 (standard deviation 0.00); in contrast, the estimates from the other two approaches are on average 0.30 (standard deviation 0.01).

The downward bias in the WTT estimator is intuitive. When applicants skip, they omit colleges with which they have almost no chance of being matched. For applicants with low priorities, popular colleges are therefore more likely to be skipped. Whenever a college is skipped, WTT assumes that it is less preferable than all the ranked colleges. Therefore, many applicants are mistakenly assumed to dislike popular colleges, which results in a downward bias in the estimator of β_1. In contrast, this bias is absent in the stability and robust estimators: whenever a college is skipped by an applicant due to its high cutoff, neither the stability nor the robust approach assumes the college is less preferable than ranked colleges.

Figures 1e–h show the DGPs in which applicants make payoff-relevant mistakes. Neither WTT nor stability is satisfied (Table 6), and therefore both estimators are inconsistent. However, the stability estimator is still less biased; the means of the estimates are close to the true value, ranging from 0.26 to 0.29 (see Table C.6 for more details). In contrast, the means of the WTT estimates are between 0.17–0.18. As predicted, the robust estimator can tolerate some payoff-relevant mistakes and results in less-biased estimates, with means between 0.27–0.29.

4.4.2 Mis-Estimated Preferences

A direct consequence of an inconsistent estimator is the mis-estimation of applicant preferences. Let us consider colleges 10 and 11. The latter is a small college as well as a special college for disadvantaged applicants, while the former is neither. For a disadvantaged applicant ($T_i = 1$) with an equal distance to these two colleges, the probability that she prefers college 11 to college 10 is $\frac{\exp(11\beta_1+\beta_3+\beta_4)}{\exp(10\beta_1)+\exp(11\beta_1+\beta_3+\beta_4)}$. Inserting the true values, $(\beta_1, \beta_3, \beta_4) = (0.3, 2, 0)$, we find that the probability is 0.91. This is depicted by the straight line in Figure 2. With the same formula, we calculate the same probability based on the three sets of estimates, and Figure 2 presents the average estimated probability from each set of estimates across the 200 samples in each DGP.

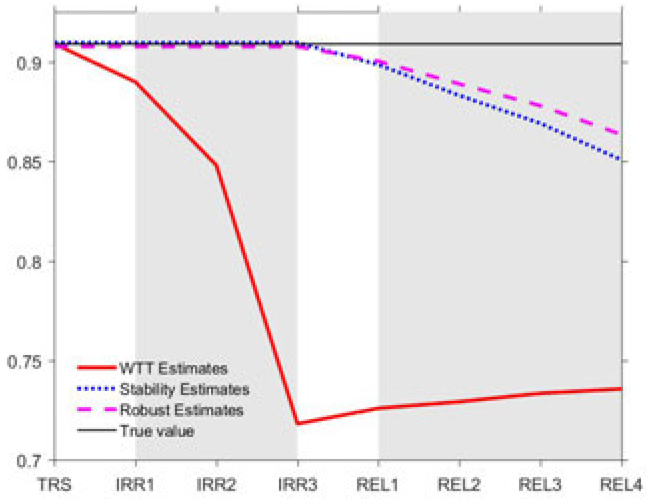


Figure 2: True and Estimated Probabilities That an Applicant Prefers College 11 to College 10

Notes: The figure presents the probability that a disadvantaged student ($T_i = 1$), with an equal distance to both colleges, prefers college 11 to college 10. The true value is 0.91 (the thin solid line), calculated as $\frac{\exp(11\beta_1+\beta_3+\beta_4)}{\exp(10\beta_1)+\exp(11\beta_1+\beta_3+\beta_4)}$. With the same formula, we calculate the estimations based on the WTT estimates, and the thick solid line presents the average over the 200 simulation samples in each DGP. Similarly, the dotted line describes those based on the stability estimates; and the dashed line depicts the average estimated probabilities based on the robust estimates.

When the DGP is TRS, all three identifying assumptions lead to consistent estimators, and the three estimated probabilities almost coincide with the true value. In IRR1-3, the stability and robust estimators are still consistent, but the estimated probabilities based on the WTT estimates (the thick solid line) deviate from the true value significantly. In IRR3 especially, the WTT estimates result in a mis-estimation of the ordinal preferences of 20% of the applicants. In REL1-4, the estimations based on the stability estimates (the dotted line) are still relatively close to the true value, despite being inconsistent. Moreover, those based on the robust estimates (the dashed line) are even less biased. The mis-estimation of preferences has direct consequences when one evaluates counterfactual policies, which we investigate in the next subsection.

4.5 Counterfactual Analysis

Making policy recommendations based on counterfactual analysis is one of the main objectives of market design research. In the following, we illustrate how the common approaches lead to mis-predicted counterfactual outcomes, while the estimations based on stability and robustness yield results close to the truth.

We consider the following counterfactual policy: applicants with $T_i = 1$ are given priority over those with $T_i = 0$, 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 $i > i'$. One may consider this as an affirmative action policy if $T_i = 1$ indicates i being disadvantaged. 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 benchmarks against which alternative methods will be evaluated. In keeping with our DGPs above, the "actual behavior" ranges from truthful reporting to varying degrees of skipping (see Section 4.2). Specifically, DGP TRS requires everyone to submit a truthful 12-college ROL; in DGPs IRR1-3, the potential skippers omit their never-matched colleges; and in DGPs REL1-4, the skippers additionally omit some colleges with which they have some 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 condition. Under the counterfactual policy, we simulate applicant preferences based on the estimates and let applicants submit truthful 12-college lists.

(iv) **Stability**: We estimate applicant preferences from the data with stability as the identifying condition. Under the counterfactual policy, we simulate applicant preferences based on the estimates and let applicants submit truthful 12-college lists.

(v) **Robustness**: We estimate applicant preferences from the data with the robust approach. Under the counterfactual policy, we simulate applicant preferences based on the estimates and let applicants submit truthful 12-college lists.

Note that we assume truthful reporting in the counterfactual in the last three approaches. This is necessary because none of these approaches estimates how applicants make mistakes, while we have to specify applicant behavior in counterfactual analysis. This assumption of truthful reporting in the counterfactual analysis is justified by Corollary 2.[32]

When simulating counterfactual outcomes, we use the same 200 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 200 simulation samples, we conduct 40 different counterfactual analyses: 8 (DGPs: TRS, IRR1-3, and REL1-4) × 5 (actual behavior and 4 counterfactual approaches—submitted ROLs, WTT, stability, and robustness).

4.5.1 Performance of the Four Approaches in Counterfactual Analysis

We first simulate the true outcomes under the counterfactual policy with the actual behavior (i.e., the true preferences with possible mistakes).[33] When doing so, we assume applicants make mistakes as they do under the current policy. As shown above, the matching outcome does not change if applicants make payoff-irrelevant skips, although payoff-relevant mistakes would lead to different outcomes.

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

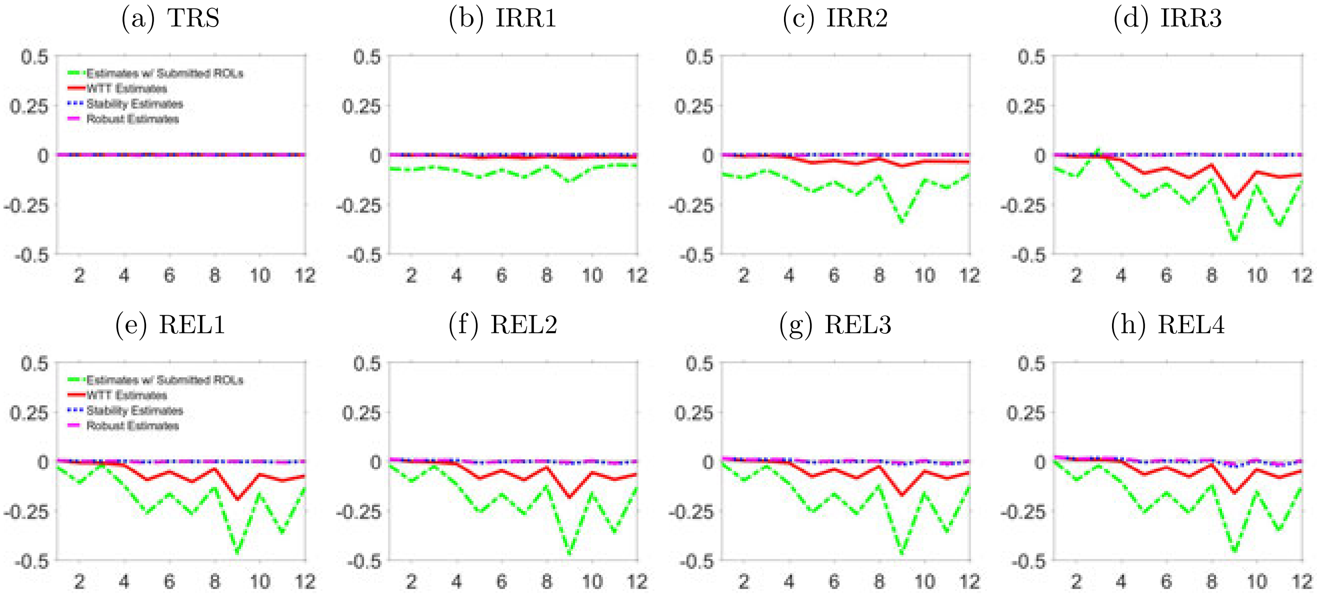
An informative statistic of a matching is the college cutoffs which summarize the joint distribution of applicant priorities and preferences. Figure 3 shows, given each DGP, how the four approaches mis-predict the cutoffs under the counterfactual policy. For each college, indexed from 1 to 12, we calculate the mean of the 200 cutoffs from the 200 simulation samples by using the actual behavior and the other four approaches. The sub-figures then depict the mean differences between the predicted cutoffs and the true ones from the actual behavior.

In Figure 3a, the DGP is TRS, 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 three approaches also lead to almost the same cutoffs.

In Figures 3b–d, corresponding to DGPs IRR1-3, only the stability and robust estimators are consistent, and indeed these estimators have the smallest mis-predictions relative to the other two. Both of the estimates based on WTT and submitted ROLs have mis-predictions increasing from IRR1 to IRR3, and those based on submitted ROLs result in larger biases. Since applicants tend to omit popular colleges from their lists, both approaches underestimate the demand for these colleges and thus result in under-predicted cutoffs. The bias is even larger for smaller colleges because they tend to be skipped more often.

[32] 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.

[33] The following results from the actual behavior, especially Figure 5 and Tables C.7 and C.8, may seem constant across DGPs REL1–4. Indeed, this variation in matching outcomes is small as predicted by our Theorem 2, and it sometimes disappears because of rounding.

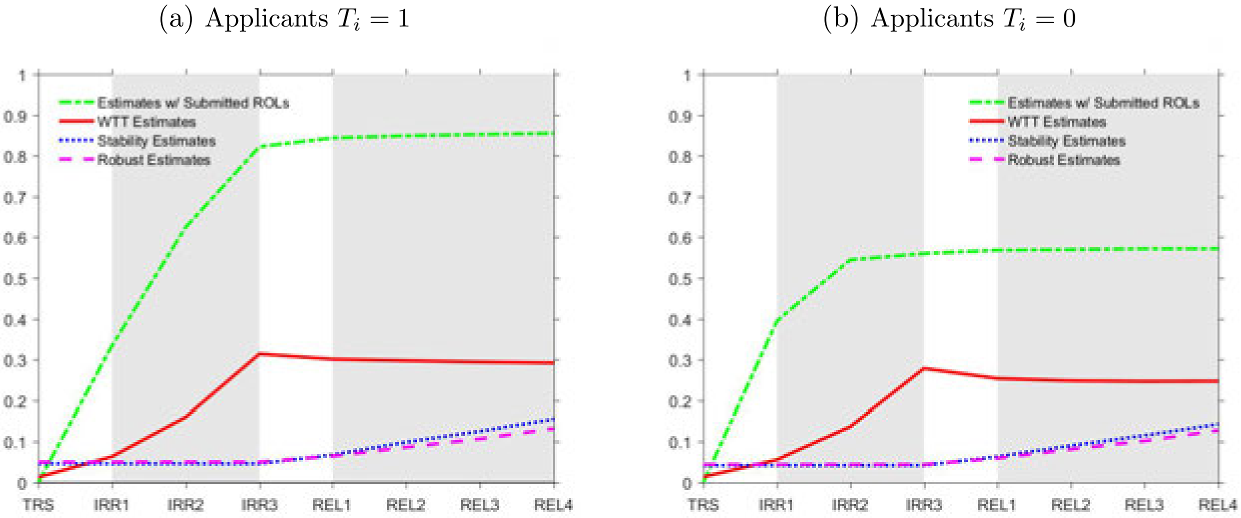


(a) TRS (b) IRR1 (c) IRR2 (d) IRR3 (e) REL1 (f) REL2 (g) REL3 (h) REL4

Figure 3: Comparison of the Four Approaches: Biases in Predicted Cutoffs

Notes: The sub-figures present 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). Each subfigure corresponds to a DGP. Given a DGP, we simulate the colleges' cutoffs following each approach and calculate the mean deviation from the true ones. The X-axis shows the college indices; the Y-axis indicates the deviation of the predicted cutoffs from the true ones.

When the DGPs contain payoff-relevant mistakes (REL1-4), none of the approaches is consistent (Figures 3e–h). However, the stability and robust estimates seem to have the negligible mis-prediction compared to the other two.



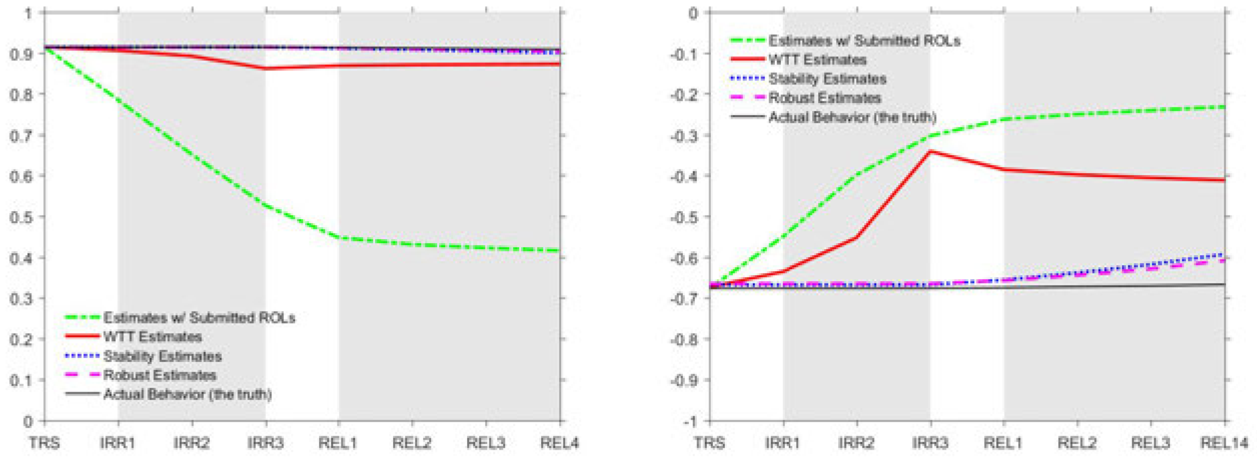
(a) Applicants $T_i = 1$ (b) Applicants $T_i = 0$

Figure 4: Comparison of the Four Approaches: Mis-predicted Match (Fractions)

Notes: The sub-figures show how each approach to counterfactual analysis mis-predicts matching outcomes under the counterfactual policy. Given a DGP, we simulate a matching outcome and compare the result to the truth which is calculated based on the actual behavior (i.e., the true preferences with possible mistakes). The sub-figures present the average rates of mis-prediction for the two groups of applicants, $T_i = 1$ and $T_i = 0$, across the 200 samples in each DGP. On average, there are 599 (1201) applicants with $T_i = 0$ ($T_i = 1$) in a simulation sample.

Figure 4 further shows how each of the four approaches mis-predicts individual outcomes. Because the counterfactual policy is intended to help applicants with $T_i = 1$, we look at two groups, $T_i = 1$ or 0, separately.

In Figure 4a, among the $T_i = 1$ applicants, the stability approach incorrectly predicts a match for 5 percent of applicants on average whenever stability is satisfied (in DGPs TRS and IRR1-3). Among REL1-4, the fraction of mis-prediction based on stability increases from 6 to 13 percent. The WTT approach has a lower mis-prediction rate in TRS but under-performs relative to stability in all other DGPs. The submitted-ROLs approach has the highest mis-prediction rates in all DGPs except TRS. Lastly, the robust estimates are almost identical to the stability estimates in TRS and IRR1-3 but perform better in REL1-4. Among the applicants with $T_i = 0$ (subfigure b), the comparison of the four approaches follows the same pattern.



(a) Applicants $T_i = 1$: (Fraction Better off)$-$(Fraction Worse off) (b) Applicants $T_i = 0$: (Fraction Better off)$-$(Fraction Worse off)

Figure 5: Comparison of the Four Approaches: Mis-predicting Welfare Effects

Notes: The sub-figures show how each approach to counterfactual analysis mis-predicts the welfare effects of the counterfactual policy for the two groups of applicants, $T_i = 1$ and $T_i = 0$. Given a DGP, we simulate matching outcomes, calculate welfare effects, and compare them to the truth that is calculated based on the actual behavior (i.e., the true preferences with possible mistakes). Welfare effects are measured by the difference between the fraction of applicants better off and that of those worse off, averaged over the 200 samples in each DGP. On average, there are 599 (1201) applicants with $T_i = 0$ ($T_i = 1$) in a simulation sample. The estimated fraction of applicants with $T_i = 1$ being worse off is close to zero in all cases, as is the estimated fraction of applicants with $T_i = 0$ being better off. There are some applicants whose welfare does not change; simulated with true preferences, this fraction is 9 percent among the $T_i = 1$ applicants and 32 percent among the $T_i = 0$ applicants. See Tables C.7 and C.8 in Appendix C for more details.

We now investigate the welfare effects on the $T_i = 1$ applicants and others when the current policy is replaced by the counterfactual one. 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 better off; she is worse off if she is matched with a "less-preferred" one. Because each approach to counterfactual analysis estimates applicant preferences in a unique way, an applicant's utility associated with a given college is estimated at a different value under each approach. Therefore, the measured welfare effects of the counterfactual policy may differ even when an applicant is matched with the same college.

Figure 5 shows the difference between the fraction of applicants better off and that of those worse off, averaged across the 200 simulation samples.[34] In Figure 5a, among the $T_i = 1$ applicants, the predictions based on the stability or robust estimates are almost identical to the true value, even in the cases with payoff-relevant mistakes (REL1-4). In contrast, the WTT approach is close to the true value only in DGPs TRS and IRR 1; those based on submitted ROLs tend to be biased toward zero effect when there are more applicants skipping or making mistakes. The reason for the biases is clear. Under WTT, the quality of popular colleges are underestimated; this leads to understatement of the benefits from affirmative action for the disadvantaged applicants. Meanwhile, the submitted-ROLs method fails to account for the likely changes in ROLs applicants submit under the affirmative action policy: for instance, disadvantaged applicants find once out-of-reach colleges now within reach, so they would start listing them in their ROLs.

The results for applicants with $T_i = 0$ are collected in Figure 5b. The general patterns remain the same, although the stability and robust estimates are more biased in REL1-4 than in Figure 5a.

In summary, the estimated welfare effects are biased toward zero for all applicants when we assume WTT or take submitted ROLs as true preferences. The stability estimates, however, are very close to the true value; even when there are some payoff-relevant mistakes, the stability estimates are much less biased than those of the other two approaches. The robust approach further improves upon the stability estimates when there are payoff-relevant mistakes. This suggests that the estimators on which the counterfactual methods build as well as the assumption of truthful reporting in the counterfactual scenario (justified by Corollary 2) work well.

5 Conclusion

Our analysis of the Australian college admissions data suggests that applicants choose not to apply to some college programs against their apparent interests, when doing so is unlikely to affect the outcome. Motivated by this evidence, we have argued theoretically using a robust equilibrium concept that an *outcome* may be more reliably predicted than *behavior*, even when participants act in a strategically straightforward environment, such as when students apply for colleges in a deferred acceptance mechanism. While this result justifies the vast theoretical literature that assumes truthful reporting behavior to analyze strategy-proof mechanisms, it calls into question any empirical method that takes truthful reporting as a literal behavioral prediction. An alternative approach focusing on the stability property of the outcome proves more robust.

Our Monte Carlo analysis indeed reveals that the empirical method based on the truth-telling assumption leads to a biased estimator of preferences when applicants make the types of mistakes consistent with our empirical evidence and our theoretical model. By contrast but in keeping with our theory, an empirical method based on the stability condition proves more robust to these mistakes. We further show that a counterfactual analysis based on stability yields more accurate

[34] There are some applicants whose outcomes do not change. See Tables C.7 and C.8 in Appendix C for more detailed summary statistics.

predictions than the one based on truth-telling.

One may worry that the outcome-based method "throws away" some information on revealed preferences of participants—albeit deemed unreliable in the presence of mistakes—and that the resulting estimation would entail efficiency loss *if* individuals make *no* mistakes. Not knowing a priori how mistake-prone actual participants are, the choice of an empirical method seems challenging. However, a statistical test similar to the Durbin-Wu-Hausman test can reveal the significance of the potential bias that would result from the use of the truth-telling assumption, and thus could provide a data-driven guide toward the appropriate method. Indeed, our simulation results show that the test has the correct size and reasonable statistical power when being used to choose between the two approaches.

References

Abdulkadiroglu, A., N. Agarwal, and P. A. Pathak (Forthcoming): "The Welfare Effects of Coordinated Assignment: Evidence from the NYC HS Match," *American Economic Review*.

Abdulkadiroglu, A., Y.-K. Che, and Y. Yasuda (2015): "Expanding 'Choice' in School Choice," *American Economic Journal: Microeconomics*, 7, 1–42.

Abdulkadiroglu, A., P. Pathak, A. E. Roth, and T. Sonmez (2006): "Changing the Boston school choice mechanism," Discussion paper, National Bureau of Economic Research.

Abdulkadiroglu, 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.

Agarwal, N. (2015): "An empirical model of the medical match," *American Economic Review*, 105(7), 1939–1978.

Agarwal, N., and P. Somaini (Forthcoming): "Demand Analysis using Strategic Reports: An Application to a School Choice Mechanism," *Econometrica*.

Ashlagi, I., and Y. Gonczarowski (2016): "Stable Mechanisms Are Not Oviously Strategy-proof," .

Ashlagi, I., Y. Kanoria, and J. D. Leshno (forthcoming): "Unbalanced Random Matching Markets," *Journal of Political Economy*.

Azevedo, E. M., and E. Budish (2015): "Strategy-proofness in the Large," *Unpublished mimeo, University of Chicago and University of Pennsylvania*.

Azevedo, E. M., and J. W. Hatfield (2012): "Complementarity and Multidimensional Heterogeneity in Matching Markets," mimeo.

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.

Calsamiglia, C., C. Fu, and M. Güell (2014): "Structural Estimation of a Model of School Choices: the Boston Mechanism vs. Its Alternatives," Barcelona GSE Working Paper No. 811.

Che, Y.-K., J. Kim, and F. Kojima (2013): "Stable Matching in Large Economies," mimeo.

Che, Y.-K., and Y. Koh (2016): "Decentralized College Admissions," *JPE*, 124(5), 1295–1338.

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 (2015a): "Efficiency and Stability in Large Matching Markets," Columbia University and PSE, Unpublished mimeo.

——— (2015b): "Payoff Equivalence of Efficient Mechanisms in Large Markets," Columbia University and PSE, Unpublished mimeo.

Chen, L., and J. S. Pereyra (2015): "Self-selection in school choice," .

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," Manuscript.

Deb, J., and E. Kalai (2015): "Stability in Large Bayesian Games with Heterogeneous Players," *Journal of Economic Theory*, 157, 1041–1055.

Drewes, T., and C. Michael (2006): "How do Students Choose a University?: An Analysis of Applications to Universities in Ontario, Canada," *Research in Higher Education*, 47(7), 781–800.

Dur, U., R. G. Hammond, and T. Morrill (Forthcoming): "Identifying the harm of manipulable school-choice mechanisms," Discussion paper.

Fack, G., J. Grenet, and Y. He (2017): "Beyond Truth-Telling: Preference Estimation with Centralized School Choice and College Admissions," Paris School of Economics and Rice University, Unpublished mimeo.

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.

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. Shorrer (2016): ""Strategic" Behavior in a Strategy-proof Environment," mimeo.

He, Y. (2017): "Gaming the Boston School Choice Mechanism in Beijing," Toulouse School of Economics and Rice University, Unpublished mimeo.

Hwang, S. I. M. (2017): "How does heterogeneity in beliefs affect students in the Boston Mechanism," *Working Paper.*

Immorlica, N., and M. Mahdian (2005): "Marriage, Honesty, and Stability," *SODA 2005*, pp. 53–62.

Kapor, A., C. Neilson, and S. Zimmerman (2016): "Heterogeneous Beliefs and School Choice," .

Kirkebøen, L. J. (2012): "Preferences for Lifetime Earnings, Earnings Risk and Monpecuniary Attributes in Choice of Higher Education," Statistics Norway Discussion Papers No. 725.

Kojima, F., and P. A. Pathak (2009): "Incentives and Stability in Large Two-Sided Matching Markets," *American Economic Review*, 99, 608–627.

Lee, S. (2014): "Incentive Compatibility of Large Centralized Matching Markets," University of Pennsylvania, Unpublished mimeo.

Lee, S., and L. Yariv (2014): "On the Efficiency of Stable Matchings in Large Markets," University of Pennsylvania, Unpublished mimeo.

Li, S. (2017): "Obviously Strategy-Proof Mechanisms," Harvard, Unpublished mimeo.

Liu, Q., and M. Pycia (2011): "Ordinal Efficiency, Fairness, and Incentives in Large Markets," *mimeo.*

McDiarmid, C. (1989): "On the method of bounded differences," *Surveys in Combinatorics*, pp. 148–188.

Pathak, P. (2011): "The mechanism design approach to student assignment," *Annual Review of Economics*, 3(1).

Pittel, B. (1989): "The Average Number of Stable Matchings," *SIAM Journal on Discrete Mathematics*, 2, 530–549.

Pycia, M., and P. Troyan (2016): "Obvious Dominance and Random Priority," UCLA, Unpublished mimeo.

Rees-Jones, A. (2017): "Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match," *Games and Economic Behavior.*

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, 748–780.

Shorrer, R. I., and S. Sóvágó (2017): "Obvious Mistakes in a Strategically Simple College-Admissions Environment," Discussion paper, Working paper.

Sun, Y. (2006): "The exact law of large numbers via Fubini extension and characterization of insurable risks," *Journal of Economic Theory*, 126, 31–69.

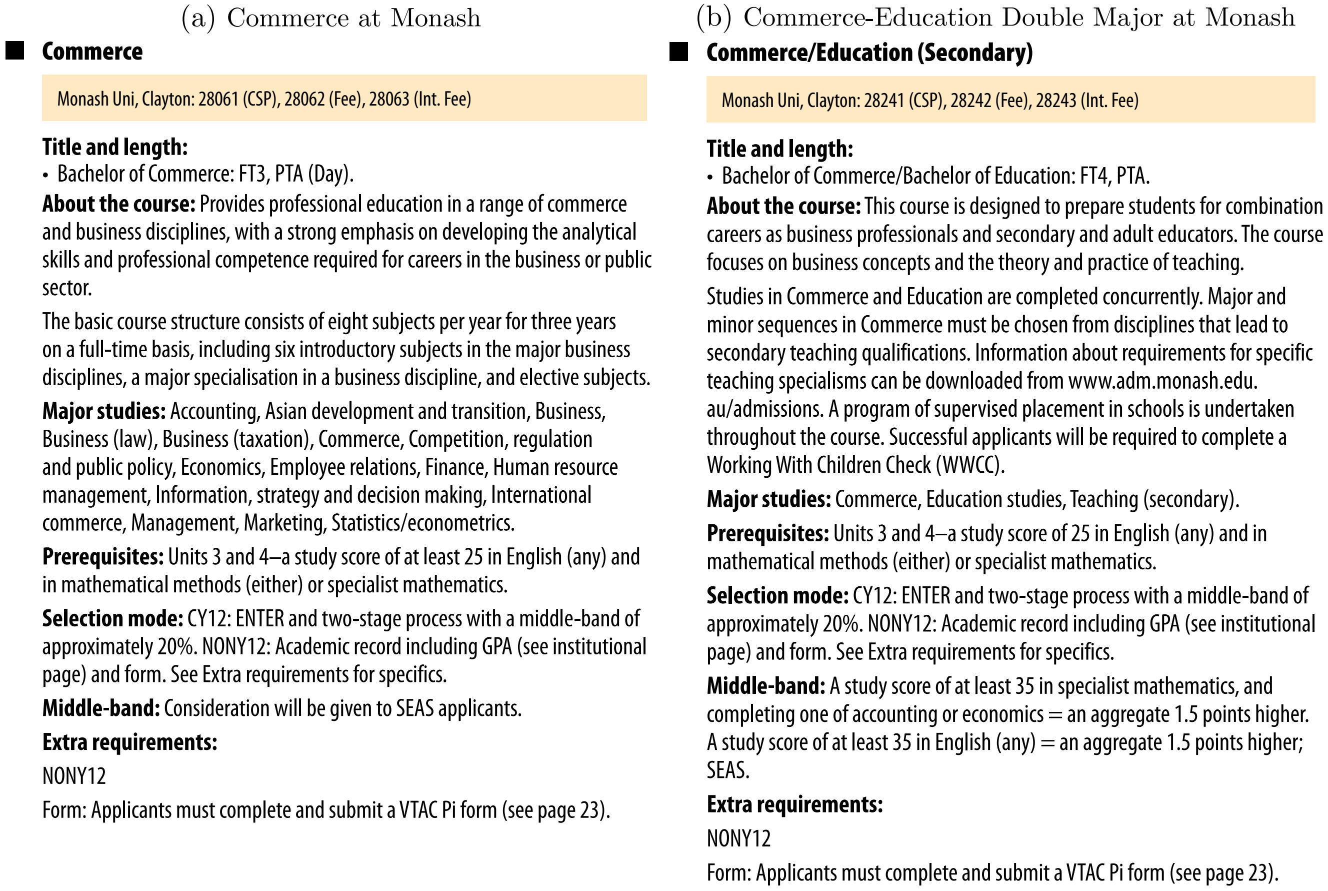
Veski, A., P. Biró, K. Poder, and T. Lauri (2016): "Efficiency and Fair Access in Kindergarten Allocation Policy Design," .

A Additional Details on Victorian Tertiary Admission

A.1 Courses

Victorian clearinghouse, Victorian Tertiary Admissions Centre (VTAC), processes applications both for undergraduate and for technical and further education (TAFE) courses. Undergraduate courses include bachelor degrees as well as a variety of diplomas and certificates. An applicant lists all types of those courses in a single ROL.

Two undergraduate course descriptions from the VTAC Guide, the main publication of the clearinghouse, are given in Figure A.1.



(a) Commerce at Monash

■ **Commerce**

Monash Uni, Clayton: 28061 (CSP), 28062 (Fee), 28063 (Int. Fee)

Title and length:

- Bachelor of Commerce: FT3, PTA (Day).

About the course: Provides professional education in a range of commerce and business disciplines, with a strong emphasis on developing the analytical skills and professional competence required for careers in the business or public sector.

The basic course structure consists of eight subjects per year for three years on a full-time basis, including six introductory subjects in the major business disciplines, a major specialisation in a business discipline, and elective subjects.

Major studies: Accounting, Asian development and transition, Business, Business (law), Business (taxation), Commerce, Competition, regulation and public policy, Economics, Employee relations, Finance, Human resource management, Information, strategy and decision making, International commerce, Management, Marketing, Statistics/econometrics.

Prerequisites: Units 3 and 4–a study score of at least 25 in English (any) and in mathematical methods (either) or specialist mathematics.

Selection mode: CY12: ENTER and two-stage process with a middle-band of approximately 20%. NONY12: Academic record including GPA (see institutional page) and form. See Extra requirements for specifics.

Middle-band: Consideration will be given to SEAS applicants.

Extra requirements:

NONY12

Form: Applicants must complete and submit a VTAC Pi form (see page 23).

(b) Commerce-Education Double Major at Monash

■ **Commerce/Education (Secondary)**

Monash Uni, Clayton: 28241 (CSP), 28242 (Fee), 28243 (Int. Fee)

Title and length:

- Bachelor of Commerce/Bachelor of Education: FT4, PTA.

About the course: This course is designed to prepare students for combination careers as business professionals and secondary and adult educators. The course focuses on business concepts and the theory and practice of teaching.

Studies in Commerce and Education are completed concurrently. Major and minor sequences in Commerce must be chosen from disciplines that lead to secondary teaching qualifications. Information about requirements for specific teaching specialisms can be downloaded from www.adm.monash.edu.au/admissions. A program of supervised placement in schools is undertaken throughout the course. Successful applicants will be required to complete a Working With Children Check (WWCC).

Major studies: Commerce, Education studies, Teaching (secondary).

Prerequisites: Units 3 and 4–a study score of 25 in English (any) and in mathematical methods (either) or specialist mathematics.

Selection mode: CY12: ENTER and two-stage process with a middle-band of approximately 20%. NONY12: Academic record including GPA (see institutional page) and form. See Extra requirements for specifics.

Middle-band: A study score of at least 35 in specialist mathematics, and completing one of accounting or economics = an aggregate 1.5 points higher. A study score of at least 35 in English (any) = an aggregate 1.5 points higher; SEAS.

Extra requirements:

NONY12

Form: Applicants must complete and submit a VTAC Pi form (see page 23).

Figure A.1: Examples of Course Description

Notes: These screen shots are from the 2008 VTAC Guide.

We treat two courses as offering the same program and differing only by the fee if the course code shares the first four digits (which implies that the courses also share the description given above). The course above is offered in three varieties: as a CSP, or reduced-fee, course; as a full-fee course; and as a course for international applicants. We are interested in the first two varieties. Although the majority of the applicants are ranked by the course according to ENTER, this course rank 20% of its applicants (see "Selection mode") based on the performance in specific courses listed in the "Middle-band" section. It also re-ranks the affirmative-action (SEAS) applicants. Most courses re-rank affirmative action applicants; whether a course re-ranks the applicants based on other criteria varies, and the criteria may be less specific. A small number of courses, such as those in performing arts, require a portfolio, an audition, or an interview. CY12 refers to current year 12 applicants (the focus of our study, category V16 applicants, is part of CY12) and NONY12 refers to non-year 12 applicants.

To determine a cutoff of a course, we select bottom 5% of accepted applicants and top 5% of rejected applicants and then take a median ENTER of applicants in this selection. Usually, the number of rejected applicants is about twice as large as the number of accepted applicants. Thus, such a selection over-weights rejected applicants. Furthermore, we do not observe special consideration applicants; the bottom 5% of accepted applicants are more likely to come from this pool. Overall, these applicants often have lower scores than the top 5% rejected. Only cutoffs of reduced-fee courses are used in the paper, as the number of accepted and rejected applicants for full-fee courses is too small.

A.2 Applicants

There are multiple categories of applicants; we focus on the category "V16", who are the most typical high school graduate in Victoria. They follow the standard state curriculum and do not have any tertiary course credits to claim. Two other categories are also of interest to us: "V14" and "V22". The former are Victorian applicants who follow International Baccalaureate curriculum, while the latter are interstate applicants. These two categories must be evaluated in the same way as V16 by the admission officers. We use them to derive course cutoffs more precisely. We do not use them in the estimations because they miss some control variables that we use. Table A.1 gives relative frequencies of these categories of applicants. Among V16 applicants, 27,922 fill less than 12 courses; they form our main sample.

Table A.1: Categories of Applicants

	All	CY12			NONY12
		V16	V14 and V22	Other CY12	
Total	74,704	37,266	4,103	9,275	24,060
% of Total	100.00	49.89	5.49	12.42	32.21

Applicants have easy access to the following information: Clearly-in ENTER, Fringe ENTER, Percentage of Offers Below Clearly-in ENTER (all three are available for Round 1 and Final Round), as well as Final Number of Offers (CY12 and Total). Clearly-in ENTER refers to the cutoff above which every eligible applicant must be admitted; Fringe ENTER refers to 5% percentile of the scores of admitted applicants. Note that this cutoff statistic does not distinguish between CY12 and NONY12 applicants; for that reason, we do not use any of these cutoffs to determine payoff-relevance of mistakes.

Even if an applicant skips a feasible RF course (that is, applicant's ENTER is above the course cutoff), the skipping mistake may not be payoff relevant: the applicant may have been admitted to a course that the applicant prefers. Thus, to determine whether the mistake is payoff-relevant, we need to complete the ROL of an applicant by adding back the skipped RF course. We report the lower and upper bounds on the number of payoff relevant mistakes.

For the lower bound, we assume that skipped RF course is the least desirable among acceptable RF courses. Specifically, a skip is considered payoff-relevant if (i) an applicant does not receive an offer from any RF course; (ii) receives an offer from an FF course; (iii) does not list the RF course corresponding to the FF course being offered (the course with the same code except for the last digit); and (iv) the RF course is feasible for this applicant. For the upper bound, we assume that skipped RF course is the most desirable. Specifically, a skip is considered payoff-relevant if (i) an applicant lists FF course; (ii) does not list a corresponding RF course; and (iii) the RF course is feasible.

The summary statistics for all applicants, applicants with skips and applicants with payoff relevant mistakes are presented in Table A.2.

Table A.2: Summary statistics for applicants by their mistake status

	All	w/Skip	w/Mistake
ENTER	65.77	61.84	78.21
GAT	61.94	59.03	66.44
Female	0.57	0.53	0.59
ln(income)	0.0024	0.0036	-0.0098
Citizen	0.98	0.95	0.94
Perm. resident	0.02	0.04	0.05
Born in			
Australia	0.90	0.85	0.86
Southern and Central Asia	0.01	0.04	0.06
Language spoken at home			
English	0.91	0.86	0.88
Eastern Asian Languages	0.02	0.04	0.06
Number of FF courses in ROL	0.22	2.40	2.93
Attends high school with tuition fees >AUD9000	0.16	0.27	0.34
Total	27,922	1,009	201

Notes: "All" refers to all V16 applicants who list fewer than 12 courses. "Mistake" refers to a payoff-relevant mistake. Numbers for citizenship status, country born and language spoken at home do not sum up to 100 as some entries have been omitted.

A.3 Expected ENTER

With the information on GAT, we predict ENTER using the following model, which is a second-order polynomial in the three parts of GAT:

$$\begin{aligned}
ENTER_i^* =& a_0 \\
&+ a_{11}\ GAT1_i + a_{12}\ GAT1_i^2 + a_{13}\ GAT1_i^3 \\
&+ a_{21}\ GAT2_i + a_{22}\ GAT2_i^2 + a_{23}\ GAT2_i^3 \\
&+ a_{31}\ GAT3_i + a_{32}\ GAT3_i^2 + a_{33}\ GAT1_3^3 \\
&+ a_{1\times 2}\ GAT1_i \times GAT2_i + a_{1\times 3}\ GAT1_i \times GAT3_i + a_{2\times 3}\ GAT2_i \times GAT3_i \\
&+ a_{12\times 2}\ GAT1_i^2 \times GAT2_i + a_{12\times 3}\ GAT1_i^2 \times GAT3_i \\
&+ a_{1\times 22}\ GAT1_i \times GAT2_i^2 + a_{1\times 32}\ GAT1_i^1 \times GAT3_i^2 + a_{22\times 3}\ GAT2_i^2 \times GAT3_i \\
&+ a_{1\times 32}\ GAT1_i \times GAT3_i^2 \\
&+ \epsilon_i, \qquad \text{(A.1)}
\end{aligned}$$

where $GAT1, GAT2$ and $GAT3$ are the results of three parts of GAT test (written communication; mathematics, science and technology; humanities, the arts and social sciences). Because ENTER is always in $(0, 100)$, we apply a tobit model to take into account the lower and upper bounds. In effect, we assume $\epsilon_i \sim N(0, \sigma^2)$. The estimated coefficients from the Tobit model are reported in Table A.3.

Table A.3: Estimation of the Model for Predicting ENTER

Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
$GAT1$	-2.48*** (0.17)	$GAT1^2 \times GAT2$	-0.00** (0.00)	$GAT2^2$	0.17*** (0.01)	$GAT3$	-1.32*** (0.28)
$GAT1^2$	0.09*** (0.01)	$GAT1^2 \times GAT3$	-0.00 (0.00)	$GAT2^3$	-0.00*** (0.00)	$GAT3^2$	0.07*** (0.01)
$GAT1^3$	-0.00*** (0.00)	$GAT1 \times GAT2^2$	-0.00** (0.00)	$GAT2 \times GAT3$	-0.04*** (0.01)	$GAT3^3$	-0.00*** (0.00)
$GAT1 \times GAT2$	0.03** (0.01)	$GAT1 \times GAT3^2$	-0.00*** (0.00)	$GAT2^2 \times GAT3$	-0.00*** (0.00)	Constant	57.41*** (0.85)
$GAT1 \times GAT3$	0.06*** (0.01)	$GAT2$	-3.30*** (0.27)	$GAT2 \times GAT3^2$	0.00*** (0.00)	σ	13.73*** (0.05)
N	37,221	Pseudo R^2	0.10				

Notes: This table reports the estimation results of the Tobit model (Equation A.1). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

With the estimated coefficients, we generate an expected/predicted ENTER for every applicant. The coefficient of correlation between the real ENTER and the expected ENTER is 0.7745.

B Proofs from Section 3

Proof of Theorem 1. Recall the cardinal type $\theta = (u, s)$ induces an ordinal preference type $(\rho(u), s)$. Recall that the strategy $\hat{R}(\theta)$ depends on u only through $\hat{\rho}(u)$, without loss we can work in terms of the "projected" ordinal type (ρ, s). Also recall that the mechanism depends only on an applicant's ROL and her score. Hence, for the current proof, we shall abuse the notation and call $\theta := (\rho, s)$ an applicant's type, redefine the type space $\Theta := \mathcal{R} \times [0,1]^C$ (the projection of the original types), and let η be the measure of the projected types (which is induced by the original measure on (u, s)). The continuum economy $E = [\eta, S]$ is redefined in this way. Likewise, the k-random economies $F^k = [\eta^k, S]$ are similarly redefined. Given this reformulation, it suffices to show that it is a robust equilibrium for each type $\theta \in \Theta^\delta$ to adopt TRS and for each type $\theta \notin \Theta^\delta$ to randomize between TRS with probability γ and $\hat{r}(\theta)$ with probability $1 - \gamma$.

We first make the following preliminary observations.

Claim 1. *Given the continuum economy $E = [\eta, S]$, let $\hat{\eta}$ denote the measure of "reported" types when the applicants follow the prescribed strategies, and let $\hat{E} = [\hat{\eta}, S]$ denote the "induced" continuum economy under that strategies. Then, $\hat{E}$ has the unique stable matching identical to that under E, characterized by the identical cutoffs P^*. The demand under that economy $\hat{D}(\cdot)$ is C^1 in the neighborhood of P^* and has $\partial\hat{D}(P^*) = \partial D(P^*)$, which is invertible.*

Proof. Let $D^{(\rho,s)}(P^*)$ be the college an applicant with type $\theta = (\rho, s)$ demands given cutoffs P^* (i.e., her most preferred feasible college given P^*). Since $\hat{r}$ ranks her favorite feasible college ahead of all other feasible colleges, it must be that $D^{(\hat{r}(\rho,s),s)}(P^*) = D^{(\rho,s)}(P^*)$ for each type (u, s). It then follows that $\hat{D}(P^*) = D(P^*)$, where $\hat{D}(P)$ is the demand at the continuum economy $\hat{E} = [\hat{\eta}, S]$. Hence, P^* also characterizes a stable matching in $\hat{E}$. Further, since η has full support and since the prescribed strategy has every type θ play TRS with positive probability, the induced measure $\hat{\eta}$ must also have full support.[35] By Theorem 1-i of AL, then the cutoffs P^* characterize a *unique* stable matching at economy $\hat{E}$. Finally, observe that, for any P with $||P - P^*|| < \delta$, each type $\theta \notin \Theta^\delta$ has the same set of feasible colleges when the cutoffs are P as when they are P^*. This means that for any such type (ρ, s) and for any P in the set, $D^{(\hat{r}(\rho,s),s)}(P) = D^{(\rho,s)}(P)$. The last statement thus follows. □

Claim 2. *Let $\hat{P}^k$ be the (random) cutoffs characterizing the DA assignment in the F^k when the applicants follow the prescribed strategies. Then, for any $\delta, \epsilon' > 0$, there exists $K \in \mathbb{N}$ such that for all $k > K$,*

$$\Pr\{||\hat{P}^k - P^*|| < \delta\} \geq 1 - \epsilon'.$$

Proof. Let $\hat{\eta}^k$ be the measure of "stated" types $(\hat{r}(\rho), s)$ under k-random economy F^k when the applicants follow the prescribed strategies. Let $\hat{F}^k = [\hat{\eta}^k, S]$ be the resulting "induced" k-random

[35] Throughout, we implicitly assume that a law of large numbers applies. This is justified by focusing on an appropriate probability space as in Sun (2006). Or more easily, we can assume that the applicants are coordinating via asymmetric strategies so that exactly γ fraction of each type plays TRS.

economy. By construction, $\hat{F}^k$ consists of k independent draws of applicants according to measure $\hat{\eta}$, so it is simply a k-random economy of $\hat{E}$. Since by Claim 1, $\hat{\eta}$ has full support, $\hat{D}(\cdot)$ is C^1 in the neighborhood of P^* and $\partial\hat{D}(P^*)$ is invertible, by Proposition 3-2 of AL, for each $\epsilon' > 0$, there exists $K \in \mathbb{N}$ such that for all $k > K$, cutoffs $\hat{P}^k$ of any stable matching of $\hat{E}^k$—and hence the DA outcome of F^k under the prescribed strategies—satisfy

$$\Pr\{||\hat{P}^k - P^*|| < \delta\} \geq 1 - \epsilon'.$$

□

We are now in a position to prove Theorem 1. Fix any $\epsilon > 0$. Take any $\epsilon' > 0$ such that $\epsilon'(\overline{u} - \underline{u}) \leq \epsilon$. By Claim 2, there exists $K \in \mathbb{N}$ such that for all $k > K$, $\Pr\{||\hat{P}^k - P^*|| < \delta\} \geq 1 - \epsilon'$, where $\hat{P}^k$ are the cutoffs associated with the DA matching in F^k under the prescribed strategies. Let $\mathcal{E}^k$ denote this event. 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 $\theta \in \Theta^\delta$ the prescribed strategy, namely TRS, is trivially optimal given the strategyproofness of DA. Hence, consider an applicant with any type $\theta \notin \Theta^\delta$, and suppose that all other applicants employ the prescribed strategies. Now condition on event $\mathcal{E}^k$. Recall that the set of feasible colleges is the same for type $\theta \notin \Theta^\delta$ when the cutoffs are $\hat{P}^k$ as when they are P^*, provided that $||\hat{P}^k - P^*|| < \delta$. Hence, given event $\mathcal{E}^k$, strategy $\hat{r}(\theta)$ is a best response—and hence the prescribed mixed strategy—attains the maximum payoff for type $\theta \notin \Theta^\delta$.

Of course, the event $\mathcal{E}^k$ may not occur, but its 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 $\overline{u} - \underline{u} (\geq \overline{u} - \max\{0, \underline{u}\})$. Hence, the payoff loss she incurs by playing the prescribed mixed strategy is at most

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

This proves that the sequence of strategy profiles for the sequence $\{F^k\}$ of k-random economies forms a robust equilibrium. □

Proof of Theorem 2. For any sequence $\{F^k\}$ induced by E, fix any arbitrary regular robust equilibrium $\{(\sigma^k_{1\leq i\leq k})\}_k$. The strategies induce a random ROL, R_i, for each player i, and (random) per capita demand

$$D^k(P) := \left(\frac{1}{k}\sum_{i=1}^{k} I\left\{c \in \arg\max_{\text{w.r.t. } R_i^k} \{c' \in C : s_{i,c'} \geq P_{c'}\}\right\}\right)_{c\in C},$$

where the set $\{c' \in C : s_{i,c'} \geq P_{c'}\}$ is the set of feasible colleges for applicant i with respect to the cutoff P and $I\{\cdot\}$ is an indicator function equal to 1 if $\{\cdot\}$ holds and 0 otherwise. (Note that the random ROLs R_i's are suppressed as arguments of $D^k(P)$ for notational ease.) Let P^k be the (random) cutoffs, satisfying $D^k(P^k) = S^k$.

Let $\bar{D}^k(P) := \mathbb{E}_{(R_1,...,R_k)}\left[D^k(P)\right]$, where the randomness is taken over the random draws of the types of the applicants and the random reported ROLs according to mixed strategy profile $\left(\sigma_i^k\right)_{1\leq i\leq k}$.

As a preliminary step, we establish a series of claims.

Claim 3. *Fix any P. Then, for any $\alpha > 0$,*

$$\Pr\left[\left\|D^k(P) - \bar{D}^k(P)\right\| > \sqrt{|C|}\alpha\right] \leq |C| \cdot e^{-2k\alpha^2}.$$

Proof. First by McDiarmid's theorem, for each $c \in C$,

$$\Pr\{|D_c^k(P) - \bar{D}_c^k(P)| > \alpha\} \leq e^{-2k\alpha^2},$$

since for each $c \in C$, $|D_c^k(P)(R_1, ..., R_k) - D_c^k(P)(R_1', ..., R_k')| \leq 1/k$ whenever $(R_1, ..., R_k)$ and $(R_1', ..., R_k')$ differ only in one component.

It then follows that

$$\begin{aligned}&\Pr\left[\left\|D^k(P) - \bar{D}^k(P)\right\| > \sqrt{|C|}\alpha\right]\\ &\leq \Pr\left[\exists\ c \in C \text{ s.t. } \left|D_c^k(P) - \bar{D}_c^k(P)\right| > \alpha\right]\\ &\leq |C| \cdot e^{-2k\alpha^2}.\end{aligned}$$

□

Claim 4. *The sequence of functions $\left\{\bar{D}^k(\cdot)\right\}_k$ are equicontinuous (in the class of normalized demand functions across all $k = 1, ..$).*

Proof. Fix $\varepsilon > 0$ and $P \in [0,1]^C$.

We want to find $\delta > 0$ (which may depend on ε and P) s.t.

$$\left\|\bar{D}^k(P') - \bar{D}^k(P)\right\| < \varepsilon$$

for all $P' \in [0,1]^C$ with $\|P' - P\| < \delta$ and all k.

Define

$$\Theta_{P,P'} := \left\{(u,s) \in \Theta : \begin{array}{c}\exists\ c \in C \text{ s.t. } s_c \text{ is weakly greater than}\\ \text{one and only one of } P_c \text{ and } P_c'\end{array}\right\}.$$

We can find $\delta > 0$ s.t. $\eta\left(\Theta_{P,P'}\right) < \varepsilon/\sqrt{|C|}$ for all P' s.t. $\|P' - P\| < \delta$. This can be guaranteed if we assume the measure η to be absolutely continuous w.r.t. Lebesgue measure.

Then we have

$$
\begin{aligned}
&\left\|\bar{D}^k\left(P'\right)-\bar{D}^k\left(P\right)\right\| \\
&=\sqrt{\sum_{c\in C}\left|\mathbb{E}\left[D_c^k\left(P'\right)-D_c^k\left(P\right)\right]\right|^2} \\
&=\sqrt{\sum_{c\in C}\left|\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left(\begin{array}{l} I\left\{c\in\arg\max_{R_i^k}\left\{c'\in C: s_{i,c'}\geq P'_{c'}\right\}\right\} \\ -I\left\{c\in\arg\max_{R_i^k}\left\{c'\in C: s_{i,c'}\geq P_{c'}\right\}\right\}\end{array}\right)\right]\right|^2} \\
&\leq\sqrt{\sum_{c\in C}\left[\mathbb{E}\left|\frac{1}{k}\sum_{i=1}^{k}\left(\begin{array}{l} I\left\{c\in\arg\max_{R_i^k}\left\{c'\in C: s_{i,c'}\geq P'_{c'}\right\}\right\} \\ -I\left\{c\in\arg\max_{R_i^k}\left\{c'\in C: s_{i,c'}\geq P_{c'}\right\}\right\}\end{array}\right)\right|\right]^2} \\
&\leq\frac{1}{k}\sqrt{\sum_{c\in C}\left(\mathbb{E}\sum_{i=1}^{k}\left|\begin{array}{l} I\left\{c\in\arg\max_{R_i^k}\left\{c'\in C: s_{i,c'}\geq P'_{c'}\right\}\right\} \\ -I\left\{c\in\arg\max_{R_i^k}\left\{c'\in C: s_{i,c'}\geq P_{c'}\right\}\right\}\end{array}\right|\right)^2} \\
&\leq\frac{1}{k}\sqrt{\sum_{c\in C}\left(\mathbb{E}\sum_{i=1}^{k}I\left\{\theta_i\in\Theta_{P,P'}\right\}\right)^2}\leq\frac{1}{k}\sqrt{\sum_{c\in C}\left(k\cdot\eta\left(\Theta_{P,P'}\right)\right)^2} \\
&=\sqrt{|C|}\eta\left(\Theta_{P,P'}\right)<\varepsilon,
\end{aligned}
$$

where the first inequality follows Jensen, and the third inequality is because the two sets, $\{c'\in C: s_{i,c'}\geq P'_{c'}\}$ and $\{c'\in C: s_{i,c'}\geq P_{c'}\}$, are identical when $\theta_i\notin\Theta_{P,P'}$. □

Claim 5. *The sequence of functions* $\left\{\bar{D}^k\right\}_{k=1}^{\infty}$ *has a subsequence that converges uniformly to some continuous function* $\bar{D}$.

Proof. Because the sequence of functions $\left\{\bar{D}^k\right\}_{k=1}^{\infty}$ defined on a compact set $[0,1]^C$ are uniformly bounded and equicontinuous (by Claim 4), by Arzela-Ascoli theorem, we can find a subsubsequence $\left\{\bar{D}^{k_j}\right\}_{j=1}^{\infty}$ uniformly convergent to some continuous function $\bar{D}$. □

Claim 6. *For any* $\epsilon'>0$, *there exists a subsequence* $\left\{D^{k_\ell}\right\}_{\ell=1}^{\infty}$ *such that* $\lim_{\ell\to\infty}\Pr\{\sup_P||D^{k_\ell}(P)-\bar{D}(P)||>\epsilon'\}=0$.

Proof. Using the argument in the proof of Glivenko-Cantelli, we can partition the space of P's into finite intervals $\Pi_{i_1,\ldots,i_C}[P_{i_j},P_{i_j+1}]$, where $i_j=0,\ldots,i_j^*$ such that $||\bar{D}(P_{\mathbf{i}+1})-\bar{D}(P_{\mathbf{i}}^-)||<\epsilon'/2$ for all $\mathbf{i}=(i_1,\ldots,i_C)$, where $\mathbf{i}+1:=(i_1+1,\ldots,i_C+1)$. Let m be the number of such intervals. Using the argument of Glivenko-Cantelli, one can show that for any P there exists $\mathbf{i}$ such that

$$||D^{k_\ell}(P)-\bar{D}(P)||\leq\max\{||D^{k_\ell}(P_{\mathbf{i}})-\bar{D}(P_{\mathbf{i}})||,||D^{k_\ell}(P_{\mathbf{i}+1})-\bar{D}(P_{\mathbf{i}+1})||\}+\epsilon'/2.$$

Suppose event $||D^{k_\ell}(P)-\bar{D}(P)||>\epsilon'$ occurs for some P. Then there must exist $\mathbf{i}$ such that $||D^{k_\ell}(P_{\mathbf{i}})-\bar{D}(P_{\mathbf{i}})||\geq\epsilon'/2$. Since $\bar{D}^{k_\ell}(\cdot)$ converges to $\bar{D}(\cdot)$ in sup norm by Claim 5, there exists K'

such that for all $\ell > K'$, $\sup_P ||\bar{D}^{k_\ell}(P) - \bar{D}(P)|| < \epsilon'/4$. Hence, for $\ell > K'$ and for $\mathbf{i}$, we must have

$$||D^{k_\ell}(P_{\mathbf{i}}) - \bar{D}^{k_\ell}(P_{\mathbf{i}})|| \geq \epsilon'/4.$$

Combining the arguments so far, we conclude:

$$\begin{aligned}
&\Pr\{\sup_P ||D^{k_\ell}(P) - \bar{D}(P)|| > \epsilon'\} \\
=&\Pr\{\exists P \text{ s.t. } ||D^{k_\ell}(P) - \bar{D}(P)|| > \epsilon'\} \\
\leq&\Pr\{\exists \mathbf{i} \text{ s.t. } ||D^{k_\ell}(P_{\mathbf{i}}) - \bar{D}(P_{\mathbf{i}})|| > \epsilon'/2\} \\
\leq&\Pr\{\exists \mathbf{i} \text{ s.t. } ||D^{k_\ell}(P_{\mathbf{i}}) - \bar{D}^{k_\ell}(P_{\mathbf{i}})|| > \epsilon'/4\} \\
=&\Pr\{\cup_{\mathbf{i}}\{||D^{k_\ell}(P_{\mathbf{i}}) - \bar{D}^{k_\ell}(P_{\mathbf{i}})|| > \epsilon'/4\}\} \\
\leq&\sum_{\mathbf{i}} \Pr\{||D^{k_\ell}(P_{\mathbf{i}}) - \bar{D}^{k_\ell}(P_{\mathbf{i}})|| > \epsilon'/4\} \\
\leq&\sum_{\mathbf{i}} e^{-k_\ell \epsilon'^2/8} \\
=&me^{-k_\ell \epsilon'^2/8} \to 0 \text{ as } \ell \to \infty,
\end{aligned}$$

where the penultimate inequality follows from Claim 3. □

Now we are in a position to prove Theorem 2.

Suppose to the contrary that the sequence of strategy profiles $\left\{\left(\sigma_i^k\right)_{1\leq i\leq k}\right\}_k$ is not asymptotically stable. Then by definition, there exists $\varepsilon > 0$ and a subsequence of finite economies $\left\{F^{k_j}\right\}_j$ such that

$$\Pr\left(\text{The fraction of applicants playing SRS against } P^{k_j} \geq 1-\varepsilon\right) < 1-\varepsilon. \qquad (*)$$

By Claim 6, we know that there exists a sub-subsequence $D^{k_{j_l}}$ that converges to $\bar{D}$ uniformly in probability. Given the regularity of the strategies employed by the applicants along with the full support assumption, $\bar{D}$ is C^1 and $\partial\bar{D}$ is invertible. Hence (using an argument by AL), we know that $P^{k_{j_l}}$ converges to $\bar{P}$ in probability, where $\bar{P}$ is a deterministic cutoff s.t. $\bar{D}\left(\bar{P}\right) = S$.

Define

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

Let's take δ to be small enough s.t. $\eta\left(\hat{\Theta}\right) > (1-\varepsilon)^{1/3}$ (this can be done since η is absolutely continuous).

By WLLN, we know that $\eta^{k_{j_l}}\left(\hat{\Theta}\right)$ converges to $\eta\left(\hat{\Theta}\right)$ in probability, and therefore there exists L_1 s.t. for all $l > L_1$ we have

$$\Pr\left(\eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2}\right) \geq (1-\varepsilon)^{1/2}.$$

For each economy $F^{k_{j_l}}$, define the event

$$A^{k_{j_l}} := \left\{ \left| P_c^{k_{j_l}} - \bar{P}_c \right| < \delta \text{ for all } c \in C \right\}.$$

Because $P^{k_{j_l}}$ converges to $\bar{P}$ in probability, there exists L_2 s.t. for all $l > L_2$ we have

$$\Pr\left(A^{k_{j_l}}\right) \geq \max\left\{ (1-\varepsilon)^{1/6}, 1 - (1-\varepsilon)^{1/2}\left[(1-\varepsilon)^{1/3} - (1-\varepsilon)^{1/2} \right] \right\}. \qquad (**)$$

Because $\left\{ \left(\sigma_i^k\right)_{1 \leq i \leq k} \right\}_k$ is a robust equilibrium, there exists L_3 s.t. for all $l > L_3$ the strategy profile $\left(\sigma_i^{k_{j_l}}\right)_{i=1}^{k_{j_l}}$ is a $\delta\left[(1-\varepsilon)^{1/6} - (1-\varepsilon)^{1/3}\right]$-BNE for economy $F^{k_{j_l}}$.

By WLLN, there exists $\hat{L}$ s.t. $\hat{L}$ i.i.d. Bernoulli random variables with $p = (1-\varepsilon)^{1/3}$ have a sample mean greater than $(1-\varepsilon)^{1/2}$ with probability no less than $(1-\varepsilon)^{1/3}$. Then we find L_4 s.t. $l > L_4$ implies $(1-\varepsilon)^{1/2} k_{j_l} > \hat{L}$.

Now let's fix an arbitrary $l > \max\{L_1, L_2, L_3, L_4\}$, and we wish to show that in economy $F^{k_{j_l}}$

$$\Pr\left(\text{The fraction of applicants playing SRS against } P^{k_{j_l}} \geq 1-\varepsilon\right) \geq 1-\varepsilon,$$

which would contradict $(*)$ and complete the proof.

First, notice that in economy $F^{k_{j_l}}$, an applicant with $\theta \in \hat{\Theta}$ plays SRS against $\bar{P}$ with probability no less than $(1-\varepsilon)^{1/3}$. To see this, suppose by contrary that there exists some applicant i and some type $\theta \in \hat{\Theta}$ s.t.

$$\Pr\left(\sigma_i^{k_{j_l}}(\theta) \text{ plays SRS against } \bar{P}\right) < (1-\varepsilon)^{1/3}.$$

Then deviating to TRS will give this applicant i with type θ at least a gain of

$$\begin{aligned}
&\delta \cdot \Pr\left(\sigma_i^{k_{j_l}}(\theta) \text{ does not play SRS against } P^{k_{j_l}}\right) \\
&\geq \delta \cdot \Pr\left(\begin{array}{c} \sigma_i^{k_{j_l}}(\theta) \text{ does not play SRS against } \bar{P} \\ \text{and event } A^{k_{j_l}} \end{array}\right) \\
&\geq \delta\left[\Pr\left(A^{k_{j_l}}\right) - \Pr\left(\sigma_i^{k_{j_l}}(\theta) \text{ plays SRS against } \bar{P}\right)\right] \\
&\geq \delta\left[(1-\varepsilon)^{1/6} - (1-\varepsilon)^{1/3}\right],
\end{aligned}$$

which contradicts the construction of L_3, which implies that the strategy profile $\left(\sigma_i^{k_{j_l}}\right)_{i=1}^{k_{j_l}}$ is a $\delta\left[(1-\varepsilon)^{1/6} - (1-\varepsilon)^{1/3}\right]$-BNE for the economy $F^{k_{j_l}}$.

Therefore, in economy $F^{k_{j_l}}$, for each applicant $i = 1, \ldots, k_{j_l}$ and each $\theta \in \hat{\Theta}$, we have

$$\begin{aligned} & \Pr\left(\sigma_i^{k_{j_l}}(\theta) \text{ plays SRS against } \bar{P} \middle| \eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2} \right) \\ = & \Pr\left(\sigma_i^{k_{j_l}}(\theta) \text{ plays SRS against } \bar{P} \right) \geq (1-\varepsilon)^{1/3}, \end{aligned} \tag{***}$$

where the first equality holds because applicant i's random report according to her mixed strategy is independent of random draws of the applicants' type.

Then we have

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

where the first inequality is because of the inequality (***) and that σ_i's are independent across applicants conditioning on the event $\eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2}$, and the second inequality is because $l > L_4$ and $\eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2}$ imply $\eta^{k_{j_l}}\left(\hat{\Theta}\right) \cdot k_{j_l} > \hat{L}$.

Comparing the finite economy random cutoff $P^{k_{j_l}}$ with the deterministic cutoff $\bar{P}$, we have

$$\begin{aligned} & \Pr\left(\begin{array}{c} \text{The fraction of applicants with } \theta \in \hat{\Theta} \\ \text{playing SRS against } P^{k_{j_l}} \geq (1-\varepsilon)^{1/2} \end{array} \middle| \eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2} \right) \\ \geq & \Pr\left(\begin{array}{c} \text{The fraction of applicants with } \theta \in \hat{\Theta} \\ \text{playing SRS against } \bar{P} \geq (1-\varepsilon)^{1/2}, \\ \text{and event } A^{k_{j_l}} \end{array} \middle| \eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2} \right) \\ \geq & \Pr\left(\begin{array}{c} \text{The fraction of applicants with } \theta \in \hat{\Theta} \\ \text{playing SRS against } \bar{P} \geq (1-\varepsilon)^{1/2} \end{array} \middle| \eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2} \right) \\ & - \Pr\left(\bar{A}^{k_{j_l}} \middle| \eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2} \right) \\ \geq & (1-\varepsilon)^{1/3} - \frac{\Pr\left(\bar{A}^{k_{j_l}}\right)}{\Pr\left(\eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2}\right)} \\ \geq & (1-\varepsilon)^{1/3} - \frac{(1-\varepsilon)^{1/2}\left[(1-\varepsilon)^{1/3} - (1-\varepsilon)^{1/2}\right]}{(1-\varepsilon)^{1/2}} \\ = & (1-\varepsilon)^{1/2}, \end{aligned}$$

where the last inequality is because of (**).

The construction of L_1 implies $\Pr\left(\eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2}\right) \geq (1-\varepsilon)^{1/2}$, and so finally we have in economy $F^{k_{j_l}}$

$$
\begin{aligned}
&\Pr\left(\text{The fraction of applicants playing SRS against } P^{k_{j_l}} \geq 1-\varepsilon\right)\\
&\geq \Pr\begin{pmatrix} \text{The fraction of applicants with } \theta \in \hat{\Theta} \\ \text{playing SRS against } P^{k_{j_l}} \geq (1-\varepsilon)^{1/2} \\ \text{and } \eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2} \end{pmatrix}\\
&= \Pr\left(\eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2}\right) \cdot \Pr\left(\begin{matrix} \text{The fraction of applicants with } \theta \in \hat{\Theta} \\ \text{playing SRS against } P^{k_{j_l}} \geq (1-\varepsilon)^{1/2} \end{matrix} \,\middle|\, \eta^{k_{j_l}}\left(\hat{\Theta}\right) \geq (1-\varepsilon)^{1/2}\right)\\
&\geq (1-\varepsilon)^{1/2} \cdot (1-\varepsilon)^{1/2}\\
&= 1-\varepsilon,
\end{aligned}
$$

which contradicts $(*)$. □

Proof of Corollary 2. Fix any regular robust equilibrium. The sequence of strategy profiles in that equilibrium induces a sequence $(r^k)_k$ of expected fractions of students not assigned their most preferred feasible colleges given the equilibrium cutoffs (P^k) (which are random). If any limit outcome is not stable in the continuum economy $E = [\eta, S]$, then there must be a subsequence $(r^{k_\ell})_\ell$ bounded away from 0 for all ℓ. However, this contradicts Theorem 2, which implies that $r^k \to 0$ as $k \to \infty$. Each limit outcome therefore must be stable in $E = [\eta, S]$. By the full support assumption, $E = [\eta, S]$ admits a unique stable matching, so the limit outcome is the unique stable matching outcome of $E = [\eta, S]$. □

C Monte Carlo Simulations

This appendix describes how we estimate applicant preferences under each of the three identifying assumptions, weak truth-telling, stability, and robustness.

We also present additional details of the Monte Carlo simulations. Figure C.2 describes the simulated spatial distribution of applicants and colleges in one simulation sample; Figure C.3 depicts the marginal distribution of each college's cutoffs across 1000 simulations under the assumption that every applicant always truthfully ranks all colleges.

Furthermore, Tables C.4 and C.5 describe the skipping behaviors and mistakes for applicants with $T_i = 1$ and $T_i = 0$, respectively. Table C.6 shows the mean and standard deviation of the estimates of each coefficient from different approaches; and Tables C.7 and C.8 present more detailed estimation results of the welfare effects among applicants with $T_i = 1$ and $T_i = 0$, respectively.

C.1 Estimation

Our formulation of estimation approaches follows Fack, Grenet, and He (2017) who also provide more details on the assumptions for identification and estimation.

We first re-write the random utility model (Equation 6) as follows:

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

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, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$. In the following, $k = 1800$ and $C = 12$.

Let $u_i = (u_{i,1}, \cdots, u_{i,C})$. Following the notations in Section 3, we use $\sigma_i(u_i, s_i)$ to denote applicant i's pure strategy but on the modified type domain: $\sigma_i : \mathbb{R}^C \times [0,1] \to \mathcal{R}$. This modification is necessary because we now allow for utility functions to take any value in $\mathbb{R}^C$, rather than $[\underline{u}, \overline{u}]^C$, and because we consider serial dictatorship, where an applicant's score is identical at every college.

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

Weak Truth-Telling. We start with formalizing the estimation under the weak truth-telling (WTT) assumption. If each applicant i submits $K_i \equiv |\sigma_i(u_i, s_i)|$ ($\leq C$) choices, under the assumption that applicants are weakly truth-telling, σ_i ranks truthfully i's top K_i preferred colleges.

The probability of applicant i submitting $R = (r^1, \ldots, r^{|R|}) \in \mathcal{R}$ is:

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

Under the assumptions that $|\sigma_i(u_i, s_i)|$ 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 $|\sigma_i(u_i, s_i)|$ and obtain:

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

where $c' \not\succ_R c$ indicates that c' is not ranked before c in R, which includes c itself and the colleges not ranked in R. This rank-ordered (or "exploded") logit model can be seen as a series of conditional logit models: one for the top-ranked college (r^1) being the most preferred; another for the second-ranked college (r^2) being preferred to all colleges except the one ranked first, and so on.

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

$$\ln L_{WTT}\big(\boldsymbol{\beta} \mid \mathbf{X}, \{|\sigma_i(u_i, s_i)|\}_i\big) = \sum_{i=1}^{k} \sum_{c \text{ ranked in } \sigma_i(u_i,s_i)} V_{i,c} - \sum_{i=1}^{k} \sum_{c \text{ ranked in } \sigma_i(u_i,s_i)} \ln\Big(\sum_{c' \not\succ_{\sigma_i(u_i,s_i)} c} \exp(V_{i,c'})\Big).$$

The WTT estimator, $\hat{\boldsymbol{\beta}}^{WTT}$, is the solution to $\max_{\boldsymbol{\beta}} \ln L_{WTT}\big(\boldsymbol{\beta} \mid \mathbf{X}, \{|\sigma_i(u_i, s_i)|\}_i\big)$.

Stability. We now assume that the matching is stable and explore how we can identify and estimate applicant preferences. Suppose that the matching is $\mu(u_i, s_i)$, which leads to a vector of cutoffs $P(\mu)$. With information on how colleges rank applicants, we can find a set of colleges that are ex-post feasible to i, $\mathcal{C}(s_i, P(\mu))$. A college is feasible to i, if i's score is above the college's cutoff.

From the researcher's perspective, $\mu(u_i, s_i)$, $P(\mu)$, and $\mathcal{C}(s_i, P(\mu))$ are all random variables because of the unobserved $\epsilon_{i,c}$. The conditions specified by the stability of μ imply the likelihood of applicant i matching with s in $\mathcal{C}(s_i, P(\mu))$:

$$\Pr\left(s = \mu(u_i, s_i) = \underset{c \in \mathcal{C}(s_i,P(\mu))}{\arg\max}\; u_{i,c} | \mathbf{X}_i, \mathcal{C}(s_i, P(\mu)); \boldsymbol{\beta}\right).$$

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

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

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

A key assumption of this approach is that the feasible set $\mathcal{C}(s_i, P(\mu))$ is exogenous to i. As shown in Fack, Grenet, and He (2017), it is satisfied when the mechanism is the serial dictatorship and when there are no peer effects.

Robustness. The robust approach is the same as the stability estimator, except that the feasible set of each applicant, $\mathcal{C}(s_i, P(\mu))$, is modified to be $\mathcal{C}(s_i, P^i(\mu))$, where $P^i(\mu)$ is such that $P^i_s(\mu) = P_s(\mu) + \delta$ if $s \neq \mu(i)$ and $P^i_s(\mu) = P_s(\mu)$ if $s = \mu(i)$. In the results we present here, we choose $\delta = 50/1800$. Recall that the 1800 applicants' scores are uniformly distributed in [0,1].

By inflating the cutoffs of some colleges, we shrink every applicant's set of feasible colleges. Therefore, we increase the probability that i is matched with her most-preferred college in $\mathcal{C}(s_i, P(\mu))$.

We can write down the likelihood function as follows:

$$\ln L_{RB}\big(\boldsymbol{\beta} \mid \mathbf{X}, \mathcal{C}(s_i, P^i(\mu))\big) = \sum_{i=1}^{k} V_{i,\mu(u_i,s_i)} - \sum_{i=1}^{k} \ln\Big(\sum_{c' \in \mathcal{C}(s_i, P^i(\mu))} \exp(V_{i,c'})\Big).$$

The stability estimator, $\hat{\boldsymbol{\beta}}^{RB}$, is the solution to $\max_{\boldsymbol{\beta}} \ln L_{RB}\big(\boldsymbol{\beta} \mid \mathbf{X}, \mathcal{C}(s_i, P^i(\mu))\big)$.

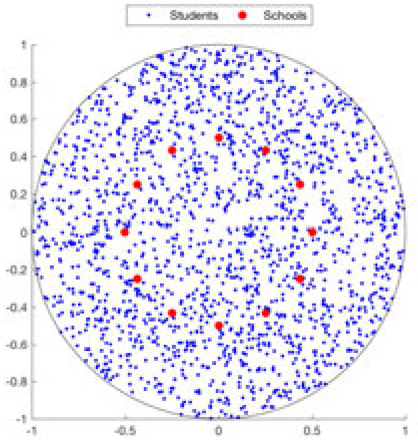


Figure C.2: Monte Carlo Simulations: Spatial Distribution of Applicants and Colleges

Notes: This figure shows the spatial configuration of the area considered in the Monte Carlo simulations with 1800 applicants and 12 colleges. The area is within a circle of radius 1. The blue and red circles show the locations of applicants and colleges, respectively, in one simulation sample. Across samples, the colleges' locations are fixed, while applicants' locations are uniformly drawn within the circle.

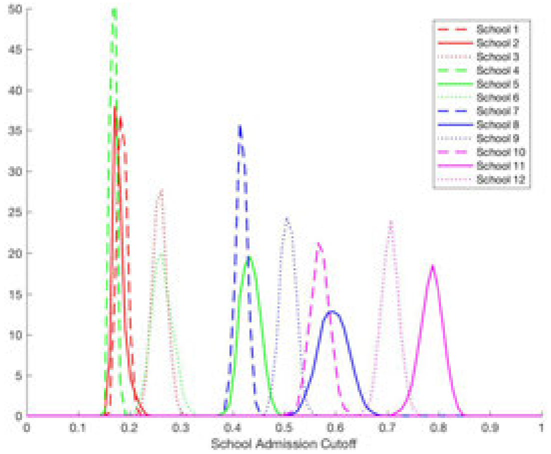


Figure C.3: Simulated Distribution of Cutoffs when Everyone is Truth-telling

Notes: Assuming everyone is strictly 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.

Table C.4: Skips and Mistakes in Monte Carlo Simulations (Percentage Points): $T_i = 1$ Applicants

	Scenarios: Data Generating Processes w/ Different Applicant Strategies							
	Truthful-Reporting Strategy	Payoff Irrelevant Skips			Payoff Relevant Mistakes			
	TRS	IRR 1	IRR 2	IRR 3	REL 1	REL 2	REL 3	REL 4
WTT: Weak Truth-Telling[a]	100	73	43	13	14	14	14	14
Matched w/ favorite feasible college[b]	100	100	100	100	96	91	85	80
Skippers[c]	0	31	63	95	95	96	95	95
By number of skips:								
Skipping 11 colleges	0	19	40	62	79	84	87	90
Skipping 10 colleges	0	7	15	22	14	10	8	5
Skipping 9 colleges	0	4	8	11	3	1	0	0
Skipping 8 colleges	0	0	0	0	0	0	0	0
Skipping 7 colleges	0	0	0	0	0	0	0	0
Skipping 6 colleges	0	0	0	0	0	0	0	0
Skipping 5 colleges	0	0	0	0	0	0	0	0
Skipping 4 colleges	0	0	0	0	0	0	0	0
Skipping 3 colleges	0	0	0	0	0	0	0	0
Skipping 2 colleges	0	0	0	0	0	0	0	0
Skipping 1 college	0	0	0	0	0	0	0	0
TRS: Truthful-Reporting Strategy[d]	100	70	37	5	5	5	5	5

Notes: This table presents the configurations of the eight data generating processes (DGPs), similar to Table 6 but only among the applicants with $T_i = 1$. Each entry is a percentage averaged over the 200 simulation samples. On average, there are 599 such applicants in each sample. [a]An applicant is "*weakly truth-telling*" 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. [b]A college is feasible to an applicant, if the applicant's index (score) is higher than the college's ex-post admission cutoff. If an applicant is matched with her favorite feasible college, she cannot form a blocking pair with any college. [c]Given that every college is acceptable to all applicants and is potentially over-demanded, an applicant is a skipper if she does not rank all colleges. [d]An applicant adopts the "*truthful-reporting strategy*" if she truthfully ranks all available colleges.

Table C.5: Skips and Mistakes in Monte Carlo Simulations (Percentage Points): $T_i = 0$ Applicants

	Scenarios: Data Generating Processes w/ Different Applicant Strategies							
	Truthful-Reporting Strategy	Payoff Irrelevant Skips			Payoff Relevant Mistakes			
	TRS	IRR 1	IRR 2	IRR 3	REL 1	REL 2	REL 3	REL 4
WTT: Weak Truth-Telling[a]	100	91	81	72	72	71	71	71
Matched w/ favorite feasible college[b]	100	100	100	100	99	98	97	95
Skippers[c]	0	23	48	70	70	70	70	70
By number of skips:								
Skipping 11 colleges	0	16	35	52	61	64	65	67
Skipping 10 colleges	0	5	10	14	8	6	5	3
Skipping 9 colleges	0	1	3	4	1	0	0	0
Skipping 8 colleges	0	0	0	0	0	0	0	0
Skipping 7 colleges	0	0	0	0	0	0	0	0
Skipping 6 colleges	0	0	0	0	0	0	0	0
Skipping 5 colleges	0	0	0	0	0	0	0	0
Skipping 4 colleges	0	0	0	0	0	0	0	0
Skipping 3 colleges	0	0	0	0	0	0	0	0
Skipping 2 colleges	0	0	0	0	0	0	0	0
Skipping 1 college	0	0	0	0	0	0	0	0
TRS: Truthful-Reporting Strategy[d]	100	78	52	30	30	30	30	30

Notes: This table presents the configurations of the eight data generating processes (DGPs), similar to Table 6 but only among the applicants with $T_i = 0$. Each entry is a percentage averaged over the 200 simulation samples. On average, there are 599 such applicants in each sample. [a]An applicant is "*weakly truth-telling*" 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. [b]A college is feasible to an applicant, if the applicant's index (score) is higher than the college's ex-post admission cutoff. If an applicant is matched with her favorite feasible college, she cannot form a blocking pair with any college. [c]Given that every college is acceptable to all applicants and is potentially over-demanded, an applicant is a skipper if she does not rank all colleges. [d]An applicant adopts the "*truthful-reporting strategy*" if she truthfully ranks all available colleges.

Table C.6: Estimation with Different Identifying Conditions: Monte Carlo Results

DGPs	Identifying Condition	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.
		A. Strict Truth-telling (All three approaches are consistent)							
TRS	WTT	0.30	0.00	-1.00	0.03	2.00	0.03	0.00	0.02
	Stability	0.30	0.01	-1.00	0.09	2.01	0.12	0.00	0.08
	Robust	0.30	0.01	-1.00	0.10	2.01	0.14	-0.01	0.08
		B. Payoff-irrelevant Skips (Only stability and the robust approach are consistent)							
IRR1	WTT	0.27	0.00	-0.93	0.03	1.85	0.04	-0.04	0.02
	Stability	0.30	0.01	-1.00	0.09	2.01	0.12	0.00	0.08
	Robust	0.30	0.01	-1.00	0.10	2.01	0.14	-0.01	0.08
IRR2	WTT	0.23	0.00	-0.83	0.04	1.58	0.04	-0.10	0.02
	Stability	0.30	0.01	-1.00	0.09	2.01	0.12	0.00	0.08
	Robust	0.30	0.01	-1.00	0.10	2.00	0.14	-0.01	0.08
IRR3	WTT	0.15	0.01	-0.66	0.05	0.99	0.07	-0.20	0.03
	Stability	0.30	0.01	-1.00	0.09	2.01	0.12	0.00	0.08
	Robust	0.30	0.01	-1.00	0.10	2.01	0.14	-0.01	0.08
		C. Payoff-relevant Mistakes (None of the approaches is consistent)							
REL1	WTT	0.17	0.01	-0.69	0.05	1.00	0.07	-0.19	0.03
	Stability	0.29	0.02	-0.98	0.09	1.94	0.22	-0.02	0.12
	Robust	0.29	0.02	-0.99	0.10	1.96	0.20	-0.03	0.11
REL2	WTT	0.17	0.01	-0.70	0.05	1.00	0.08	-0.18	0.03
	Stability	0.28	0.03	-0.96	0.09	1.84	0.30	-0.04	0.14
	Robust	0.29	0.02	-0.98	0.10	1.89	0.27	-0.05	0.13
REL3	WTT	0.17	0.01	-0.70	0.05	1.02	0.08	-0.18	0.03
	Stability	0.27	0.03	-0.94	0.09	1.77	0.37	-0.06	0.16
	Robust	0.28	0.03	-0.96	0.10	1.83	0.33	-0.06	0.15
REL4	WTT	0.18	0.01	-0.71	0.05	1.02	0.08	-0.17	0.03
	Stability	0.26	0.04	-0.92	0.10	1.66	0.43	-0.08	0.16
	Robust	0.27	0.03	-0.94	0.10	1.74	0.38	-0.08	0.15

Notes: This table presents estimates (mean and standard deviation across 200 samples) of the random utility model described in equation (6). The true values are $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.3, -1, 2, 0)$, and the coefficient on the small college dummy is zero. It shows results in the eight data generating processes (DGPs) with three identifying assumptions, WTT, stability, and the robust approach. WTT assumes that every applicant truthfully ranks her top K_i ($1 < K_i \leq 12$) colleges, where K_i is the observed number of colleges in i's ROL. Stability implies that every applicant is matched with her favorite feasible college, given the ex-post cutoffs. The robust approach inflates some cutoffs and re-runs the stability estimator.

Table C.7: Welfare Effects of the Counterfactual Policy on Applicants with $T_i = 1$

DGP	Approach to Counterfactual	Worse Off		Better Off		Indifferent	
		mean	s.d.	mean	s.d.	mean	s.d.
		A. Strict truth-telling					
TRS	Submitted ROLs	0	0	91	1	9	1
	WTT	0	0	91	1	9	1
	Stability	0	0	91	1	9	1
	Robust	0	0	91	1	9	1
	Actual Behavior (the Truth)	0	0	91	1	9	1
		B. Payoff-irrelevant skips					
IRR 1	Submitted ROLs	0	0	79	2	21	2
	WTT	0	0	91	1	9	1
	Stability	0	0	91	1	9	1
	Robust	0	0	91	1	9	1
	Actual Behavior (the Truth)	0	0	91	1	9	1
IRR 2	Submitted ROLs	0	0	65	2	35	2
	WTT	0	0	89	1	11	1
	Stability	0	0	91	1	9	1
	Robust	0	0	91	1	9	1
	Actual Behavior (the Truth)	0	0	91	1	9	1
IRR 3	Submitted ROLs	0	0	53	3	47	3
	WTT	0	0	86	1	14	1
	Stability	0	0	91	1	9	1
	Robust	0	0	91	1	9	1
	Actual Behavior (the Truth)	0	0	91	1	9	1
		C. Payoff-relevant mistakes					
REL 1	Submitted ROLs	0	0	45	3	55	3
	WTT	0	0	87	1	13	1
	Stability	0	0	91	1	9	1
	Robust	0	0	91	1	9	1
	Actual Behavior (the Truth)	0	0	91	1	9	1
REL 2	Submitted ROLs	0	0	43	2	57	2
	WTT	0	0	87	1	13	1
	Stability	0	0	91	2	9	2
	Robust	0	0	91	1	9	1
	Actual Behavior (the Truth)	0	0	91	1	9	1
REL 3	Submitted ROLs	0	0	42	2	58	2
	WTT	0	0	87	1	13	1
	Stability	0	0	90	2	10	2
	Robust	0	0	91	2	9	2
	Actual Behavior (the Truth)	0	0	91	1	9	1
REL 4	Submitted ROLs	0	0	42	2	58	2
	WTT	0	0	87	1	13	1
	Stability	0	0	90	2	10	2
	Robust	0	0	90	2	10	2
	Actual Behavior (the Truth)	0	0	91	1	9	1

Notes: This table presents the estimated effects of the counterfactual policy (giving $T_i = 1$ applicants priority in admission) on applicants with $T_i = 1$. On average, there are 599 such applicants (standard deviation 14) in each simulation sample. The table shows results in the eight data generating processes (DGPs) with five 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); stability implies that every applicant is matched with her favorite feasible college, given the ex-post cutoffs; and the robust approach inflates some cutoffs and re-runs the stability estimator. 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 200 simulation samples.

Table C.8: Welfare Effects of the Counterfactual Policy on Applicants with $T_i = 0$

DGP	Approach to Counterfactual	Worse Off		Better Off		Indifferent	
		mean	s.d.	mean	s.d.	mean	s.d.
		A. Strict truth-telling					
TRS	Submitted ROLs	68	2	0	0	32	2
	WTT	68	2	0	0	32	2
	Stability	67	2	1	0	32	2
	Robust	67	2	1	0	32	2
	Actual Behavior (the Truth)	68	2	0	0	32	2
		B. Payoff-irrelevant skips					
IRR 1	Submitted ROLs	55	2	0	0	45	2
	WTT	65	2	2	0	33	2
	Stability	67	2	1	0	32	2
	Robust	67	2	1	0	32	2
	Actual Behavior (the Truth)	68	2	0	0	32	2
IRR 2	Submitted ROLs	40	2	0	0	60	2
	WTT	60	2	5	1	35	2
	Stability	67	2	1	0	32	2
	Robust	67	2	1	0	32	2
	Actual Behavior (the Truth)	68	2	0	0	32	2
IRR 3	Submitted ROLs	30	1	0	0	70	1
	WTT	47	2	13	1	40	2
	Stability	67	2	1	0	32	2
	Robust	67	2	1	0	32	2
	Actual Behavior (the Truth)	68	2	0	0	32	2
		C. Payoff-relevant mistakes					
REL 1	Submitted ROLs	26	1	0	0	74	1
	WTT	50	2	11	1	39	2
	Stability	67	3	1	1	32	2
	Robust	67	3	1	1	32	2
	Actual Behavior (the Truth)	68	2	0	0	32	2
REL 2	Submitted ROLs	25	1	0	0	75	1
	WTT	51	2	11	1	38	2
	Stability	66	3	2	2	32	2
	Robust	66	3	2	1	32	2
	Actual Behavior (the Truth)	67	2	0	0	32	2
REL 3	Submitted ROLs	24	1	0	0	76	2
	WTT	51	2	11	1	38	2
	Stability	65	4	3	2	33	2
	Robust	65	4	2	2	32	2
	Actual Behavior (the Truth)	67	2	0	0	32	2
REL 4	Submitted ROLs	23	1	0	0	76	2
	WTT	52	2	11	1	37	2
	Stability	63	5	4	3	33	3
	Robust	64	4	3	2	33	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 $T_i = 0$. On average, there are 1201 such applicants (standard deviation 14) in each simulation sample. The table shows results in the eight data generating processes (DGPs) with five 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); stability implies that every applicant is matched with her favorite feasible college, given the ex-post cutoffs; and the robust approach inflates some cutoffs and re-runs the stability estimator. 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 200 simulation samples.