Strategic ‘Mistakes’: Implications for Market Design
Research*

Georgy Artemov†  Yeon-Koo Che‡  Yinghua He§

March 3, 2020

Abstract

We document that a non-negligible fraction of Australian college applicants adopt unambiguously dominated strategies in strategically straightforward situations; however, the majority of these “mistakes” are payoff irrelevant. We propose an equilibrium concept that accommodates such mistakes. Under a strategy-proof mechanism, equilibrium strategies need not be truth-telling, but every equilibrium outcome is asymptotically stable. Therefore, mistakes have limited influence on outcomes but complicate empirical analyses. Using Li (2017)’s experimental data, we illustrate the vulnerability of the empirical methods based on truth-telling or stability; our proposed novel approach is more robust to mistakes and can extract reliable information from data containing mistakes.

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.

*We thank Xingye Wu, who has provided excellent research assistance for the theory part, and Julien Grenet for his generous help with the Monte Carlo simulations. We are grateful to the seminar/conference participants at ANU, Asia-Pacific IO Conference, ASSA Meeting, Barcelona GSE Summer Forum, Boston College, Deakin, Columbia, Conference on Economic Design, “Econometrics Meets Theory” Conference at NYU, European Meeting on Game Theory, “Dynamic Models in Economics” Workshop at NUS, Hitotsubashi, Higher School of Economics, Game Theory and Management Conference, MIT, NBER Market Design Group Meeting, PET Conference, Workshop “Matching in Practice”, Paris School of Economics, Stony Brook Conference on Game Theory, UC Irvine, University of Queensland, and Waseda for their comments. Authors acknowledge support from the Australian Research Council (DP160101350) and the University of Melbourne (Artemov); the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (2016S1A2A2912564) through Global Research Network program (Che); and National Science Foundation (1730636; He).

†Department of Economics, University of Melbourne, Australia. Email: georgy@gmail.com
‡Department of Economics, Columbia University, USA. Email: yeonkooche@gmail.com.
§Department of Economics, Rice University, USA; Toulouse School of Economics, France. Email: yinghua.he@gmail.com.
1 Introduction

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

However, alarming evidence has been accumulating that strategic mistakes are common, even in strategy-proof environments. Laboratory experiments show that a significant fraction of subjects mis-report their preferences 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 importantly, similar mistakes occur in high-stakes real-world contexts. Chen and Pereyra (2015), Hassidim, Romm, and Shorrer (2016) and Shorrer and Sóvágó (2018) find that 10–20% of applicants for Mexico City high schools, Israeli graduate programs in psychology, and Hungarian colleges use obviously dominated strategies in DA; 17% of surveyed applicants in the National Resident Matching Program report using a dominated strategy (Rees-Jones, 2017).

Participants’ mistakes have recently attracted significant attention in the theoretical literature. Li (2017) posits that participants do not comprehend every detail of a mechanism and may not play a dominant strategy when its dominance is not “obvious”; Pycia and Troyan (2019) argue that even some obviously-dominant strategies may not be simple enough.

If participant mistakes are common and arbitrary, they will pose a challenge to the market-design literature, both theoretical and empirical. Most theoretical results depend on the assumption that participants play their unique dominant strategy: truth-telling. Much of the empirical literature relies on similar assumptions and ignores mistakes, potentially rendering increasingly-available data from strategy-proof mechanisms useless.

Our aim is to better understand the implications of strategic mistakes. We start by observing that, in the studies cited above, participants can usually predict the consequences of their actions; for example, applicants can recognize which colleges are likely out-of-reach when admissions are based on test scores. Building on this observation, we reach two conclusions. First, the mistakes documented in the literature have a limited impact on the outcome of a strategy-proof mechanism: despite the participants’ obvious mistakes, the

---

1 A voluminous literature documents non-optimal behaviors in experimental auctions (for a survey, see Kagel and Levin, 2008). However, the field evidence in auctions is less conclusive as it is more difficult to establish the non-optimality of participant behaviors (see, for example, Malmendier and Lee (2011) and Schneider (2016)).

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.
outcome is similar to the one that would have emerged had everyone been truth-telling. Therefore, as far as market outcomes are concerned, theoretical predictions are not far from reality. Second, participant behavior may be wildly different from truth-telling, leading to our second conclusion: empirical methods need to pay close attention to mistakes. On the one hand, commonly used empirical methods recover useful information from the data, contradicting the extreme view that the data is useless due to the documented mistakes. On the other hand, the researcher need to consider the prevalence of mistakes in the data. If mistakes are likely to be prevalent, the truth-telling assumption may bias empirical results, as we illustrate using simulated and Li (2017)’s experimental data. For these cases, we propose a novel empirical approach that is robust to mistakes.

To reach these conclusions, we conduct an empirical analysis of strategic mistakes, develop a theoretical model allowing for mistakes, and evaluate empirical methods for estimating applicant preferences. The results are summarized below.

**Empirical analysis of strategic mistakes.** Using a dataset from the centralized tertiary assignment in Victoria, Australia, we identify a form of strategic mistake, called a *skip*: an applicant applying to a full-tuition option for a “course”—a university-major pair—while failing to apply to the half-tuition option for the same course even though doing so would incur no cost. Following Hassidim, Romm, and Shorrer (2016), we classify such behavior as a dominated strategy, or a “mistake.” Note that the analysis relies on unambiguously identifying some dominated strategies, but not on specific details of the mechanism. Both papers find that such mistakes are common, but the vast majority of them are payoff *irrelevant*: correcting a payoff-irrelevant mistake would not change applicant’s outcome. Going beyond what they study, we investigate what correlates with payoff-relevant mistakes and find that payoff-relevant and payoff-irrelevant mistakes are of different nature. This finding is reinforced by the analysis of changes in applicants’ rank-order lists (ROLs) of colleges submitted before and after colleges’ ranking of applicants is released, a unique feature in our dataset.

This finding is consistent with the hypothesis that the observed “mistakes” are a result of applicants simply not bothering to rank either (i) “out-of-reach” options—those they would stand little chance of getting in even if they applied to them—or (ii) options that are irrelevant because they are worse than some clearly “within-reach” (feasible) options.

**Theoretical model allowing for mistakes.** We develop a theoretical model of applicant behavior in a large matching market operated by a DA mechanism. Each college ranks applicants by some score, and every applicant knows her own score before submitting an ROL. Consistent with the empirical findings, we focus on an equilibrium concept—called *robust equilibrium*—which lets applicants make mistakes as long as the payoff consequences of these mistakes vanish when the market grows large.

We show that a dramatic departure from truth-telling—all but a vanishing fraction of
applicants submitting untruthful ROLs—is supported as a robust equilibrium (Theorem 1). Hence, if the equilibrium concept must tolerate the type of behavior observed in the data, then we cannot trust truth-telling as a reliable prediction. Indeed, no sharp prediction of applicant behavior is possible because a wide range of behaviors is consistent with robust equilibria. This result thus questions the empirical methods that rely on truth-telling.

Nevertheless, not all is lost due to non-truthful behavior. Despite the behavioral multiplicity and ambiguity, under a mild condition, all robust equilibria yield a virtually unique outcome for a sufficiently large economy. The outcome is asymptotically stable and approximates the outcome arising from the truth-telling behavior (Theorem 2). These latter results calls for different empirical methods, which we turn to next.

**Empirical methods for preference estimation** rely crucially on assumptions on participant behavior. Much of the existing literature assumes that applicants are truthful. Specifically, the weak truth-telling (WTT) hypothesis assumes that an applicant ranks her most-preferred colleges truthfully, but may not rank all colleges. WTT makes a correct inference when an applicant omits a less-preferred college (type-ii options as described earlier), but makes an incorrect inference when she omits a more-preferred out-of-reach college (type-i options). Thus, WTT tends to underestimate the preferences for popular colleges.

This problem can be addressed, at least for a large economy, if one restricts preference inference to feasible colleges. **Stability** assumes that an applicant is assigned her most-preferred feasible college but “refuses” to infer anything about the infeasible ones. However, as our theory shows, stability is justified only in the limit economy. In any finite economy, applicants can be uncertain about colleges’ cutoffs. An applicant may consider a college infeasible to her at the time of the application, even though the college turns out feasible ex-post. To address this problem, we propose a novel approach called **robust stability**. Our approach “declares” a college infeasible to an applicant if its cutoff is close to her score, thus making no inference about this college.

We show that the three methods exhibit a nesting structure with respect to the misclassification of inferred preferences. WTT misclassifies applicant preferences the most; stability eliminates misclassifications related to payoff-irrelevant mistakes; robust stability further reduces the misclassification of preferences related to payoff-relevant mistakes. The robustness is achieved by using less information, thus increasing the variance of the estimation.

We illustrate the performance of the three methods in Monte Carlo simulations. Robust stability leads to smaller biases than stability in datasets that contain payoff-relevant mistakes; the WTT-based estimator is inconsistent in the presence of any mistakes. When the estimated preferences are taken to predict counterfactual outcomes, an inconsistent estimator produces misleading predictions.

We next take these three approaches to the data obtained from Li (2017)’s experiment
in which applicants compete for colleges of different monetary values. We confirm our conclusions by in-sample fit (i.e., predicting the true preference order and matching outcome). The WTT-based estimator substantially understates the preferences for top colleges. The stability-based estimator remains vulnerable to payoff-relevant mistakes; it performs worse than the robust-stability-based estimator whenever payoff-relevant mistakes are common.

**In sum**, the current paper reinforces the recent literature that highlights participants’ mistakes, but it differs in its focus on their nature and, more importantly, their implications for market design and its research methods from the following two aspects.

First, our contribution complements the perspective developed by Li (2017). Consistent with his perspective, the non-obviousness of strategy-proof mechanisms may have contributed to the prevalence of mistakes documented here. However, Li’s theory does not tell us what mistakes we could expect. We find in the data that mistakes are mostly inconsequential. We then study a solution concept that allows for such mistakes\(^3\) and find that equilibrium outcome is close to the one that would emerge under optimal play. This is reassuring because it is often impossible to implement the outcome of popular mechanisms, such as the DA, by an obviously-strategy-proof mechanism (Ashlagi and Gonczarowski, 2018).\(^4\)

Second, while largely harmless in terms of the performance of the affected mechanism, strategic mistakes can nevertheless impact the empirical methods for estimating preferences. In particular, the popular method based on WTT is not robust to mistakes, whereas the new method we propose, robust stability, is more robust to them.

The rest of the paper is organized as follows. Sections 2, 3, 4 and 5 present three main parts in detail. The relationship with the literature, which will become clear with the main parts already presented, will be discussed in Section 6. Section 7 concludes.

## 2 Mistakes in Australian College Admissions

### 2.1 Institutional Details and Data

We use the data from the Victorian Tertiary Admission Centre (VTAC) in 2007. VTAC is a centralized clearinghouse for admissions to tertiary courses in Victoria, Australia. Applicants are required to rank tertiary courses by which they wish 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

---

\(^3\)This interpretation is similar in spirit to the perspectives of rational inattention and quantal response equilibria, which view individuals as making mistakes but at the same time “economizing” in their frequencies based on their payoff consequences.

\(^4\)A direct comparison with Li (2017) on the performance of such mechanisms is not possible, since unlike the current paper, its objective is more normative and does not study the performance of strategy-proof mechanisms that are not obviously strategy-proof.
institution may be either a university (including programs not granting bachelor’s degrees) or a technical school. A field of study is roughly equivalent to a major in US universities. In 2007, the full tuition (FT) is approximately AUD17,000 (USD13,000) per year. The reduced tuition (RT) is set by the government and has a median about AUD7,000 (USD5,500) per year. The normal duration of a university course is three years. The government offers student loans, which are subsidized for students in RT courses. Apart from tuition payments, there is no difference between FT and RT courses. Furthermore, both FT and RT courses share the same course description (see Figure A.1 for an example), so there is no information friction associated with finding the RT course corresponding to an FT course. In total, there were 1899 tertiary programs in 2007. 881 of them offered both the FT and RT options, among which 97 percent were university programs as described above. There were also approximately 800 RT-only programs and 200 FT-only programs.

Applicants are required to submit an ROL, along with other information, in September. In December, applicants receive their Equivalent National Tertiary Entrance Ranks (ENTER), as a number between 0 and 99.95 in 0.05 increments, which we will refer to as Score. For applicant $i$, $Score_i$ is the percentage of applicants with scores below $Score_i$. For most applicants, it is the sole determinant of the admission. As ROLs are initially submitted before score is known, applicants can revise their ROL after scores are released. Applicants receive their offers in January-February and have two weeks to enroll in the course they are offered.

In early January, courses rank applicants and send the list of acceptable applicants to VTAC. Using an applicant’s ROL, VTAC picks the highest-ranked course to which the applicant is acceptable, one for each tuition type (FT or RT), and transmits the offer(s) to the applicant. That is, if an applicant’s ROL contains both RT and FT courses, this applicant may receive two offers, one of each type. This feature means that ranking an FT course ahead of the corresponding RT course is not a dominated strategy.

When ranking applicants, courses follow a published set of rules. For the largest category of applicants, labeled “V16” by VTAC, admission is based almost exclusively on their scores. Our analysis focuses on these applicants. 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 by a “cutoff”: the lowest score sufficient for admission to the course.\(^5\)

Applicants can rank up to 12 courses. If an applicant ranks exactly 12 courses, it is possible that the applicant has more than 12 acceptable courses, but is unable to rank all of them. We exclude such applicants because we cannot make any inference about their optimal strategy. Applicants who rank less than 12 courses comprise 75 percent of V16 applicants.\(^5\)

\(^5\)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.
Out of the 27,922 V16 applicants who rank fewer than 12 courses (below we refer to these applicants as the “full sample”), 24,625 applicants rank at least one program that offers both RT and FT courses. Among them, 2,915 applicants rank at least one FT course that has a corresponding RT version. These 2,915 applicants constitute the “FT subsample.”

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

<table>
<thead>
<tr>
<th></th>
<th>Full sample</th>
<th>FT subsample</th>
<th>Skips</th>
<th>Payoff-relevant mistakes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of applicants</td>
<td>27,922</td>
<td>2,915</td>
<td>1,009</td>
<td>201</td>
</tr>
<tr>
<td>Percentage of the full sample</td>
<td>100.00</td>
<td>10.44</td>
<td>3.61</td>
<td>0.72</td>
</tr>
<tr>
<td>Percentage of the FT subsample</td>
<td>100.00</td>
<td>100.00</td>
<td>34.6</td>
<td>6.90</td>
</tr>
<tr>
<td>Percentage of the “Skips” subsample</td>
<td>100.00</td>
<td>100.00</td>
<td>19.92</td>
<td>19.92</td>
</tr>
<tr>
<td>Average length of submitted ROLs</td>
<td>6.61</td>
<td>7.67</td>
<td>7.47</td>
<td>7.65</td>
</tr>
<tr>
<td>Average number of RT courses in submitted ROLs</td>
<td>6.20</td>
<td>5.34</td>
<td>4.72</td>
<td>4.55</td>
</tr>
<tr>
<td>Average number of FT courses in submitted ROLs</td>
<td>0.41</td>
<td>2.33</td>
<td>2.75</td>
<td>3.10</td>
</tr>
</tbody>
</table>

Notes: “Full sample” refers to all V16 applicants who rank fewer than 12 courses in the 2007 college admissions. “FT subsample” are the applicants from the full sample who rank at least one full-tuition course which has a corresponding reduced-tuition version. “Skips” refers to the applicants who rank a full-tuition course but not the corresponding reduced-tuition 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-tuition course.

2.2 Skips and Payoff-Relevant Mistakes

If an applicant who does not exhaust her ROL ranks an FT course but does not rank the corresponding RT course, we say that the applicant “skips.” Even though a skip is a dominated strategy, the skip may not affect the applicant’s assignment for two reasons. First, the applicant’s Score may be below the course’s cutoff, hence the course is “out of reach” for the applicant. Second, the applicant may have been assigned 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 skip is payoff-relevant. We will demonstrate that most of the skips are not payoff-relevant.

As we do not know how in the applicant’s ROL a skipped course should have been ranked, we report the lower and upper bounds of payoff-relevant mistakes. To calculate the lower bound, we assume that the skipped RT course would have been ranked below all the RT courses in the applicant’s ROL. To calculate the upper bound, we assume that the skipped RT course would have been ranked first in the applicant’s ROL. Please see Appendix A for details on calculating the bounds.

In Table 1, we report the number of applicants who skip at least one RT course and the number of applicants for whom skipping an RT course becomes a payoff-relevant mistake.

---

6In 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 following tables is usually smaller than the original size of the full sample or the FT subsample.
at least once, showing both the upper and lower bounds. The table suggests that, among applicants who rank at least one FT 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.

### 2.2.1 Correlation between Score and Skip

Our hypothesis, denoted by $H_1$, is that applicants omit out-of-reach courses. A course is out of reach for an applicant if its cutoff is above her Score. Omitting such a course has no adverse effect on the applicant as she would not be admitted to it even if she had ranked it.

Hypothesis $H_1$ implies a negative correlation between skips and Score. Indeed, an RT course normally has a higher cutoff than its FT counterpart. An RT course may be out-of-reach and dropped, while its FT counterpart may still be within-reach and retained in her ROL. That would be identified as a skip. As more courses are out of reach for an applicant with a low score, such an applicant is more likely to drop an RT course while keeping its FT counterpart hence be identified as making a skip.

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

$$
\text{Skip}_i \times 100 = \alpha + \beta \text{Score}_i + \text{Controls}_i + \epsilon_i,
$$

where $\text{Skip}_i = 1$ if applicant $i$ has made at least one skip and is zero otherwise, and $\text{Controls}_i$ includes $i$’s demographics. We expect $\beta$ to be negative.

### Table 2: Probability of Skipping a Reduced-tuition Course

<table>
<thead>
<tr>
<th></th>
<th>Full sample</th>
<th>FT subsample</th>
<th>Full sample</th>
<th>FT subsample</th>
<th>Full sample</th>
<th>FT subsample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
</tr>
<tr>
<td><strong>Score</strong></td>
<td>-0.06***</td>
<td>-0.71***</td>
<td>-0.04***</td>
<td>-0.55***</td>
<td>-0.04***</td>
<td>-0.56***</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.06)</td>
<td>(0.01)</td>
<td>(0.08)</td>
<td>(0.01)</td>
<td>(0.07)</td>
</tr>
<tr>
<td><strong>GAT</strong></td>
<td>-0.05***</td>
<td>-0.35***</td>
<td>-0.04***</td>
<td>-0.33***</td>
<td>(0.01)</td>
<td>(0.10)</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.12)</td>
<td>(0.01)</td>
<td>(0.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(School fees &gt;AUD11,000) $\times$ Score</td>
<td>-0.03**</td>
<td>-0.05**</td>
<td></td>
<td></td>
<td>(0.01)</td>
<td>(0.02)</td>
</tr>
<tr>
<td># of Applicants</td>
<td>26,325</td>
<td>2,766</td>
<td>26,325</td>
<td>2,766</td>
<td>26,325</td>
<td>2,766</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.37</td>
<td>0.29</td>
<td>0.37</td>
<td>0.30</td>
<td>0.36</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Notes: The dependent variable in every regression is $\text{Skip} \times 100$; $\text{Skip} = 1$ 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 FT 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-tuition 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 FT

---

7We use $\text{Skip}_i \times 100$ to make the results more readable. If we use $\text{Skip}_i$ instead, every estimate will be one percent of those reported here.
subsample. The control variables include school fixed effects and application demographics, e.g., the applicant’s gender, median income (in logarithm) of the residence area (at the postal code level), citizenship status, region born, and language spoken at home. We also include eleven dummy variables that correspond to the number of FT courses ranked by an applicant, because the number of FT courses ranked may have a “mechanical” effect on the number of skips: if no FT course is ranked, then no RT course can be skipped, by definition; so, the more FT courses are ranked, the more opportunities there are for skipping. Our results are robust to the exclusion of these dummies.

Columns (1) and (2) are baseline regressions showing the negative relation between skips and scores. To account for applicant ability, the General Achievement Test (GAT) is included as an explanatory variable. Although GAT and Score are both correlated with an applicant ability, GAT is not used in admissions and does not affect the assignment, while Score does. 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. While GAT is an ability test, Score aggregates the grades achieved by an applicant from her high school classes. These classes may differ significantly among applicants with the same Score. Score is the most similar to a grade point average in the U.S. system. The generic and standardized nature of GAT likely makes it a better measure of the comprehension of the mechanism compared to Score.

When both GAT and Score are included in the regression, the coefficient on Score shows how much more likely, in percentage terms, an applicant with the same GAT but a different admission probability (captured by 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 cognitive abilities may play a role in explaining skips. Even after controlling for cognitive ability, Score continues to be negatively correlated with skip, which is consistent with Hypothesis H1.

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

---

8The 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 Score.

9We cannot include the dummy variable “School fees > AUD11,000” alone because of the inclusion of high school fixed effects in all regressions.
2.2.2 Payoff-relevant Mistakes

Unlike skips, which we expect to vary systematically with Score, Hypothesis $H_1$ 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: the mistake is payoff-relevant if adding a skipped RT course at the top of applicant ROL would change her assignment.

Table 3 presents the results of the following regression:

$$Payoff\text{-}relevant\ Mistake_i \times 100 = \nu + \gamma Score_i + Controls_i + \epsilon_i,$$

where $Payoff\text{-}relevant\ Mistake_i = 1$ if $i$ makes a payoff-relevant skip, 0 otherwise.

Table 3: Probability of Making A Payoff-relevant Mistake

<table>
<thead>
<tr>
<th></th>
<th>Full sample</th>
<th>Subsample including applicants who</th>
<th>FT subsample</th>
<th>Skip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Score</td>
<td>0.01***</td>
<td>0.01***</td>
<td>0.16***</td>
<td>0.19***</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>GAT</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>(School fees &gt;AUD11,000) \times Score</td>
<td>0.01</td>
<td>-0.11*</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.06)</td>
<td></td>
<td>(0.27)</td>
</tr>
<tr>
<td>Other controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td># Applicants</td>
<td>26,325</td>
<td>26,325</td>
<td>2,766</td>
<td>2,766</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.14</td>
<td>0.14</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Notes: The dependent variable, $Payoff\text{-}relevant\ Mistake_i \times 100$, 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 FT 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-tuition courses. Standard errors clustered at high school level are in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Note that the coefficients of two variables that were significantly negative in the regressions of skips – GAT and interaction of Score and school fees – are no longer robustly significant. Only the interaction is significantly negative (at a 10% level) in one specification (column 4). Combining the regression results of equations (1) and (2), which are reported in Tables 2 and 3 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 may be explained “mechanically”: skipping an RT course is more likely to be payoff-relevant for an
applicant with a high \textit{Score}.\footnote{A higher \textit{Score} may make a skipped RT course feasible. When we put such a course back at the top of an ROL, it will be counted as a payoff-relevant mistake.}

\subsection*{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” ROLs several months before applicants learn their scores. After learning the scores, applicants are allowed to change their ROLs. If not changed, a preliminary ROL becomes final and is used for admission. Because a small effort is needed to change an ROL, we treat the preliminary ROL as the best estimate of the final ROL that the applicant would submit given the information that she has at the time.

We investigate the change in the frequencies of skips and payoff-relevant mistakes using the following two regressions:

\begin{align*}
\Delta(\text{#Skips}_i) &= \tau^s + \text{Demeaned Controls}_i + \epsilon_i, \\
\Delta(\text{#Payoff-relevant Mistakes}_i) &= \tau^m + \text{Demeaned Controls}_i + \epsilon_i,
\end{align*}

(3) (4)

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

Columns (1)–(4) of Table 4 show the results. As the number of ranked FT courses may have a mechanical effect on skips and mistakes, we add the change in the number of FT courses as a control variable in columns (3) and (4). The main finding is that, while payoff-relevant mistakes become less frequent in the final ROL, skips become more frequent. The constant is negative and significant in columns (2) and (4), implying that the number of payoff-relevant mistakes decreases over time on average in the sample. By contrast, the effect of the revision of ROL on skips (columns 1 and 3) is significantly positive. Our conclusions remain the same when GAT is included as a control variable. The results are consistent with applicants responding to new information regarding the set of within-reach colleges and re-optimizing their ROLs. Alternative explanations, such as learning about the mechanism, would not predict the increase in the number of skips.

Furthermore, to capture an applicant’s response to an unexpectedly high or low \textit{Score}, we introduce an additional explanatory variable, shock to \textit{Score}, into equations (3) and (4). It is defined as the difference between an applicants’ realized and expected \textit{Scores}, where the expected \textit{Score} is predicted using GAT (see Appendix A.3 for details).

We report the results in columns (5)–(8) of Table 4. Mechanically, if an applicant does...
Table 4: Skips and Payoff-relevant Mistakes: Changes over Time

<table>
<thead>
<tr>
<th>Changes over time</th>
<th>Effects of a shock to Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Skips</td>
<td>#Mistakes</td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.05***</td>
</tr>
<tr>
<td>(0.14)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Shock to Score/100</td>
<td>-2.33</td>
</tr>
<tr>
<td>(2.28)</td>
<td>(0.72)</td>
</tr>
<tr>
<td>Change in # FT courses</td>
<td>43.24***</td>
</tr>
<tr>
<td>(2.75)</td>
<td>(1.26)</td>
</tr>
<tr>
<td>Other Controls</td>
<td>Yes</td>
</tr>
<tr>
<td># of Applicants</td>
<td>26,325</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Notes: “Mistake” means a payoff-relevant mistake. “FT courses” means full-tuition 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. * \(p < 0.10\), ** \(p < 0.05\), *** \(p < 0.01\).

not re-optimize her ROL, 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 skips 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).

3 Theoretical Implications of Strategic Mistakes

Suppose, as have been found in the preceding analysis, applicants indeed make mistakes but rarely with any real payoff consequences. What does this mean, in terms of applicants’ behavior and outcome? In this section, we address this question theoretically. We consider a large matching market operated by the deferred acceptance algorithm (Gale and Shapley, 1962), but adopt an equilibrium concept that permits participants to make mistakes as long as the mistakes become virtually payoff irrelevant as the market size grows arbitrarily large.

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 colleges, \(C = \{c_1, \ldots, c_C\}\) and a set of applicants. Each applicant has a type \(\theta = (u, s)\), where \(u = (u_1, \ldots, u_C) \in [\underline{u}, \bar{u}]^C\) is a vector of von-Neumann Morgenstern utilities of attending colleges for some \(0 \leq u < \bar{u}\), and \(s = (s_1, \ldots, s_C) \in [0, 1]^C\) is a vector of scores representing the colleges’ preferences or applicants’ priorities at colleges, such that an applicant with a higher score has a higher priority at a college. A vector \(u\) induces ordinal preferences over colleges, denoted by an ROL of colleges, \(\rho(u)\), of
length 0 ≤ ℓ ≤ C. We assume that the outside option for the applicants has a zero utility, which means that all colleges are acceptable. This assumption facilitates our analysis but is not crucial for our results. Let Θ = [u, π]C × [0, 1]C denote the set of applicant types. One special case is the serial dictatorship, in which colleges’ preferences are given by a single score. We incorporate this case by imposing an additional restriction on scores s₁ = ... = s_C. Australian college admissions can be seen as a case of the serial dictatorship.

A continuum economy consists of the same finite set of colleges and a unit mass of applicants with type θ ∈ Θ and is given by E = [η, 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 = (S₁, ..., S_C) available at the colleges, where S_j > 0 and ∑_{j=1}^{C} S_j < 1. We assume that η admits continuous density that is bounded below by zero (i.e., full support). In the case of the 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 applicants’ or colleges’ preferences arises only for a measure 0 set of applicants. The 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 µ : C ∪ Θ → 2^Θ ∪ (C ∪ Θ) 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 to satisfy individual rationality and no-blocking.

According to AL, a stable matching is characterized via market-clearing cutoffs, P = (P₁, ..., P_C) ∈ [0, 1]^C, satisfying the demand-supply condition D_c(P) ≤ S_c with equality in case of P_c > 0 for each c ∈ C, where the demand D_c(P) for college c is the measure of applicants who prefer c to all other feasible colleges (i.e., colleges with cutoffs less than their scores). 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 vector P* and a unique stable matching µ*. Given the continuous density assumption, D(·) is C¹ and ∂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 = [η^k, S^k] be

---

11Since we shall assume that the distribution of the types is atomless, the tie-breaking becomes immaterial.
12At 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 (Abdulkadiroğlu, Agarwal, and Pathak, 2017).
13Our full-support assumption is stronger than that of AL. We use the stronger version for Theorem 2.
14Individual rationality requires that no participant (an applicant or a college) receives an unacceptable match, which holds by assumption here. No blocking means that no applicant-college pair exists such that the applicant prefers the college over her 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.
15The uniqueness of the matching under serial dictatorship is obtained directly; see Abdulkadiroğlu, Che, and Yasuda (2015).
a \( k \)-random economy that consists of \( k \) applicants each with type \( \theta \) drawn independently according to \( \eta \), with \( \eta^k \) denoting the resulting empirical measure, 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). A matching is defined in the usual way.

Consider a sequence of \( k \)-random economies \( \{ F^k \} \). Applicants are assigned to colleges using an applicant-proposing DA. We assume that colleges are acting passively and report their preferences and capacities truthfully\(^{16} \) and focus on applicants’ behavior.

In one sense, this task is trivial: since DA is strategy-proof, it is a weakly dominant strategy for each applicant to rank colleges truthfully. We call such a strategy the truth-telling (TT). A weaker notion, weak truth-telling (WTT), allows some suboptimal behavior and is often adopted in empirical research. According to WTT, applicants rank their top \( k \) colleges truthfully, where \( k \) is any integer between 0 and \( C \); unranked colleges are worse than those ranked, but may still be acceptable. However, a more robust solution concept is needed to account for mistakes we observe in the field; we define it below.

We assume that each applicant observes her own type \( \theta \) but not the types of other applicants; as usual, all applicants understand as common knowledge the structure of the game. Given this structure, an applicant’s strategy is a measurable function \( \sigma : \Theta \to \Delta(\mathcal{R}) \), where \( \mathcal{R} \) is the set of all possible ROLs an applicant may submit. For any given \( k \)-random economy, we let \( \sigma^k \) denote a \( k \)-tuple strategy profile \( (\sigma, \ldots, \sigma) \). Note that assuming symmetric strategies is not restrictive, as the probability that two applicants have the same type is zero. When applicants play \( \sigma^k \), they induce a distribution of ROLs, which leads to cutoffs \( P \in [0, 1]^C \) of colleges. 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 refer to the assignment of applicants to colleges as an outcome.

We are interested in the following solution concept:\(^{17} \)

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

Without relaxing the exact Bayesian Nash equilibrium, one cannot explain the departures from the dominant strategy observed in the preceding section. To see why, it is useful to contrast the optimal strategies in the continuum and finite economies. Recall that in the

---

\(^{16}\)In the context of VTAC, the common college preferences make this an ex-post equilibrium strategy (see Che and Koh, 2016).

\(^{17}\)A number of authors adopted a similar \( \epsilon \)-based solution concept to analyze approximate equilibrium behavior (see Kalai (2004), Deb and Kalai (2015), Azevedo and Budish (2018), and Che and Tercieux (2019), for instance).
continuum economy, the cutoffs are degenerate and applicants face no uncertainty about the set of feasible colleges. Hence, dominated strategies, such as ranking only the most preferred feasible college, may do just as well as the dominant truthful strategy. However, such a strategy is not optimal in any finite economy because the cutoffs are not degenerate and any non-truthful strategy results in a payoff loss with a positive probability. Hence, to explain departures from a dominant strategy, one must relax exact optimality of applicant 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 mistakes with payoff losses that vanish as the market grows large.

3.2 Analysis of Robust Equilibria

Does the robustness concept imply that most applicants report their preferences truthfully in a large market? We show below that this need not be the case. Specifically, we construct a robust equilibrium in which all but a vanishing fraction of applicants do not adopt WTT strategies. The adopted strategies need not even respect the true preference order among the colleges that are included in an applicant’s ROL.

To begin, recall the unique market-clearing cutoffs $P^*$ 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 $C = \{1, 2, 3\}$, 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 $1 \leq \ell \leq C$ feasible colleges, then the number of SRSs is $\sum_{a \leq \ell - 1, b \leq C - \ell} (a + b + 1) a! b!$. For each type $\theta = (u, s)$, there exists at least one SRS that violates WTT.\footnote{This can be shown as follows. If an applicant’s most preferred college is infeasible (i.e., its cutoff at $P^*$ is above the applicant’s score at that college), then she can simply drop that college and rank order the remaining colleges truthfully. The resulting strategy is SRS but fails WTT. If an applicant most preferred college is feasible, then she can rank that college at the top of her ROL, but rank the remaining colleges untruthfully (in relative rankings). Again the resulting strategy is SRS but violates WTT.}

For the next result, we construct such a strategy. 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 property that $\hat{r}(\rho, s)$ is an SRS that violates WTT, given the true preference $\rho$ and cutoffs $P^*$. The existence of such a strategy is established above. We then define a strategy $\hat{R} : \Theta \to \mathcal{R}$, given by $\hat{R}(u, s) := \hat{r}(\rho(u), s)$, for all $\theta = (u, s)$.\footnote{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.} 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 $\delta$-close to its market-clearing cutoff in the continuum economy.

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

- all applicants with types $\theta \in \Theta^\delta$ play TT and
- all applicants with types $\theta \notin \Theta^\delta$ randomize between TT (with probability $\gamma$) and the strategy $\hat{R}(\theta)$ that violates WTT (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-TT strategies, vanishes in a large economy. Specifically, the sequence of strategy profiles that we construct satisfies two properties: (i) it prescribes a large fraction of participants to deviate from WTT with a large enough probability, and yet, (ii) it gives rise to the same cutoffs $P^*$ in the limit continuum economy as if all applicants played TT. That these two properties can be satisfied simultaneously is not trivial and requires some care, since the cutoffs may change as applicants deviate from TT. Indeed, the feature that all applicants play TT with some small probability $\gamma$ is designed to ensure that the same unique stable matching obtains under the prescribed strategies. Given these facts, 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 the applicants whose scores are $\delta$ away from $P^*$ will suffer very little payoff loss from playing any SRS that deviates (possibly significantly) from WTT. These are precisely the applicants who deviate from WTT by construction. Since $(\delta, \gamma)$ is arbitrary, the following striking conclusion emerges.

**Corollary 1.** There exists a robust equilibrium in which every applicant does not play WTT with probability arbitrarily close to one.

To the extent that a robust equilibrium is a reasonable solution concept, this 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. Importantly, even among the colleges that an applicant includes in her equilibrium ROL, the order may not respect her true preferences. This result also calls into question any empirical method relying on WTT—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 WTT, do they also undermine the stability of the outcome? This is an important question on two accounts. First, if mistakes jeopardize stability in a significant way, the rationale for DA should be called into question. If stability remains largely intact despite the presence of
mistakes, however, then the presence/prevalence of mistakes may not raise a fundamental concern. Second, stability is also important as an empirical identification assumption, which has been used by a number of authors for preference estimation (see Fox, 2009; Agarwal, 2015; Fox and Bajari, 2013; Chiappori and Salanié, 2016; Fack, Grenet, and He, 2019, for instance). Our second theorem shows that 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.** A strategy $\sigma : \Theta \rightarrow \Delta(\mathcal{R})$ is **asymptotically stable** if the fraction of applicants assigned their most preferred feasible colleges (given the equilibrium cutoffs) in economy $F_k$ under $\sigma^k$ converges in probability to 1 as $k \rightarrow \infty$.

We call a strategy $\sigma$ **regular** if there exists some $\gamma > 0$ such that for each $\theta \in \Theta$, $\sigma(\theta)$ assigns probability at least of $\gamma$ to playing TT.

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

This result is intuitive because robust equilibrium requires that an applicant’s loss vanishes as the economy grow larger. However, a non-vanishing proportion of applicants may still obtain non-stable assignments—and thus asymptotic stability fails—without suffering significant payoff losses in a large economy. To rule out this possibility, one must gain a deeper understanding about the structural properties of the matching arising in a large economy.

The key argument to this end is to demonstrate that in any regular robust equilibrium the uncertainty about colleges’ DA cutoffs vanishes as the economy becomes large. Then, applicants must **virtually** know what the true cutoffs are in the limit as the economy grows large, so 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. Asymptotic stability therefore follows.

The argument is nontrivial because the cutoffs in the $k$-random economy depend on the equilibrium strategies, but our robust equilibrium concept imposes very little structure on these strategies. To prove the argument, we fix an arbitrary regular robust equilibrium strategy $\sigma$ and study the sequence of (random) demand vectors $\{D^k(P)\}$ for any fixed cutoffs.

---

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

21 The regularity of the strategy profile need not always mean that every applicant always report truthfully with some probability. We can “purify” the strategy by define a richer type space $\tilde{\Theta} := \{TT, NTT\} \times \Theta$ such that type $(TT, \theta)$ always adopts TT.
P induced by $\sigma^k$. 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 the law of large numbers: namely, for each vector $P$, $D^k(P)$ converges pointwise to its expectation $\bar{D}^k(P) := \mathbb{E}[D^k(P)]$ 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 $\bar{D}(\cdot)$ is not too elastic, which allows us to prove that cutoffs converge in probability to some degenerate cutoffs $\bar{P}$ along that sub-subsequence as $k \to \infty$.

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

**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.

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

---

22A vector $D^k(P) = (D^k_{c_1}(P), \ldots, D^k_{c_K}(P))$ describes the fractions of applicants who “demand” alternative colleges at vector $P$ of cutoffs given $\sigma^k$ 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 applicant types are random and they may play mixed strategies.

23It is tempting to conclude that this convergence can be obtained from Proposition 3 in Azevedo and Leshno (2016). For instance, one might define, for any arbitrary strategy $\sigma$, a hypothetical economy in which “true” ordinal preferences are ROLs generated by $\sigma$, and use their convergence result for this hypothetical economy. There are two difficulties with this approach. First, AL’s analysis requires a continuum economy with a well-defined measure of types. In our setup, $\sigma$ may involve an arbitrary mixing. Thus, in the hypothetical continuum economy, the measure of types, which now depends on both $\eta$ and $\sigma$, requires a law of large numbers on the continuum economy to be well-defined, which has usual conceptual difficulties (see, e.g., Judd, 1985). We avoid this issue by proving that the outcomes (in terms of demand maps) of a subsequence of random finite economies have a well-defined limit. Second, the convergence-of-cutoffs result that we use is guaranteed in AL with an additional assumption that $\partial D(P^\ast)$ is invertible (see Proposition 3, part 2 of AL). It is not clear, however, that this property holds under an arbitrary regular robust equilibrium strategy $\sigma$. Instead, we derive the property in the limit (this is shown in Claims 7 and 8 in our proof; see Online Appendix B). Finally, even if $D^k$ converges to some $P$ in probability, additional arguments are required to establish asymptotic stability.

24This 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.
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 do in the next section.

Taken together, our two theorems provide very different implications for the behavior and outcome under DA. On the one hand, the behavioral prediction exhibits multiplicity, and a drastic departure from truth-telling is consistent with robust equilibrium. On the other, the prediction in terms of the outcome is virtually unique, and the outcome would be virtually the same as if all applicants reported their preferences truthfully. This latter finding should ultimately be reassuring about the performance of DA and perhaps other strategy-proof mechanisms. Despite the lack of obviousness in the strategy-proofness, and significant departure from the truth-telling behavior, the outcome, which is what one ultimately cares about, need not suffer or differ much from what the theory suggests. At the same time, the first result raises a serious question about taking “expressed” preferences at their face value, when it comes to interpreting the data. As we propose next, the outcome of a mechanism, rather than the expressed preferences of the participants, should provide a more reliable guide for learning true preferences in the presence of strategic mistakes by participants.

4 Empirical Analysis of Data with Mistakes

Building on the theoretical results from the previous section, we develop an empirical method for analyzing data on matching that is robust to participants’ mistakes.

Consider a dataset on a random finite economy $F^k$ in which $k$ applicants apply for admissions to $C$ colleges. They adopt robust equilibrium strategies, $\sigma$. Each college has a finite capacity, and applicants are assigned through the DA mechanism. Besides the submitted ROLs and the outcome, the researcher observes scores, applicant characteristics, and college attributes.

4.1 Inferred Preferences and Robust Stability

In the following, we review two common assumptions from the literature and propose a new one. To aid explanation, we use an example in which 6 colleges select applicants based on a common score. Their cutoffs are described in column (2) of Table 5. We consider applicant $i$ who has a score 0.5, submits ROL $c_3\prec c_1\prec c_4$, and is matched with $c_3$.

- The first assumption from the literature is weak truth-telling (WTT) (Hällsten, 2010; Kirkebøen, 2012). Applicant $i$ is WTT if her submitted ROL orders her most preferred colleges according to her true preferences, while every unranked college is less desirable to $i$ than any ranked college. Let $\succ_i$ denote the inferred preference relation of $i$. In Table 5, WTT infers that $c_3 \succ_i c_1 \succ_i c_4 \succ_i c_2, c_5, c_6$. This inference may be incorrect:
### Table 5: Inferred Preferences of Applicant i with Score s = 0.5—Comparison of Assumptions

<table>
<thead>
<tr>
<th>Submitted ROL</th>
<th>Weak truth-telling</th>
<th>Stability Cutoffs Feasible?</th>
<th>Robust Stability Overestimated Cutoffs Clearly feasible?</th>
</tr>
</thead>
<tbody>
<tr>
<td>First choice: c₃</td>
<td>Most preferred</td>
<td>0.4 √</td>
<td>0.4 √</td>
</tr>
<tr>
<td>Second choice: c₁</td>
<td>Second preferred</td>
<td>0.9 ×</td>
<td>0.9 + ∆₃₁ ×</td>
</tr>
<tr>
<td>Third choice: c₄</td>
<td>Third preferred</td>
<td>0.2 √</td>
<td>0.2 + ∆₄₁ √</td>
</tr>
</tbody>
</table>

Unranked colleges:
- c₂ Dominated by c₄
- c₅ Dominated by c₄
- c₆ Dominated by c₄

Inferred preferences of applicant i: c₃ ≻ᵢ c₁ ≻ᵢ c₄ ≻ᵢ c₂, c₅, c₆

Notes: In this table, i is one of the many applicants in the market. Cutoff overestimation ∆ₖᵢ is a small positive value for c = 1, 2, 4, 6 and can be of the same across these colleges; ≻ᵢ denotes the inferred preference relation of i under a given assumption. i’s score at each college is the same, s = 0.5.

- **c₂** is out of reach for i, hence she may not have ranked c₂ even though she prefers it to some college in her ROL. For a similar reason, i may have ranked c₁ arbitrarily. Consequently, WTT may bias the estimation of her preference over c₁ and c₂.

- The second assumption is **stability** (Akyol and Krishna, 2017; Fack, Grenet, and He, 2019; Bucarey, 2018). An outcome is stable if every applicant is matched with her most-preferred college among the feasible ones (i.e., those with an observed cutoff below her score). In Table 5, only colleges c₃, c₄, c₅, and c₆ are feasible to i; stability infers that c₃ ≻ᵢ c₁, c₄, c₅, c₆. Since c₁ and c₂ are infeasible, stability infers no preferences over them in relationship with c₃ or c₄. Consequently, stability avoids erroneous inference made by WTT. Meanwhile, c₅’s cutoff is exactly i’s score. Even though c₅ is feasible, i may not have anticipated it. For that reason, i may have regarded c₅ out of reach and not ranked it, even though she may prefer it to some college she ranked. In this case, stability makes incorrect inferences of preferences on c₅.

- Our proposed assumption is **robust stability**. An outcome is robustly stable if every applicant is matched with her most-preferred college among those that are “clearly” feasible, i.e., feasible even when she slightly “overestimates” the cutoffs. Formally, given an outcome, a college is **clearly feasible** to applicant i if her score at that college exceeds the college’s (slightly) overestimated cutoff, Pᵢ + ∆ᵢ, where overestimation ∆ᵢ is non-negative and, importantly, ∆ᵢ = 0 if i and c are matched.²⁵ The positive

---

²⁵Note we treat an applicant’s matched college differently from others. This asymmetry serves several purposes. First, by keeping an applicant’s matched college always clearly feasible to her, the preference relations inferred by robust stability are a subset of those implied by stability. This makes the current assumption weaker than the other common assumptions (see Proposition 1). Second, this way of overestimating cutoffs works even when data on ROLs are not available; otherwise, if we also overestimate the matched college’s cutoff, the matched college may become infeasible, which requires the researcher to use the ROLs
Δc_i’s are independent of i’s ROL and can be set at the same value. By construction, this method does not rely on any information on ROLs; in effect, it may shrink an applicant’s feasible set, while always keeping her matched college in the set. In Table 5, we posit that an applicant may overestimate cutoffs as in column (4), and c_5 is not clearly feasible to i. Hence, robust stability infers c_3 \succ_i c_4, c_6. Recall that stability infers c_3 \succ_i c_5, a potential error. By contrast, relying on overestimated cutoffs, robust stability avoids this error by not inferring any preferences regarding c_5.

Let \( O(\text{robust stability}_{\{\Delta c_i\}_{ci}} \mid F^k, \sigma) \) denote the set of all preference relations inferred by robust stability with overestimation \( \{\Delta c_i\}_{ci} \) given economy \( F^k \) and strategy profile \( \sigma^k = (\sigma, ..., \sigma) \). For example, if there were only one applicant in the market described in Table 5, \( O(\text{robust stability}_{\{\Delta c_i\}_{ci}} \mid F^k, \sigma) = \{c_3 \succ_i c_4; c_3 \succ_i c_6\} \). \( O(\text{stability} \mid F^k, \sigma) \) and \( O(\text{WTT} \mid F^k, \sigma) \) are similarly defined for stability and WTT, respectively.

We call an inferred preference relation misclassified if it contradicts the applicant’s true preferences. For example, if in Table 5, i’s true preferences rank c_2 above c_3, WTT leads to a misclassification of c_3 \succ c_2.

**Proposition 1.** Given any \( F^k, \sigma, \) and cutoff overestimation \( \{\Delta c_i\}_{ci}, \)

\[
O(\text{robust stability}_{\{\Delta c_i\}_{ci}} \mid F^k, \sigma) \subseteq O(\text{stability} \mid F^k, \sigma) \subseteq O(\text{WTT} \mid F^k, \sigma).
\]

That is, WTT admits highest number of misclassifications, followed by stability and then robust stability. Moreover, payoff-irrelevant mistakes lead to misclassified preference relations only in \( O(\text{WTT} \mid F^k, \sigma) \), while payoff-relevant ones lead to misclassifications under any of the three assumptions.

In sum, moving from WTT to stability and then to robust stability, we gain robustness to mistakes but utilize less information. Therefore, the estimation based on robust stability will be less efficient than the other two, yet less likely to be biased because it uses fewer misclassified preference relations. One may be concerned with the identification power of the robust stability assumption, which we address in the next subsection.

### 4.2 Identification and Consistency

Identification deals with the limit economy in which stability is satisfied. As exemplified by Table 5, the robust-stability assumption transforms the matching game into a discrete choice model. An applicant is matched with her favorite college in a personalized choice set comprising the colleges that are clearly feasible to her. As long as the personalized choice sets are determined exogenously, the nonparametric identification arguments for discrete choice and all overestimated cutoffs to recalculate a new match for every applicant.
models (Matzkin, 1993) still apply (also see discussions in Agarwal and Somaini (2018); Fack, Grenet, and He (2019)). The exogeneity is plausible, since applicants take cutoffs as given in a large economy and cutoff overestimation is chosen exogenously. Exogeneity of choice sets also requires that an applicant does not “choose” to have low or high scores based on her preference, an assumption likely to hold in practice.

A potential concern is the variation in personalized choice sets across applicants. For example, an applicant’s score may enter her utility functions directly or be correlated with unobserved preference heterogeneity, if, for example, scores are determined by applicant ability. In this case, the robust stability assumption may not reveal low-scoring applicants’ preferences over popular colleges, because such colleges are often infeasible to them. This may lead to a failure of identification with regard to the effect of an applicant’s ability on her preferences. This issue is mitigated if another measure of applicant ability is available, such as GAT in our VTAC data. It is plausible that conditional on applicant ability, scores do not determine preferences and only affect the feasibility of colleges. If, in addition, scores have full support (i.e., can take any possible value) at any value of applicant ability, some low-ability applicants will have all colleges feasible. This restores nonparametric identification. Alternatively, when scores are determined by a lottery as in the experiment of Li (2017), personalized choice sets are exogenously determined and have sufficient variation conditional on applicant characteristics.

When taking the above discussion to data, we inevitably study a finite economy in which the outcome may not be robustly stable. Indeed, our theoretical results show that stability may not be satisfied in a finite economy. This raises the question whether the estimator based on robust stability is consistent. In Appendix C.1, we show that the robust-stability-based estimation is consistent when stability is asymptotically satisfied (Proposition C.1).

4.3 Limitations and Advantages of the Robust-stability-based Estimator

Relative to the estimators based on WTT and stability, the robust-stability-based estimator uses less information on inferred preferences and thus lead to a larger estimation variance. For example, it does not use applicants’ submitted ROLs in the estimation, because even the ranking among the colleges included in an applicant’s ROL in a robust equilibrium may not respect her ordinal preferences. Therefore, we may lose some precision in estimating the substitution patterns when we allow for more flexible random utility models (Berry, Levinsohn, and Pakes, 2004; Abdulkadiroğlu, Agarwal, and Pathak, 2017). Our Monte Carlo simulation results illustrate the bias-variance tradeoff (see, for more details, Section 4.4 and Appendix D). It is therefore important to have a systematic way to select the optimal method, and, in the case of robust stability, the optimal level of cutoff overestimation, which is the purpose of the testing procedure described in Appendix C.2.
Our proposed estimator shares two notable limitations with the estimator based on stability. First, they both require that an applicant’s feasible set of colleges and any determinant of her preferences are conditionally independent, as discussed in Section 4.2. This can be restrictive in some contexts, e.g., test-based admissions. Second, both estimators require that either an applicant can predict her feasible set with a reasonable accuracy or can rank as many colleges as possible without any costs. In this regard, the robust-stability-based estimator is less demanding, as it can tolerate some prediction errors.

On the other hand, the estimator based on robust stability enjoys several advantages. First, it does not require data on the ROLs submitted by applicants. As long as the researcher observes an applicant’s score and assigned college, as well as each college’s cutoff, she can construct the set of colleges that are clearly feasible to the applicant given a cutoff overestimation. Second, since it transforms the matching problem into a discrete choice problem, one may obtain a consistent estimator even with a random sample of applicants. Lastly, the nesting structure in Proposition 1 leads to a statistical test, described in Appendix C.2, which can help us choose the optimal method, or the optimal cutoff overestimation if necessary, that balances bias and variance in the estimation even in the presence of strategic mistakes.

4.4 Monte Carlo Simulations: Summary

To further assess the implications of our theoretical results, Appendix D presents an extensive set of Monte Carlo simulations. The simulated environment has 1800 applicants competing for admissions to 12 colleges. Applicant preferences follow a logit random utility model, and colleges rank applicants by an ex ante known score. We consider three types of behaviors in the data generating process by introducing various mistakes. The first is truth-telling: every applicant truthfully ranks all colleges. The second type is payoff-irrelevant skips: a fraction of applicants skip colleges with which they are never matched according to a simulated distribution of cutoffs. The last type is payoff-relevant mistakes: in addition to skipping never-matched colleges, some applicants make payoff-relevant mistakes.

We then apply the three types of estimators (WTT, stability, and robust stability) to the simulated data. The results display several noticeable patterns. First, WTT leads to an inconsistent estimator whenever some applicants are not truthful. In particular, it systematically underestimates the average quality of popular or small colleges, as it incorrectly regards them as inferior whenever an applicant chooses not to apply to these out-of-reach colleges. Second, stability performs well when there are no payoff-relevant mistakes but suffers from inconsistency when there are payoff-relevant mistakes. Third, the robust-stability-based estimator is consistent whenever there are no payoff-relevant mistakes; moreover, even when there are some payoff-relevant mistakes, it successfully reduces biases.

We further consider a counterfactual analysis of an affirmative-action policy. Measured
by how close the estimation results are to the truth, the estimates based on robust stability
dominate WTT and stability, especially when there are payoff-relevant mistakes.

Finally, we consider a common method for counterfactual analysis based on submitted
ROLs. That approach assumes that the submitted ROLs under the existing policy are true
ordinal preferences and that applicants will submit the same ROLs after the existing policy
has been replaced by the counterfactual. Our simulation results show that this approach is
highly vulnerable to both types of mistakes, payoff-relevant and -irrelevant, even when the
mechanisms are strategy-proof.

5 Analysis of Experimental Data from Li (2017)
We evaluate the above three empirical methods in Li (2017)’s experimental data. This is an
ideal testbed. First, the true payoffs are known in Li’s data. Hence, we can measure the
goodness of fit associated with the competing methods directly against the truth. This would
not be possible if we were to use observational data such as VTAC’s. Second, the choice
environment is much simpler in Li’s experiment than in VTAC, which helps us keep this
paper concise and on point.

His experiment includes two versions of the Serial Dictatorship; one is strategy-proof,
while the other is obviously strategy-proof. The former is a special case of our setting, and
we use it to test our estimation methods. We use the latter for the counterfactual analysis in
Section 5.3.

For a given game in the experiment, there are four subjects competing for four distinct
cash prizes drawn from \{$1.25, $1, $0.75, $0.5, $0.25, $0\}. In the following, we use “colleges”
to refer to the cash prizes and call subjects in the experiment “applicants.” Each college has
one seat and is worth the same cash value for all applicants. Applicant scores are private
information and drawn uniformly from \{1, 2, ..., 10\}. The discrete uniform distribution is
common knowledge. A serial dictatorship is implemented in each game as follows: After
observing her own score, every applicant is required to submit an ROL of all four available
colleges; the applicant with the highest score is assigned the top-ranked college in her
submitted ROL, then the one with the second highest score is assigned the highest-ranked
college in her submitted ROL that is still available, and so on. In this process, ties in scores
are broken randomly.

The dataset is ideal for testing our empirical approaches, because 29% of submitted ROLs
contain mistakes despite the strategy-proofness of the serial dictatorship. We perform the
same analyses as in Section 2 and reach similar conclusions. In particular, the majority of
the mistakes (64%) are payoff irrelevant.\textsuperscript{26}

Since applicant preferences and mistakes are known in the experimental data, we can
\textsuperscript{26}Hassidim, Romm, and Shorrer (2016) perform similar analysis of Li’s data; however, their focus is not
on the estimation of applicant preferences or on how misreporting affects empirical analysis.
assess the implications of our theoretical results. In particular, we demonstrate that the WTT-based estimator is biased, while the estimator based on robust stability is always less vulnerable to mistakes. More importantly, in the dataset that is plagued by strategic mistakes, robust stability still manages to extract reliable information for empirical analysis.

In the following, Section 5.1 describes the experimental data and assesses the assumptions; Section 5.2 specifies an econometric model and discusses the estimation results; and, finally, Section 5.3 presents a counterfactual analysis.

5.1 Data and Assessment of the Assumptions

Each applicant plays 10 rounds. In each round, she is grouped with three randomly chosen applicants to play the game with newly drawn colleges and scores.

To investigate how mistakes affect estimation, we divide the data into two subsamples: early rounds (rounds 1-5) and late rounds (rounds 6-10). In each subsample, there are 72 applicants, and we observe 5 game plays of each applicant. Among all the individual game plays, there are more mistakes in the early rounds (31%) than in the late rounds (27%). The same comparison holds for payoff-relevant mistakes (13% versus 8%). These numbers are comparable to those in the VTAC data, where 34.6% make mistakes, with 1.4–20.0% being payoff relevant (Table 1).

Table 6: Assessment of the Assumptions: Early versus Late Rounds

<table>
<thead>
<tr>
<th></th>
<th>Early Rounds (Rounds 1–5)</th>
<th>Late Rounds (Rounds 6–10)</th>
</tr>
</thead>
<tbody>
<tr>
<td># individual actions</td>
<td>360 (72 applicants × 5 rounds)</td>
<td>360 (72 applicants × 5 rounds)</td>
</tr>
<tr>
<td><strong>Panel A. WTT of applicant action</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A submitted ROL is truth-telling.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency (Percent)</td>
<td>249 (69%)</td>
<td>262 (72%)</td>
</tr>
<tr>
<td><strong>Panel B. Stability of outcome (w.r.t. observed cutoffs)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An applicant’s observed outcome is her favorite feasible college (i.e., no payoff-relevant mistake).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency (Percent)</td>
<td>312 (87%)</td>
<td>331 (92%)</td>
</tr>
<tr>
<td><strong>Panel C. Robust stability of outcome (w.r.t. overestimated cutoffs)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An applicant’s observed outcome is her favorite “clearly” feasible college.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency (Percent)</td>
<td>320 (89%)</td>
<td>335 (93%)</td>
</tr>
</tbody>
</table>

Notes: The dataset is from Li (2017). WTT is equivalent to “strict truth-telling” in this dataset, as every applicant is required to rank all colleges. For robust stability, we add 0.5 to the observed cutoffs.

We assess the three assumptions for estimation: WTT, stability, and robust stability. Note that WTT is equivalent to strict truth-telling in this dataset, since every applicant is required to rank all colleges (as required by the experimental design). Panel A of Table 6 shows that WTT is satisfied in 69% of the observations in the early rounds and 72% in the late rounds. This implies that the WTT-based estimator is likely to be inconsistent, as it relies on a
substantial fraction of misclassified preference relations.

The stability assumption is evaluated in Panel B of the table. Relative to WTT, stability is satisfied by many more observations. In the early rounds, 87% of the applicants are matched with their favorite feasible college; this fraction increases to 92% in the late rounds. This indicates that the stability-based estimator can still be biased, especially in early rounds, albeit likely less biased than the WTT-based estimator.

Lastly, Panel C of Table 6 assesses robust stability. By adding a small overestimation, 0.5, we correct some misclassifications of inferred preferences. The fraction of applicants who are matched with their favorite colleges among the clearly feasible ones increases to 89% in the early rounds and 93% in the late rounds. One may increase the overestimation and obtain a higher rate of correctly inferred preferences, and our choice of a small magnitude illustrates that correcting a small fraction of misclassification can already improve the estimation. Particularly, in the early rounds, although the change is merely 2 percentage points, from 87% under stability to 89% under robust stability, our estimation results in the next section show a substantial reduction in estimation bias.

5.2 Econometric Model and Estimation Results

We approximate applicant preferences by a random utility model. The six colleges are ordered by their cash value from the highest to the lowest. That is, $c = 1$ is of $1.25$ and $c = 6$ is of $0$. Applicant $i$’s utility for college $c$ is

$$u_{i,c} = \alpha_c + \beta_c \times Female_i + \epsilon_{i,c}, \text{ for } c = 1 \ldots 5;$$

$$u_{i,6} = \epsilon_{i,6};$$

where $\alpha_c$ is college $c$’s fixed effect, or average quality, for males; $\alpha_c + \beta_c$ is $c$’s fixed effect for females; and $\epsilon_{i,c}$ is an idiosyncratic shock and a type-I extreme value. Table 7 presents the estimates of $\alpha$’s and $\beta$’s under the three assumptions separately for the two subsamples.

Recall that we know the true cash value of each college, so we can test how well different methods predict choice probabilities based on the true cash values. Taking $c = 3$ ($0.75$) as the benchmark, we calculate the choice probability of $c'$ when college 3 and $c'$ are the

---

27Recall that the gap between any two different scores is at least one. The overestimation, 0.5, only makes a difference when an applicant loses a college to another applicant with the same score.

28Given the common ordinal preferences of the applicants and the to-be-defined random utility model, (almost) 100% correctly inferred preferences imply that there will be no variation in applicant choices (e.g., college 1 of value $1.25$ is always chosen whenever it is available), and this will lead to non-identification of the random-utility model. It should be noted that this may not happen in real-life college admissions or school choice, as applicant preferences are more heterogeneous. In real-life settings, one may apply the testing procedure in Appendix C.2 to select the value of overestimation.

29As robustness checks, we include other controls such as an applicant’s GPA, whether the applicant is in a STEM major, whether she is an economics major or has taken microeconomics or game theory course, as well as some interactions. The results are qualitatively the same as those from the model specified in equation (5).
Table 7: Estimation with the Experimental Data under the Three Assumptions

<table>
<thead>
<tr>
<th></th>
<th>Early Rounds (Rounds 1–5)</th>
<th>Late Rounds (Rounds 6–10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) WTT</td>
<td>(2) Stability</td>
</tr>
<tr>
<td></td>
<td>(4) WTT</td>
<td>(5) Stability</td>
</tr>
<tr>
<td>College 1’s FE for males: $\alpha_1$</td>
<td>5.17***</td>
<td>6.11***</td>
</tr>
<tr>
<td></td>
<td>(0.79)</td>
<td>(0.98)</td>
</tr>
<tr>
<td></td>
<td>College 2’s FE for males: $\alpha_2$</td>
<td>4.66***</td>
</tr>
<tr>
<td></td>
<td>(0.60)</td>
<td>(0.80)</td>
</tr>
<tr>
<td></td>
<td>College 3’s FE for males: $\alpha_3$</td>
<td>3.70***</td>
</tr>
<tr>
<td></td>
<td>(0.54)</td>
<td>(0.84)</td>
</tr>
<tr>
<td></td>
<td>College 4’s FE for males: $\alpha_4$</td>
<td>3.03***</td>
</tr>
<tr>
<td></td>
<td>(0.46)</td>
<td>(0.71)</td>
</tr>
<tr>
<td></td>
<td>College 5’s FE for males: $\alpha_5$</td>
<td>1.94***</td>
</tr>
<tr>
<td></td>
<td>(0.33)</td>
<td>(0.60)</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>-2.14**</td>
</tr>
<tr>
<td></td>
<td>(0.99)</td>
<td>(1.20)</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>-1.99***</td>
</tr>
<tr>
<td></td>
<td>(0.77)</td>
<td>(0.98)</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>-1.60**</td>
</tr>
<tr>
<td></td>
<td>(0.67)</td>
<td>(1.01)</td>
</tr>
<tr>
<td></td>
<td>$\beta_4$</td>
<td>-1.36**</td>
</tr>
<tr>
<td></td>
<td>(0.54)</td>
<td>(0.85)</td>
</tr>
<tr>
<td></td>
<td>$\beta_5$</td>
<td>-1.01**</td>
</tr>
<tr>
<td></td>
<td>(0.39)</td>
<td>(0.77)</td>
</tr>
<tr>
<td># Applicants</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td># Individual actions</td>
<td>360</td>
<td>360</td>
</tr>
<tr>
<td></td>
<td>360</td>
<td>360</td>
</tr>
</tbody>
</table>

Notes: This table presents the estimates of the model, $u_{i,c} = \alpha_c + \beta_c \times Female_i + \epsilon_{i,c}$, for $c = 1 \ldots 5$, (equation 5). Columns 1-3 use data from the early rounds, while columns 4-6 use the late rounds. Each subsample is estimated under the three assumptions (WTT, stability, and robust stability). Standard errors clustered at the applicant level are in parentheses. * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

only available colleges. Among the male applicants, the choice probability for $c'$ is simply $\exp (\alpha_{c'}) / \left[ \exp (\alpha_{c'}) + \exp (\alpha_3) \right]$. If the estimation is perfect, the choice probability (against college 3 or $0.75$ college) should be zero if college $c'$ is of a lesser value, and one if $c'$ is more valuable.

Figure 1 shows the sources of biases under each assumption. In the early rounds (Figure 1a), the WTT-based estimates substantially understate the preferences for top colleges. Against the $0.75$ college, the $1.25$ college is only chosen about 81% of the times according to the WTT-based estimates. The reason is that when some low-scoring applicants consider the $1.25$ college out of reach and do not rank it highly, WTT incorrectly classifies the college as less-preferred for them. Such misclassification is reduced under stability and robust stability: the estimated choice probability increases to 94% under the former, and to 97% under the latter. This shows the advantage of the estimator based on robust stability when payoff-relevant mistakes are a concern. Similar patterns are observed for the $1$ college. The differences among the three assumptions are not as pronounced for the low-valued colleges, although the robust-stability-based estimates are the closest to the true value, zero.

Figure 1b shows the results from the late rounds. WTT still seriously mis-predicts, while
Figure 1: Predicted Choice Probabilities based on WTT, Stability, and Robust Stability

Notes: The figures present the predicted choice probabilities for male applicants based on the estimates from three methods, weakly truth-telling (WTT), stability, and robust stability. Each choice probability is the predicted likelihood of a male applicant choosing a given college when that college and the $0.75 college are the only options available. If predicted perfectly, the colleges more valuable than $0.75 should have a choice probability equal to one, while those worth less than $0.75 should never be chosen. We assume that the choice probability of a $0.75 college is 0.5 when there is another $0.75 college. Subfigure (a) is for the early rounds (rounds 1-5), while subfigure (b) is for the late rounds (rounds 6-10). The results for females display similar patterns.

stability and robust stability almost coincide, as payoff-relevant mistakes become rarer.

We further evaluate the assumptions by their model fit. First, we use the estimated utility functions to calculate the probability that an applicant in a given game ranks the four available colleges according to their monetary values. Figure 2a shows that the WTT-based estimates perform the worst in both the early and the late rounds, while robust stability performs the best. Stability dominates WTT, but it leads to a worse fit than robust stability in the early rounds when there are non-negligible payoff-relevant mistakes.

Figure 2b shows that similar comparison holds with the model fit in terms of predicting individual outcomes. Together with the parameter estimates, we draw 10,000 sets of type I extreme values as inputs into equation (5) and let applicants re-play each game by submitting the simulated preference order; we then compare the simulated outcome of each applicant in each game to the observed one. The figure presents the rate of correct prediction averaged across applicants and simulated games. The robust-stability-based estimates can correctly predict individual outcomes at a rate of 56% for the early rounds and 69% for the late rounds, despite the mistakes and the model parsimony. As a comparison, a uniformly random prediction obtains a rate of 25%.

Taken together, WTT leads to the worst model fit to the data, while robust stability can significantly improve upon stability when payoff-relevant mistakes are more frequent.

It is also worth noting that useful information is extracted from the data under all assumptions, even though a significant fraction of applicants make mistakes. For example, the worst-performing assumption, WTT, can lead to a 34% rate of correct prediction of a
Figure 2: Predicting True Preference Order and Outcomes based on WTT, Stability, and Robust Stability

Notes: The figures show the in-sample fit of the three methods, weak truth-telling (WTT), stability, and robust stability. To obtain subfigure (a), for a given game, we use the estimates to calculate the likelihood that an applicant’s preferences over the four colleges in the game coincide with the order of their cash values; we then average these likelihoods across applicants and games. For subfigure (b), we draw 10,000 sets of type I extreme values as inputs into equation (5) and let applicants re-play each game by submitting the simulated preference order. Comparing the simulated outcome of each applicant in each game to the observed one, the figure presents the average rate of correct prediction over all applicants and simulated games.

preference order. By comparison, a uniformly random preference order for an applicant leads to a $1/24$, or 4%, rate of correct prediction. This implies that empirical analyses of matching data should not be precluded by potential mistakes.

5.3 Counterfactual Analysis

One of the important objectives of empirical market design is to conduct counterfactual analysis, for example, under a new policy or a changed environment. We take advantage of the presence of another mechanism in Li’s experiment, namely, the sequential serial dictatorship. Similar to the game described above, four applicants compete for four colleges. Each of them is given a score, and in the descending order of score, each applicant chooses among the remaining colleges. They are not required to submit an ROL. Notably, this mechanism is obviously strategy-proof, so applicants make fewer mistakes.

With the estimates from the previous game, we predict outcomes of this new game and compare our prediction to the observed outcomes. We draw 10,000 sets of type I extreme values as inputs into equation (5) and let applicants re-play each game by choosing the most-preferred college among the remaining ones according to the simulated preference order. Comparing the simulated outcome of each applicant in each game to the observed one, Figure 3 shows the average rate of correct prediction.

Robust stability still outperforms the other two, especially in the early rounds. The robust-stability-based estimates can correctly predict individual outcomes at the rate of 68% for the early rounds and 76% for the late rounds. These rates are substantially higher than
Figure 3: Predicting Counterfactual Outcomes based on WTT, Stability, and Robust Stability

Notes: The figure presents the out-of-sample goodness of fit of the estimates from three approaches, weakly truth-telling (WTT), stability, and robust stability. We draw 10,000 sets of type I extreme values as inputs into equation (5) and let applicants re-play each game by choosing the most-preferred college among the remaining ones according to the simulated preference order. We then compare the simulated outcome of each applicant in each game to the observed one, and the figure presents the average rate of correct prediction over all applicants and simulated games.

In sum, the above analysis demonstrates three points: (i) because there are many mistakes, outcome-based analyses work better than action-based ones; (ii) our robust-stability-based estimator works especially well when there are payoff-relevant mistakes; and (iii) although mistakes can be non-negligible in the data, robust stability can still extract reliable information in the data to conduct meaningful empirical analysis.

6 Related Literature


The setup of our theoretical model builds upon Azevedo and Leshno (2016), which studies asymptotics of cutoffs of stable matchings. Our results differ from theirs in that Azevedo and Leshno perform price-theory analysis of stable matchings without a game-theoretic framework. We, in contrast, allow much richer game-theoretical behavior, making our results more relevant to the analysis of the real-life markets where applicants’ mistakes are common. This richer setup makes AL results inapplicable and requires a different proof strategy (see the discussion after Theorem 2, especially footnote 23). Technically, our main theoretical result (Theorem 2) has the same flavor as Deb and Kalai (2015), whose Theorem 2 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. This result would lead to the same conclusion as our Theorem 2, since most of the applicants would be matched with their favorite feasible college. Nevertheless, their theorem is not applicable in our setting. Specifically, a crucial condition needed for their result—LC2, which requires the effect that any participant can unilaterally have on an opponent’s payoff to be uniformly bounded and decreases with the number of participants in the game—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.

In terms of the empirical message, the paper closest to ours is Fack, Grenet, and He (2019) (FGH, hereafter). FGH argue that applicants may not submit truthful ROLs due to positive application costs. While useful in explaining some departures from truth-telling, FGH do not explain many others documented in the literature. First, the cost of adding the RT version of a course is plausibly zero in the Australian college admissions, given its FT version is already included. Second, excluding a course previously included in an ROL – a time trend we report in Section 2 – cannot be explained by such costs. Third, several other papers (Hassidim, Romm, and Shorrer, 2016; Li, 2017) report that applicants rank less-preferred colleges above more-preferred ones in their ROLs, which is a dominated strategy inconsistent with FGH’s model. In our paper, non-truthfulness is a suboptimal behavior with negligible consequences; our model is more permissive and is able to accommodate all the documented mistakes, because Theorem 1 imposes relatively little restrictions on applicant behavior in a large economy. Despite the multiplicity of equilibrium behaviors, we derive a sharper prediction regarding the equilibrium outcome: Theorem 2 shows that every robust equilibrium leads to asymptotic stability. In contrast, FGH show that there exists one such sequence of equilibria. This is an important difference because a mere existence does not guarantee that such an outcome would be observed in the field. Our uniqueness result (Theorem 2) provides a strong justification for using (asymptotic) stability for estimation. The proof of the theorem is also more challenging than in FGH and uses different techniques.

Additionally, our paper differs from FGH in the proposed empirical method. Payoff-relevant mistakes are observed in our dataset; our theory also suggests that such mistakes may be present in any finite economy. Building upon this observation, we propose an estimator based on robust stability. This estimator is more robust to payoff-relevant mistakes, a task that FGH’s method cannot accomplish. As illustrated in the analysis of the experimental data from Li (2017), our estimator can accommodate payoff-relevant mistakes and is less biased than the stability-based estimator.
Our proposed estimator can be readily applied to real-life markets where applicants know their own scores before application. The markets that satisfy our assumptions, yet have been analyzed under the assumption of truth-telling, include the National Resident Matching Program (Roth and Peranson, 1999), teacher assignment in France (Combe, Tercieux, and Terrier, 2016), kindergarten allocation in Estonia (Veski, Biró, Pöder, and Lauri, 2017), as well as college admissions in Sweden (Hällsten, 2010), Norway (Kirkebøen, 2012) and Ontario, Canada (Drewes and Michael, 2006). Our paper suggests that their approach may entail a bias and their conclusions from counterfactual analysis may be misleading.\footnote{These papers differ significantly in the exact version of the truth-telling assumption that is adopted. For example, some assume that exactly the same ROL will be submitted by an applicant under a counterfactual policy; some exclude certain infeasible colleges from each applicant’s choice sets. The magnitude of the bias will also depend on the counterfactual policy considered. In our simulations (Appendix D), we evaluate an affirmative action policy that prioritizes some applicants. By contrast, Roth and Peranson (1999) and Combe, Tercieux, and Terrier (2016) evaluate alternative mechanisms without changing the priority structure. One may argue that the resulting bias can be smaller in these two papers compared to the bias we report.} There is, however, a strand of literature that is beyond the scope of the present paper, which analyzes markets with priorities induced by lotteries, such as school choice in New York (Abdulkadiroğlu, Pathak, and Roth, 2009; Abdulkadiroğlu, Agarwal, and Pathak, 2017; Che and Tercieux, 2019). 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 feasibility of a college (at the time of application) may not be degenerate in a large economy as in our model.

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, forthcoming), Beijing (He, 2017), Boston, MA (Abdulkadiroglu, Pathak, Roth, and Sonmez, 2006), Cambridge, MA (Agarwal and Somaini, 2018), New Haven, CT (Kapor, Neilson, and Zimmerman, 2016), Seoul (Hwang, 2017), and Wake County, NC (Dur, Hammond, and Morrill, 2018). Their empirical approaches to identifying mistakes are different from ours, because without detailed information on participants’ true preferences, it often requires more assumptions to identify those who play a sub-optimal strategy under a non-strategy-proof mechanism.

Several studies proposed explanations for the mistakes reported in our paper and in the literature. In addition to Li (2017), who argues that strategy-proof mechanisms can be too difficult for applicants to understand, Dreyfuss, Heffetz, and Rabin (2019) show that the documented mistakes can be optimal for a loss-averse applicant. Unlike these papers, our model is agnostic about the causes of such behavior, but develops an equilibrium concept that accommodates it. Similar to our study, Dreyfuss, Heffetz, and Rabin (2019) use Li (2017)’s experimental data to estimate participants’ preferences. However, their goal is to
quantify the degree of loss aversion that can explain the observed mistakes, assuming the researcher’s knowledge of the exact value of each college. Our purpose is to estimate the values of colleges, as perceived by applicants, in the presence of strategic mistakes (which could well have arisen from loss aversion).

7 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 argue 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. While this result justifies the vast theoretical literature that assumes truthful reporting behavior to analyze outcomes of strategy-proof mechanisms, it calls into question empirical methods that take truthful reporting as a literal behavioral prediction.

Our Monte Carlo analysis and empirical analysis of Li (2017)’s experimental data indeed reveal that the empirical method based on the truth-telling assumption leads to a biased estimation 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 robust stability proves more robust to these mistakes. We further show that a counterfactual analysis based on robust stability yields more accurate predictions than the one based on truth-telling. These results reassure us that one can still use data on matching markets to conduct reliable empirical studies, even if participants make mistakes.
References


