PEER PREFERENCES IN CENTRALIZED SCHOOL CHOICE MARKETS: THEORY AND EVIDENCE*

Natalie Cox† Ricardo Fonseca‡ Bobak Pakzad-Hurson§ Matthew Pecenco¶

First Draft: April 2021 This Draft: May 2025

Abstract

School-choice clearinghouses instruct applicants to "rank preferences truthfully" without allowing them to express preferences over peers. Empirically, we show college applicants value relative peer ability using data from Australia. Theoretically, we prove stable matchings exist under mild conditions when peer preferences are included, yet standard assignment mechanisms typically fail to find them. The status-quo procedure frequently employed by clearinghouses, which reveals prior cohort peer information, inadvertently induces a tâtonnement process that shapes peer expectations and fuels instability—an outcome we confirm empirically. To resolve this, we introduce a modified assignment mechanism that consistently ensures stability and incentivizes truth-telling.

---

*For helpful comments we thank Atila Abdulkadiroglu, Mohammad Akbarpour, Eduardo Azevedo, Oğuzhan Çelebi, Yan Chen, Aram Grigoryan, Guillaume Haeringer, YingHua He, Richard Holden, Clemence Idoux, Adam Kapor, Fuhito Kojima, Maciej Kotowski, Jacob Leshno, Shengwu Li, Margaux Luflade, George Mailath, Ellen Muir, Paul Milgrom, Samuel Norris, Parag Pathak, Alex Rees-Jones, Al Roth, Tayfun Sönmez, Ran Shorrer, Satoru Takahashi, Utku Ünver, Rakesh Vohra, Bumin Yenmez, and seminar audience members at Baruch College, Boston College, Brown, CMU/Pitt, Penn, MSR New England, ITAM, Stanford, UCSB, University of Tokyo Market Design Center, UNSW, VIVE, Wharton, EAMMO, EC, Matching in Practice Workshop, SITE (Market Design), and NBER Summer Institute (Education). We are grateful to Camilla Adams, Clemens Lehner, Sergio Nascimento, Joanna Tasmin, and Jiayue Zhang for excellent research assistance.

This paper subsumes"Do Peer Preferences Matter in School Choice Market Design? Theory and Evidence" which appeared as an extended abstract at EC '22.

†Princeton University, Bendheim Center for Finance, 20 Washington Rd, Princeton, NJ 08540. Email: nbachas@princeton.edu

‡Departamento de Economía, Pontificia Universidad Javeriana, Bogotá, Colombia. Email: ricardobafonseca@gmail.com

§Brown University, 64 Waterman Street, Providence, RI 02912. Email: bph@brown.edu

¶Brown University, 64 Waterman Street, Providence, RI 02912. Email: matthew_pecenco@brown.edu

I Introduction

Centralized matching mechanisms are used to allocate seats at schools and colleges around the world (Neilson, 2019). Creating a stable matching—one in which no agent wants to "block" the matching by deviating with a willing partner—is often viewed as a chief concern in these settings (Roth, 2002). Student preferences over educational programs may depend on a variety of factors, including both predetermined characteristics such as location, and the endogenous-to-the-matching characteristics of peers in their cohort. However, matching mechanisms used in school choice markets are designed to create a stable matching under the assumption that student preferences do not depend on the characteristics of their peers.

In this paper, we address four questions essential to understanding the role of peer preferences in contemporary school choice markets: Do students have peer preferences? Do stable matchings exist when students have peer preferences? Do status quo matching markets deliver stable matchings and what are the consequences if not? Do better mechanisms exist in the face of peer preferences?

We study these questions theoretically and empirically, uncovering clear evidence that peer preferences significantly shape student decisions. Importantly, we find that stable matchings exist under realistic conditions, but existing matching procedures fail to ensure stability because they neglect peer-dependent preferences. To address this gap, we develop a novel mechanism that explicitly accounts for peer preferences, consistently achieves near-stability, and improves fairness by reducing the scope for strategic manipulation. Thus, by recognizing peer effects and explicitly integrating them into mechanism design, we provide a robust solution for centralized school-choice markets.

The centralized admissions market in New South Wales (NSW), Australia's largest state, provides an ideal context to examine peer preferences and stability in matching markets. As is typical in school-choice settings, each student submits a Rank Order List (ROL) indicating their preferred college-subject programs, and programs submit rankings over students. NSW generates final assignments using the student-proposing deferred acceptance algorithm of Gale and Shapley (1962), which is well-known to produce a stable matching when peer characteristics do not influence preferences. Critically, the NSW clearinghouse explicitly displays peer-related information—specifically, a summary statistic of standardized test scores of the previous year's cohort—and advises students to use this information "as a guide when deciding on [their] preferences."[1] This transparency makes NSW particularly suitable for investigating whether and how

[1]This practice is common worldwide. For example, *U.S. News and World Report* annually publishes standardized test scores of the prior year's entering class at U.S. universities, and popular guidebooks in China's college admissions market similarly publish information on the test scores of previous entering classes.

peer characteristics shape students' preferences, as well as evaluating the implications of stability. This transparency also informs our theoretical analysis by demonstrating that the mechanism design question of how best to match students to programs interacts with the broader market design question of how students' information about payoff-relevant peer characteristics is aggregated over time. Consequently, the status quo matching problem is best studied as a dynamic process.

The initial step in our investigation of peer preferences answers the fundamental question: *Do students have preferences over peer ability?* Our primary research design exploits a unique feature of the NSW market—students submit initial ROLs before learning their own scores and have the opportunity to revise them afterward. Observing multiple ROLs from the same student enables us to identify changes in their rankings. By focusing specifically on "switches"—inversions in the relative ranking of two programs—we relax the typical assumptions of truthtelling prevalent in the literature. Instead, we require only the weak assumption that students do not play dominated strategies. We directly test how a student's revealed score, relative to prospective peers, influences their ranking decisions. We identify an asymmetric preference pattern: students prefer programs where peer scores are somewhat below their own, while demonstrating markedly reduced interest in programs with peers scoring above them. We further confirm peer preferences using an event-study approach, examining how student demand responds to changes in publicly observable peer characteristics. Finally, we directly evaluate and rule out alternative explanations for these patterns, such as strategic responses to admissions probabilities, by examining discontinuities in these probabilities.

Guided by our empirical evidence that students do indeed have peer preferences, we ask: *Does a stable matching exist in the presence of peer preferences?* To study this question, we construct a matching model with a continuum of students and finitely many programs as in Abdulkadiroğlu et al. (2015) and Azevedo and Leshno (2016). The presence of many students and a relatively small number of programs in NSW comports with these modeling choices. We depart from standard models by assuming that student preferences depend on the distribution of peer abilities at each program. We allow these preferences to be arbitrary, encompassing cases in which, for example, students wish to attend programs that: enroll high-ability peers, low-ability peers, or peers of similar ability. Our analysis extends in a straightforward way to student preferences over the distribution of other peer characteristics.

While a market designer may have specific desires (e.g., to maximize value added), an axiomatic characterization provides guidance on incorporating peer preferences. We show three desirable matching properties: individual rationality, non-wastefulness, and fairness (Balinski and Sönmez, 1999) are jointly identical in our setting to (pairwise) stability *taking into account students'*

preferences over programs given the distribution of peers. This characterization leads us to take a positive rather than normative view of peer preferences, following a long-standing tradition of the market design literature (Roth, 2002; Abdulkadiroğlu and Sönmez, 2003).[2] As in an equilibrium of a club good economy (see e.g., Ellickson et al. (1999)), a stable matching is endogenously supported by the set of students at each program. We show that a stable matching exists under a mild condition: a sufficiently small change in the matching changes the ordinal preferences of a small fraction of students. Unlike standard large market matching models, the set of stable matchings is not generally a singleton in our model.

Guided by evidence that both peer preferences and stable matchings exist, we ask: *Do "status quo" matching markets deliver a stable matching?* We first show theoretically that canonical static mechanisms—including deferred acceptance—are unlikely to result in a stable matching without correctly specified beliefs about peer types.

Therefore, our second (and primary) analysis of status quo matching markets studies the evolution of beliefs in a discrete-time dynamic process in which students observe the distribution of student abilities at each program in the previous cohort and then submit an ROL to a centralized matchmaker who generates a stable matching with respect to the ROLs. This market forms a discrete-time process similar to a tâtonnement process in exchange economies, where the distribution of student abilities serves the role of "prices," and students best respond to the previous period's "prices."[3] Unfortunately, we show that the status quo procedure is not guaranteed to produce a stable matching, even though a nominally stable matching mechanism is used. This failure occurs because the tâtonnement process can cycle, as in Scarf (1960), due to a failure of a gross-substitutes-like condition, i.e. the distribution of peer abilities never converges.[4] Moreover, due to unit demand and student preferences over peers, the failure of the gross-substitutes-like condition can easily occur; we

[2] Stability is also a desirable property from a market efficiency standpoint, as it may lead to lower attrition rates. We discuss this point later in the paper.

[3] Best responding to the previous period's distribution is analogous to the Cournot updating procedure in exchange economies. Additionally, as Berger (2007) remarks, the simultaneous decisions made within cohort are indeed a variant of the original fictitious play framework proposed by Brown (1951). Experimental evidence further supports our assumption. In a dynamic experimental matching setting wherein students learn about the existing matching before submitting their own preferences, Dur et al. (2021) find that students best respond given the information of previous movers 84% of the time.

[4] This cyclic pattern of "great and small years" in which programs alternate between having more and less competitive student bodies has been previously observed in China's college admissions market, which operates similarly to the NSW market. Specifically, students are told, "if the university has a history of great and small years, you should pay particular attention to this cyclic factor" when submitting ROLs (p. 210 Qiu and Zhao, 2007). We are indebted to Yan Chen for this reference, and for the translation from Mandarin. As we discuss further below, we also find non-convergence in the NSW market.

show that even a social planner who can either change peer preferences freely, or change the market structure freely, cannot guarantee stability. Taken together, our results find that the status quo procedure can generate a matching that is far from stable in all time periods in nearly any matching market.

Our theoretical results provide a simple tool for an observer to ex-post judge whether a sequence of matchings converges to stability in the status quo process: the distribution of student abilities at each program is (approximately) in steady state if and only if the market creates a (approximately) stable matching. We empirically show that the NSW market fails this test in every year, meaning that the market never yields a stable matching.

To understand the empirical implications of instability, we link program-years with higher predicted instability to student attrition. Using a panel fixed effects approach, we find that an increase in the realized vs presented peer ability leads to a reduction in the completion rate. Specifically, a one standard deviation increase in this potential belief difference translates to a 0.07 standard deviation decrease in completion, and these effects are robust to the inclusion of differential time trends across fields of study, accounting for changing labor market trends. Therefore, the status quo instability from peer preferences results in Pareto losses either from student "transfer costs" between programs (Larroucau and Rios, 2020b) or through unfilled slots.

Given the failure of the status quo in finding a stable matching, we ask: *How can a market be designed to better account for peer preferences?* We propose a mechanism that solicits ordinal preferences of students over programs, and in particular, does not require or ask students to report detailed information about the "functional form" of peer preferences.[5] The proposed mechanism induces a controlled tâtonnement process *within* each cohort of students, and unlike the status quo process, it is guaranteed to nearly converge, meaning that our mechanism always generates a (approximately) stable matching. Moreover, our proposed mechanism incentivizes truth-telling, reducing the advantages sophisticated students have when strategically gaming the mechanism. This enhances fairness by placing sophisticated and unsophisticated students on a more equal footing (Pathak and Sönmez, 2008; Song et al., 2020). Reporting costs remain low in that the vast majority of students experience no change in how they interact with the new mechanism compared to the status quo process.

Related Literature

Several theoretical papers have studied peer preferences in centralized matching frameworks, and typically focus on showing the existence of stable matchings. One literature studies the effects

[5]Budish and Kessler (2021) suggest that students may not be capable of accurately stating functional preferences, and Carroll (2018) suggests that any such mechanism may be outside the realm of consideration for many centralized clearinghouses.

of couples in matching markets (see e.g. Roth and Peranson, 1999; Kojima et al., 2013; Nguyen and Vohra, 2018). These papers differ from ours in that peer preferences depend only on the presence of an agent's spouse, not on the entire set of peers. Another literature (Echenique and Yenmez, 2007; Bykhovskaya, 2020; Pycia, 2012; Pycia and Yenmez, 2023) studies general forms of peer preferences with small sets of students, and they primarily focus on identifying conditions under which stable matchings exist. Unlike in our setting, stable matchings frequently do not exist. Greinecker and Kah (2021) study stability with peer preferences in large finite matching markets, and a contemporaneous paper (Leshno, 2022) considers a continuum market.

There are several key differences between our paper and Leshno (2022). First, our model allows students to have preferences over the entire distribution of peer abilities, whereas Leshno (2022) assumes students care only about summary statistics of peer abilities. As a result, our existence result is shown using an infinite-dimensional fixed point theorem, which contrasts with their approach. The added generality is potentially valuable because we show that certain reasonable forms of peer preferences cannot be expressed via (any finite number of) summary statistics. Second, our results also apply to the special case in which students have preferences only over summary statistics of the distribution of peers. Importantly, due to the generality of our base model, our results apply to the case in which students have preferences over their ordinal rank in the class, which reflects our empirical setting but is not supported in the analysis of Leshno (2022). Third, other than initial existence results, the focuses of our papers diverge. They show that the continuum model is a valid approximation of large, finite models while we study the consequences of peer preferences in present-day school choice markets.

Our theoretical results on the use of deferred acceptance with peer preferences mirror some findings on the static and dynamic application of the Boston (immediate acceptance) mechanism under standard, non-peer preferences. While the Boston mechanism does not generate a stable matching with respect to the submitted preferences, the set of equilibria of the static game induced by the Boston mechanism corresponds to the set of stable matchings under complete information (Ergin and Sönmez, 2006). Similarly, our Proposition 1 finds that complete information is sufficient and almost necessary to generate a stable matching with peer preferences in a static setting. In a dynamic model in which a sequence of student cohorts each simultaneously submit preferences to the Boston mechanism after observing a summary statistic from the previous cohort's matching, Çelebi (2022) shows conditions under which the sequence of matchings generated by iterative best responses converges to a stable matching. Our Proposition 2 provides an empirical test for convergence to stability under iterative best responses, and our Theorem 2 states that such

convergence is not guaranteed except in special cases. Overall, this comparison suggests that peer preferences may cause bigger barriers to generating a stable matching over time than does the use of an unstable matching mechanism in markets without peer preferences.

Empirically, we contribute to a literature studying peer preferences in school choice settings (Rothstein, 2006; Allende, 2020; Abdulkadiroğlu et al., 2020; Ainsworth et al., 2023; Che et al., 2022; Beuermann and Jackson, 2022; Beuermann et al., 2023; Campos, 2024). The majority of these papers find that (parents of) students prefer, on average, programs where peers have higher ability. We provide evidence that *relative* peer ability affects the desirability of a program holding fixed other program attributes.[6] While there has been less evidence for preferences over relative peer ability in school choice, relative peer comparisons have been shown to affect academic performance (Frank, 1985; Azmat and Iriberri, 2010; Tincani, 2018), psychic costs (Pop-Eleches and Urquiola, 2013; Ainsworth et al., 2023) in schools and satisfaction at work (Card et al., 2012).

We also contribute methodologically to preference identification in school choice markets. Existing literature either assumes that student ROLs fully represent ordinal preferences due to mechanisms incentivizing truthful reporting (e.g., Hastings et al., 2009; Luflade, 2019), or posits that ROLs capture only a subset of true preferences (Fack et al., 2019). We employ identification strategies to address both scenarios.

In one approach, we exploit temporal variation in peer information presented to applicants. Under the assumption that ROLs accurately reflect student preferences, we identify peer preferences through shifts in application rates across the test-score distribution. In our second approach, we leverage test score revelations to induce variation in relative peer ability. Building on discrete choice methods and institutional specifics, we examine "switches" in students' relative rankings of programs, controlling for pre-revelation ROLs in a manner similar to Ferreira and Wong (2023). To our knowledge, ours is the first analysis to exploit multiple ROLs from the same student within a single matching market.[7] Importantly, this approach relies solely on ranking switches and does not require that students fully disclose their preferences, aligning with Fack et al. (2019). Both strategies reveal a pattern aligning with "big fish in a little pond" preferences wherein students prefer not to be overmatched by peers.

The remainder of the paper is structured as follows: Section II presents our model of peer

[6] Beuermann et al. (2023) and Abdulkadiroğlu et al. (2020) estimate preferences for average peer ability separately for high- and low-ability students, which is closer to an analysis of relative peer preferences. They find more muted effects of high peer ability on program desirability for low ability students, which is consistent with our findings.

[7] Narita (2018) and Larroucau and Rios (2020b) study multiple ROLs from the same student, but they investigate students reapplying to schools in subsequent markets after being matched. Preference and program drift may be a concern in such situations. Our shorter time frame within the same application period and model assumptions mitigate this issue.

preferences in a matching environment and shows the existence of a stable matching; Section III theoretically assesses instability in status-quo matching processes and provides an empirical test for stability; Section IV presents a new mechanism that finds a stable matching in the presence of peer preferences; Section V discusses the NSW Tertiary Education System, provides evidence of peer preferences, and presents evidence of instability in the status quo; Section VI concludes. Omitted proofs are housed in Appendix A, while the remainder of the appendix considers robustness checks, additional results, and details on our empirical approach.

II Model Setup and Existence of Stable Matchings

A continuum of students is to be matched to a finite set of $N \geq 1$ programs $C = \{c_1, c_2, ..., c_N\} \cup \{c_0\}$, where c_0 represents the "outside option" of being unmatched. Each student is represented by a type θ, and Θ denotes the set of all possible student types. We further describe set Θ below. η is a non-atomic measure over Θ in the Borel $\sigma-$algebra of the product topology of Θ. We normalize $\eta(\Theta) = 1$. Each program $c \in C$ has capacity $q^c > 0$ measure of seats, with $q^{c_0} \geq 1$. Let $q = \{q^c\}_{c \in C}$.

To capture that student preferences depend on their peers, we first characterize potential peer groups as a useful building block. Informally, this construction allows us to isolate student preferences without concern for capacity constraints. An *assignment* of students to programs α is a measurable function $\alpha : C \cup \Theta \to 2^{\Theta} \cup 2^{C}$ such that:

1. for all $\theta \in \Theta$, $\alpha(\theta) \subset C$,
2. for all $c \in C$, $\alpha(c) \subset \Theta$ is measurable, and
3. $\theta \in \alpha(c)$ if and only if $c \in \alpha(\theta)$.

Condition 1 states that a student is assigned to a subset of programs, Condition 2 states that a program is assigned to a subset of students, and Condition 3 states that a student is assigned to a program if and only if the program is also assigned to that student. We denote the set of all assignments by $\mathcal{A}$.

Each student is characterized by a type $\theta = (u^\theta, r^\theta)$. $u^\theta(c|\alpha) \in \mathbb{R}$ represents the cardinal utility the student derives from being assigned to only program c given that other students are assigned according to assignment $\alpha \in \mathcal{A}$. That is, $u^\theta(c|\alpha) = u^\theta(c|\alpha(\theta) = c \text{ and } \{\alpha(\theta')\}_{\theta' \in \Theta \setminus \{\theta\}})$. We normalize $u^\theta(c_0|\alpha) = 0$ for all $\theta \in \Theta$ and all $\alpha \in \mathcal{A}$, that is, each student receives a constant utility from being unassigned. It will often be useful to denote ordinal preferences. Let $\mathcal{P}$ be the set of all possible linear orders over programs $c \in C$. Let $\succeq^{\theta|\alpha} \in \mathcal{P}$ represent θ's ordinal preferences over programs at assignment α, that is $c_i \succeq^{\theta|\alpha} c_j$ ($c_i \succ^{\theta|\alpha} c_j$) if and only if $u^\theta(c_i|\alpha) \geq u^\theta(c_j|\alpha)$ ($u^\theta(c_i|\alpha) > u^\theta(c_j|\alpha)$).

$r^{\theta,c} \in [0,1]$ is θ's score at program $c \in C$. We write r^{θ} to represent the vector of scores for student θ at each program. To ensure smoothness over the distribution of student scores, we assume that the measure over scores induced by η is absolutely continuous for each $c \in C$. Scores only convey ordinal information in our analysis with this restriction, so without loss of generality we assume that $\eta\{\theta | r^{\theta,c} < y\} = y$ for all $y \in [0,1]$ and all $c \in C$, that is, the marginal distribution of every program's scores is uniform. Therefore, no set of students of positive measure have the same scores: for any $\theta \in \Theta$ and any $c \in C$, $\eta(\{\hat{\theta} \in \Theta | r^{\hat{\theta},c} = r^{\theta,c}\}) = 0$.

We define the set of all student types as $\Theta = \mathbb{R}^{N+1} \times \mathcal{A} \times [0,1]^{N+1}$. We denote a market by $E = [\eta, q, N, \Theta]$.

We will focus our study on markets in which peer preferences depend on the distribution of peer characteristics. Because program c_0 represents the outside option without a binding capacity constraint, r^{θ,c_0} is a free variable without a clear meaning. We refer to r^{θ,c_0} as student θ's "ability." For each $x \in [0,1]$, $c \in C$, and $\alpha \in \mathcal{A}$, let $\lambda^{c,x}(\alpha) := \eta(\{\theta \in \alpha(c) | r^{\theta,c_0} \leq x\})$. Let $\lambda^{c}(\alpha)$ be the resulting non-decreasing function from $[0,1]$ to $[0,1]$ and let Λ be the set of all such functions. Let $\lambda(\alpha) := (\lambda^{c_1}(\alpha), ..., \lambda^{c_N}(\alpha), \lambda^{c_0}(\alpha))$. In words, for given assignment α each $\lambda^{c}(\alpha)$ is a CDF-like object that reveals the measure of students at program c with abilities below each $x \in [0,1]$, and $\lambda(\alpha)$ represents the vector of ability distributions for all programs.

Three observations are in order regarding our peer-preference machinery. First, ability distributions $\lambda(\cdot)$ are infinite-dimensional objects, which leads to a more general analysis. Specifically, we show in Appendix B that certain homophilic peer preferences can only be represented by allowing peer preferences to depend on the entire distribution of peer abilities, as opposed to a finite collection of summary statistics. Second, all of our results generalize straightforwardly if we add more dimensions for a student's type (e.g. race or gender) and allow preferences to additionally depend on the distribution of peer characteristics along these added dimensions. However, and third, adding additional dimensions is unnecessary because the distribution η is assumed to be atomless, and the marginal distribution of student abilities is assumed to have full support over $[0,1]$. In other words, the set of student "abilities" has cardinality equal to the continuum, allowing us to represent many aspects of a student's peer-related characteristics into a one-dimensional measure.[8]

In what follows, we restrict our focus to markets E satisfying regularity conditions **A1**-**A4** in order to remove nuisance cases and to better reflect our desired setting.

[8] For example, an alternative modeling choice would be that students care about peers' scores across all N schools in addition to their "ability." This would result in peer preferences depending on an object in $[0,1]^{N+1}$ instead of in $[0,1]$. But note that both of these sets have the same cardinality, meaning no generality is gained by this alternative.

A1 Strict preferences for all α: for any $\alpha \in \mathcal{A}$, $\eta(\{\theta | \succeq^{\theta|\alpha} \text{ is a strict ordering}\}) = 1$.

A2 Student preferences depend only on $\lambda(\alpha)$: for any $\alpha, \alpha' \in \mathcal{A}$ such that $\lambda(\alpha) = \lambda(\alpha')$, $\succeq^{\theta|\alpha} = \succeq^{\theta|\alpha}$ for all $\theta \in \Theta$. We will therefore write $\succeq^{\theta|\lambda(\alpha)}$ to mean $\succeq^{\theta|\alpha}$ for $\theta \in \Theta$.

A3 Rich support for all α: There exists $\omega > 0$ such that for any $[b_1, b_2] \subset [0,1]$, any $\alpha \in \mathcal{A}$, and any $c \in C \setminus \{c_0\}$: $\eta(\{\theta \in \Theta | r^{\theta,c} \in [b_1, b_2] \text{ and } c \succ^{\theta|\alpha} c_0 \succ^{\theta|\alpha} c' \text{ for all } c' \in C \setminus \{c, c_0\}\}) > \omega(b_2 - b_1)$.

A4 Peer preferences are separable and smooth: For all $\theta \in \Theta$, all programs $c \in C \setminus \{c_0\}$, and all assignments $\alpha \in \mathcal{A}$, $u^\theta(c|\alpha) = v^{\theta,c} + f^{\theta,c}(\lambda^c(\alpha))$, where $v^{\theta,c} \in \mathbb{R}$ is an exogenous component of preferences, and for all $\theta \in \Theta$ and all $c \in C \setminus \{c_0\}$:

- $f^{\theta,c}(\cdot)$ is uniformly continuous: for any $\epsilon > 0$ there exists $\delta > 0$ such that if $\lambda^c, \hat{\lambda}^c \in \Lambda$ satisfy $||\lambda^c - \hat{\lambda}^c||_\infty < \delta$ then $|f^{\theta,c}(\lambda^c) - f^{\theta,c}(\hat{\lambda}^c)| < \epsilon$, and
- $f^{\theta,c}(\cdot)$ is uniformly bounded: there exists $a < b$ such that $f^{\theta,c}(\cdot) \in [a, b]$.

A1 is a standard assumption in the literature of almost no indifferences in preferences, extended to hold for any collection of peers. **A2** implies that peer preferences are anonymous and depend only on the ability of peers. This rules out, for example, the well-known couples matching problem in which a student prefers being assigned to the same program as her spouse. **A3** and **A4** are richness assumptions. **A3** states that for any program c and any assignment there exist a positive fraction of students across the score distribution who are only willing to attend program c—a similar condition appears in Grigoryan (2022). **A4** states that student preferences change smoothly in small changes in peer composition at each program.

II.A Stable matchings

A matching is an assignment that satisfies capacity constraints. Formally, a *matching* μ is a measurable function $\mu : C \cup \Theta \to 2^\Theta \cup C$ such that:

1. for all $\theta \in \Theta$, $\mu(\theta) \in C$,
2. for all $c \in C$, $\mu(c) \subset \Theta$ is measurable and $\eta(\mu(c)) \leq q^c$, and
3. $\theta \in \mu(c)$ if and only if $c = \mu(\theta)$.

Compared to an assignment, Condition 1 adds that a student can only be matched to one program, and Condition 2 adds that the measure of students matched to a program cannot exceed the capacity

of that program. We will often refer to a student θ for whom $\mu(\theta)=c_0$ as being "unmatched." Let $\mathcal{M}$ be the set of all matchings. To reduce a multitude of essentially identical matchings that differ only for a measure zero set of students, throughout the paper we only consider matchings $\mu\in\mathcal{M}$ that are *right continuous*: for any c and θ, if $c\succ^{\theta|\mu}\mu(\theta)$ then there exists $\epsilon>0$ such that $\mu(\theta')\neq c$ for all θ' with $r^{\theta',c}\in[r^{\theta,c},r^{\theta,c}+\epsilon)$.

A student-program pair (θ,c) *blocks* matching μ if $c\succ^{\theta|\mu}\mu(\theta)$ and either (i) $\eta(\mu(c))<q^c$, or (ii) there exists $\theta'\in\mu(c)$ such that $r^{\theta,c}>r^{\theta',c}$. In words, θ and c block matching μ if θ prefers c to her current program (given peer preferences at μ) and either c does not fill all of its seats, or it admits a student it ranks lower than θ. A matching is *(pairwise) stable* if there do not exist any student-program blocking pairs. Throughout, we shorten the name of this solution concept to "stability."

Remark 1. *The following axioms are jointly equivalent to stability. A matching μ is:* individually rational *if $\mu(\theta)\succeq^{\theta|\mu}c_0$ for all θ;* non-wasteful *if for some θ and c it is the case that $c\succ^{\theta|\mu}\mu(\theta)$ then $\eta(\mu(c))=q^c$;* fair *if there do not exist θ,θ' and c such that $\mu(\theta')=c$, $c\succ^{\theta|\mu}\mu(\theta)$, and $r^{\theta,c}>r^{\theta',c}$.*

Note that if the (ordinal) preferences of all students are constant for all $\alpha\in\mathcal{A}$ then our definition of stability collapses to the standard definition. Also, our analysis is largely unchanged other than notational complications by relaxing non-wastefulness to allow a program to reject sufficiently low-scoring students even when it has an excess supply of seats.

We specify a class of assignments defined by admission cutoffs. This construction will be used to characterize stable matchings, as in Azevedo and Leshno (2016). A cutoff vector $p\in[0,1]^{N+1}$ is subject to $p^{c_0}=0$. One can construct an assignment given a cutoff vector p as follows. First, fix an arbitrary assignment α', and corresponding ability distribution $\lambda=\lambda(\alpha')$. Second, let each student θ choose her favorite program among those where her program-specific score is weakly above the cutoff.[9] We refer to this program as the *demand of* θ, and denote it by

$$D^\theta(p,\lambda)=\underset{\succeq^{\theta|\lambda}}{\operatorname{argmax}}\{c\in C|r^{\theta,c}\geq p^c\}.$$

Any θ can demand to be unmatched because $p^{c_0}=0$. We define the *demand for program* c as

$$D^c(p,\lambda)=\eta(\{\theta\in\Theta|D^\theta(p,\lambda)=c\}).$$

The assignment $\alpha=A(p,\lambda)$ is defined by setting $\alpha(\theta)=D^\theta(p,\lambda)$ for every student θ. By construction, each student is assigned to exactly one program in assignment $\alpha=A(p,\lambda)$, but a

[9] If a student does not have a unique favorite program, break ties arbitrarily. By Assumption **A1**, ties only occur for a negligible set of students, therefore, we proceed as if each student has a unique top choice.

program may be assigned to a larger measure of students than its capacity. The following two conditions link this construction to stable matchings.

Definition 1. *A pair* (p,λ) *of cutoffs and ability distributions is* market clearing *if for all programs* $c \in C$ *it is the case that* $D^c(p,\lambda) \leq q^c$*, and* $p^c = 0$ *whenever this inequality is strict.*

Lemma 1. *If a pair* (p,λ) *is market clearing, then* $A(p,\lambda)$ *is a matching.*

The proof of this result is immediate, as for each $c \in C$, $\eta(\alpha(c)) \leq q^c$ and for each $\theta \in \Theta$, $\alpha(\theta) \in C$. If (p,λ) is market clearing, we refer to matching $\mu = A(p,\lambda)$ as being *market clearing*, and we denote by M the set of all market clearing matchings, that is $M = \{\mu | \mu = A(p,\lambda)$ for some market clearing $(p,\lambda)\}$. By construction, $M \subset \mathcal{M}$.

Definition 2. *A pair* (p,λ) *satisfies* rational expectations *if it induces an assignment* $\alpha = A(p,\lambda)$ *such that* $\lambda = \lambda(\alpha)$.

The following is a direct corollary of the supply and demand lemma of Azevedo and Leshno (2016):[10]

Lemma 2. *If a pair* (p,λ) *is market clearing and satisfies rational expectations, then* $\mu = A(p,\lambda)$ *is a stable matching. For each* $c \in C$ *let* $\hat{p}^c := inf\{r^{\theta,c} | \theta \in \mu(c)\}$ *and let* $\hat{p} = (\hat{p}^{c_1},...,\hat{p}^{c_N},0)$*. If* μ *is a stable matching, then* $(\hat{p},\lambda)$ *is market clearing and satisfies rational expectations for* $\lambda = \lambda(A(\hat{p},\lambda(\mu)))$.

The following result finds that stable matchings exist in markets satisfying our regularity conditions. We prove this result by constructing an operator over the space of cutoffs p and ability distribution λ whose fixed points corresponds to stable matchings. We then show, using Schauder's fixed-point theorem for infinite-dimensional spaces, that at least one fixed point exists.[11]

Theorem 1. *There exists a stable matching in any market* E *satisfying* ***A1****-****A4****.*

The proof of Theorem 1 in the appendix shows a stable matching exists if we replace **A4** with a weaker, ordinal condition. Additionally, we can weaken the requirement that peers in any program c affect preferences only over program c to allow student preferences for c to depend on the vector of ability distributions at all programs. Therefore, our result is more general and shows the existence of stable matchings even under "externality" preferences, as discussed in Sasaki and Toda (1996).

[10] Also see Leshno (2022) for a similar finding.

[11] Fixed-point arguments are often used to show existence results in the literature (see, for example, Pycia and Yenmez, 2023; Leshno, 2022). One technical difficulty is that we must deal with an infinite-dimensional space, reflecting the cardinality of the space of (peer ability) distributions. This necessitates the use of Schauder's fixed point theorem as opposed to a more-standard approach using Brower's theorem. To our knowledge, only Grigoryan (2021) uses Schauder's fixed point theorem in an existence argument in a matching problem, although the focus of his study is not peer preferences.

Remark 2. *There need not be a unique stable matching. For example, if $N \geq 2$, all programs offer students similar intrinsic utility (i.e. $v^{\theta,c} \approx v^{\theta,c'}$ for all θ and all c,c'), and students desire classmates with higher abilities, then the "best" program is endogenously determined by the coordination of top-ability students, the "second best" program by coordination of the next-highest-ability students, and so on.*

III Does the status-quo result in a stable matching?

Theorem 1 tells us that a stable matching exists. Does the status quo matching process find a stable matching? We show that the answer is generally "no." First, we study a static environment, where, given student beliefs, the market designer uses a "canonical" matching mechanism. We show that unless students' beliefs are "sufficiently correctly specified," no reasonable matching mechanism will deliver a stable matching. We then study a dynamic process mirroring our empirical setting, where student beliefs are formed by empirical observation. We show that student beliefs may never become sufficiently correctly specified. As a result, the status quo matching procedure never generates a stable matching in the long run.

III.A Static setting

In any market E, define a *one-shot matching mechanism* φ as a simultaneous-move, deterministic game in which each student θ submits an ROL $\tilde{\succ}^{\theta}$ over programs $c \in C$. φ maps ROLs $\tilde{\succ} = \{\tilde{\succ}^{\theta}\}_{\theta \in \Theta}$ and scores into a matching, that is $\varphi : (\mathcal{P} \times [0,1]^{N+1})^{\Theta} \to \mathcal{M}$. We suppress dependency on scores when there is no risk of confusion and represent the resulting matching from report $\tilde{\succ}$ as $\varphi(\tilde{\succ})$, the matched partner for student θ as $\varphi^{\theta}(\tilde{\succ})$, and the set of students matched to program c as $\varphi^{c}(\tilde{\succ})$. A one-shot mechanism φ *respects rankings* if for any $\tilde{\succ}$ the following is satisfied: if $r^{\theta,c} > r^{\theta',c}$ for all c and $|\{c|c \,\tilde{\succ}^{\theta} \varphi^{\theta'}(\tilde{\succ})\}| \leq |\{c|c \,\tilde{\succ}^{\theta'} \varphi^{\theta'}(\tilde{\succ})\}|$, then $\varphi^{\theta}(\tilde{\succ}) \tilde{\succeq}^{\theta} \varphi^{\theta'}(\tilde{\succ})$. That is, a student is not matched to a program c' that she ranks below some program c if she has a higher score (across all programs) than another student who is matched to c, and she ranks c at least as high as the student with lower scores.[12] We refer to φ as "canonical" if it is a one-shot mechanism which respects rankings.

A stronger requirement is stability. A one-shot mechanism φ is *stable* if for any $\tilde{\succ}$, $\varphi(\tilde{\succ})$ is stable *with respect to* $\tilde{\succ}$. Note that any one-shot stable mechanism φ must respect rankings.[13]

[12] The requirement that she ranks c at least as high as the student with lower scores (i.e. $|\{c|c\tilde{\succ}^{\theta}\varphi^{\theta'}(\tilde{\succ})\}| \leq |\{c|c\tilde{\succ}^{\theta'}\varphi^{\theta'}(\tilde{\succ})\}|$) is included to expand the class of covered mechanisms to include the immediate acceptance mechanism. Removing this additional requirement would not otherwise change our results.

[13] Proof: Suppose not. Then for some $\tilde{\succ}$ there exist θ, θ' with $r^{\theta,c} > r^{\theta',c}$ for all c, and $c^* = \varphi^{\theta'}(\tilde{\succ}) \tilde{\succ}^{\theta} \varphi^{\theta}(\tilde{\succ})$. But

The following result says that we can expect a clearinghouse to generate a stable matching by using a stable mechanism if students have full knowledge of the distribution of student types.[14] In this case, the set of stable matchings is Bayes Nash implemented by any stable mechanism φ as students are able to "roll in" peer considerations into their ROLs. That is, for any stable matching μ_*, there is an equilibrium in which each student θ reports $\tilde{\succ}^{\theta} = \succeq^{\theta|\mu_*}$.[15] On the other hand, if students' beliefs about the distribution of types are sufficiently misspecified, then we should not expect a clearinghouse to generate a stable matching using a canonical mechanism.

Suppose student θ believes the measure over Θ is given by σ^{θ}. Let $\tilde{\succ}$ be a strategy profile, and let $\mu(\sigma^{\theta},\tilde{\succ})$ be the anticipated matching of student θ. Then θ's expected ordinal rankings over programs given σ^{θ} and $\tilde{\succ}$ is $\succeq^{\theta|\mu(\sigma^{\theta},\tilde{\succ})}$. We say that student θ *lacks rationality for the top choice at* $(\tilde{\succ},\varphi)$ if the $\succeq^{\theta|\mu(\sigma^{\theta},\tilde{\succ})}$-maximal program is not a $\succeq^{\theta|\varphi(\tilde{\succ})}$-maximal program. For any $r \in [0,1)^{N+1}$ let $L_{\tilde{\succ},\varphi,r} := \{\theta | r^{\theta} \geq r$ and θ lacks rationality for the top choice at $(\tilde{\succ},\varphi)\}$.

Proposition 1. *Consider a one-shot matching mechanism φ.*

1. *Let φ be stable and suppose $\sigma^{\theta} = \eta$ for all $\theta \in \Theta$. Then the set of all stable matchings of market E is identical to the set of all Bayes-Nash equilibrium outcomes of φ.*

2. *Let φ respect rankings, let μ_* be a stable matching, and let $\bar{p} \in [0,1)^{N+1}$ be such that $\bar{p}^c = \max\{1-q^c,0\}$ for all $c \in C$. If for all $\tilde{\succ}$ and some $\tilde{p} \in [0,1)^{N+1}$ satisfying $\tilde{p}^c > \bar{p}^c$ for all $c \in C \setminus \{c_0\}$ it is the case that $\eta(L_{\tilde{\succ},\varphi,\tilde{p}}) > 0$, then there is no Bayes Nash equilibrium of φ that generates μ_*.*

The presence of some students with incorrect beliefs is not necessarily enough to lead to an unstable matching; a number of additional conditions must be met. First, these students must have sufficiently strong peer preferences so that their incorrect beliefs change their ROLs. Second, these students must have scores above the admission thresholds at these programs. Third, the incorrect beliefs affect the preferences at the "top" of some students' rankings, because, for example, changes

since $r^{\theta,c^*} > r^{\theta',c^*}$, it is the case that (θ,c^*) form a blocking pair with respect to $\tilde{\succ}$. Contradiction with φ being stable.

[14] Full knowledge of the distribution of types is not a necessary condition for the clearinghouse to generate a stable matching. As the distribution of peers within programs is the only payoff-relevant feature of the market (Esponda and Pouzo, 2016) (in a strategy-proof mechanism), a stable matching can be generated in equilibrium if students anticipate the distribution of peers at each program with sufficient accuracy. We explore this in Section III.

[15] As we discuss in the proof of Proposition 1, for any stable mechanism φ, if all students θ report $\tilde{\succ}^{\theta} = \succeq^{\theta|\mu_*}$ then $\varphi(\tilde{\succ}) = \mu_*$, as μ_* is the only stable matching associated with these preferences. Moreover, we show the existence of an equilibrium yielding μ_* in which each student lists only one program as acceptable. Therefore, even if there is a cap on the number of programs that students can list, which is common in many school choice markets around the world, stable matchings can be generated in equilibrium, under full knowledge of the distribution of student types.

in the ranking order of programs that are deemed unacceptable do not affect the final matching. Informally speaking, these conditions are likely satisfied if students have a sufficiently rich set of beliefs across the ability distribution.

III.B Dynamic setting and belief updating

Given Proposition 1, an important question is how students form beliefs when submitting ROLs to a centralized mechanism. We model belief formation in a tâtonnement-like process, in which beliefs update given the assignment of the previous cohort of students.

Formally, we consider a discrete-time, infinite horizon model, where at every time $t = 1,2,3,...$, the same programs are matched to a new cohort of students. For any $t,t' \geq 1$, markets E_t and $E_{t'}$ are identical. We therefore omit all time indices when denoting market fundamentals.

The following dynamic process—which we call *Tâtonnement with Intermediate Matching (TIM)* process—generates the matching in each period. The market is initialized with an arbitrary assignment $\mu_0 \in \mathcal{A}$.[16] At each time period $t \geq 1$, a matching μ_t is constructed as follows. Incoming students at time t observe μ_{t-1}. A matchmaker solicits an ROL from each student, and then uses a stable matching mechanism to construct matching μ_t. We assume students use information from the previous period in a Cournot-updating fashion; that is, each period t student has a Dirac belief that $\lambda(\mu_t)$ will equal $\lambda(\mu_{t-1})$.[17]

To study how matchings evolve over time, fix some $t \geq 1$. Given μ_{t-1}, there is a unique stable matching in an otherwise-identical matching market in which each student θ's preferences is given by $\succeq^{\theta|\mu_{t-1}}$. That is, if we "roll in" peer preferences and treat them as exogenously fixed given the matching in period $t-1$, we recover the well-known uniqueness result of stable matchings in markets without peer preferences. The following remark states this formally, and follows directly from Assumption **A3** and Grigoryan (2022).

Remark 3. *Fix a market $E = [\eta, q, N, \Theta]$ and let μ_{t-1} be an arbitrary assignment. Let $\tilde{E}_t = [\zeta^{\eta,\mu_{t-1}}, q, N, \Theta^{\mu_{t-1}}]$ where $\Theta^{\mu_{t-1}}$ and $\zeta^{\eta,\mu_{t-1}}$ jointly satisfy the following condition for any open set $R \subset [0,1]^{N+1}$, any assignment α, and any $\succeq$: $\zeta^{\eta,\mu_{t-1}}(\{\theta \in \Theta^{\mu_{t-1}} | r^\theta \in R$ and $\succeq^{\theta|\alpha} = \succeq$*

[16] We initialize the market with an assignment instead of a matching so as not to require students in the first cohort to be fully informed of all particulars in the market, for example, the capacity at each program; our results are qualitatively unchanged if we instead allow students to have (potentially heterogeneous) beliefs over the initial assignment μ_0, but the exposition would become more cumbersome.

[17] Supporting this assumption, Dur et al. (2021) find that when students are shown information about previous matchings, 84% of them submit ROLs that best respond to the previous matching. Similar conclusions hold if beliefs over $\lambda(\mu_t)$ are Dirac over a linear combination of ability distributions during a finite look-back of $k > 1$ periods, $\lambda(\mu_{t-1}),...,\lambda(\mu_{t-k})$.

$\}) = \eta(\{\theta \in \Theta | r^\theta \in R \text{ and } \succeq^{\theta|\mu_{t-1}} = \succeq\})$. *Then there is a unique stable matching* μ^* *in market* $\tilde{E}_t$ *and* $\lambda^{c,x}(\mu_t) = \zeta^{\eta,\mu_{t-1}}(\{\theta \in \mu^*(c) | r^{\theta,c_0} \leq x\})$ *for all* $c \in C$ *and all* $x \in [0,1]$.

Returning to our sequential matching setting with peer preferences, the previous remark implies that for each $t \geq 1$, every student θ has a weakly dominant strategy, given her (potentially misspecified) beliefs, to submit her "true" preferences $\succeq^{\theta|\mu_{t-1}}$ in any stable matching mechanism used by the designer (Abdulkadiroğlu et al., 2015). We adopt the assumption that every student θ reports $\succeq^{\theta|\mu_{t-1}}$ going forward. Therefore, for $t \geq 1$, $\mu_t = A(p_t, \lambda_{t-1})$, where $p_t \in [0,1]^{N+1}$ is the (unique) cutoff vector such that (p_t, λ_{t-1}) is market clearing, and $\lambda_t := \lambda(A(p_t, \lambda_{t-1}))$ where $\lambda_0 := \lambda(\mu_0)$. Note that the entire sequence of TIM matchings $\{\mu_t\}_{t \geq 1}$ is uniquely determined by μ_0.

A central goal in school choice markets is to reliably produce a stable matching, ensuring no student-program pairs have incentives to block the matching. Our main theoretical result delivers a negative finding: the status-quo TIM process does not, in general, guarantee stability in any time period. This result is perhaps surprising because, given Remark 3, it implies that mechanisms designed to ensure a stable matching—and widely recommended to policy makers—can fail in the presence of peer preferences, even given an arbitrary amount of time to aggregate information from previous cohorts. We present this result here in an informal manner before interpreting it, describing economic intuitions, and providing a related empirical test for stability. In the appendix, we formally state and prove this result.

Theorem 2. *For almost any collection of peer preference functions, there exists a market and a starting condition* μ_0 *for which the TIM process does not yield a stable matching at any time* t*. For any number of programs, any capacity vector, and any distribution of student scores, there exists a market such that the TIM process does not yield a stable matching at any time* t *for almost any starting condition* μ_0*.*

The two claims contained in this result imply that there are "no guarantees" of stability in the status-quo TIM process, even when accompanied by powerful market interventions. The first claim states that assigning each student a particular peer preference function is unlikely to be sufficient to guarantee stability; markets exist for which the selected peer preference functions almost surely (in the space of all possible peer preferences functions) fail to yield stability. In other words, an institution that can affect students' preferences over their peers (e.g. via a media campaign) may not be successful in generating stability. The second claim states that pinning down all aspects of the "program-side" market structure is insufficient to guarantee stability. In other words, a market clearinghouse that can regulate the number of programs, their capacities, and their admissions

criteria may fail to induce a stable matching.

Theorem 2 does not imply that the TIM process *cannot* ever yield a stable matching, only that it *need not* do so. Therefore, an important consideration in proving Theorem 2 is identifying when a stable matching is generated. The following result, Proposition 2, connects stability in the TIM process to classical intuitions surrounding price convergence and equilibrium in exchange economies. It finds that if and only if the ability distribution vector is in steady state does the TIM process generate a stable matching. Moreover, if and only if the ability distribution vector is in "approximate" steady state does the TIM process generate an "approximately" stable matching. Therefore, the following result provides an empirical test of stability for an observer with only panel data on the ability distribution of entering classes at programs.

Before stating Proposition 2, we give a definition of ϵ-stability; our notion of approximate stability comes from selecting a small ϵ.

Definition 3. *A matching μ is* ϵ-stable *if the measure of students involved in blocking pairs at μ is strictly smaller than ϵ, that is, $\eta(\{\theta|(\theta,c) \text{ block } \mu \text{ for some } c \in C\}) < \epsilon$.*

Proposition 2. *Let $\mu_1,\mu_2,...$ be the sequence of matchings constructed in the TIM process given an initial assignment μ_0 in market E.*

1. *Let $t \geq 1$. If $\lambda_t = \lambda_{t-1}$ then μ_t is stable. Moreover, $\lambda_t = \lambda_{t+1}$ only if μ_t is stable.*
2. *Let $t \geq 1$. For any $\epsilon > 0$ there exists $\delta > 0$ such that if $||\lambda_t - \lambda_{t-1}||_\infty < \delta$ then μ_t is $\epsilon-$stable. Moreover, for any $\delta > 0$ there exists $\epsilon > 0$ such that $||\lambda_t - \lambda_{t+1}||_\infty < \delta$ only if μ_t is $\epsilon-$stable.*

Remark 4. *Proposition 2 is particularly amenable to empirical testing for two reasons. First, suppose students have preferences over* summary statistics *of the ability distribution, as in our empirical setting. A summary statistic of abilities at program c is defined as a function $s^c : \Lambda \to [0,1]$. We represent the vector of summary statistics across all programs given an ability distribution λ as $s(\lambda)$. The following continuity condition on summary statistics is sufficient to apply Proposition 2 under the assumption that peer preferences are determined by summary statistics, i.e. the TIM process results in a (approximate) stable matching if and only if the summary statistics of all programs are in (approximate) steady state: For any $\epsilon > 0$ there exists $\delta > 0$ such that for any assignments α,α' that satisfy $\alpha = A(p,\lambda)$, $\alpha' = A(p',\lambda')$ for some $(p,\lambda),(p',\lambda') \in [0,1]^{N+1} \times \Lambda^{N+1}$ and $||\lambda(\alpha) - \lambda(\alpha')||_\infty < \delta$, we have that $||s(\lambda(\alpha)) - s(\lambda(\alpha'))||_\infty < \epsilon$.*

The second point of Proposition 2 also implies that small changes over time in the market do not affect the predictions of our empirical test. For example, student preference distributions could

drift slightly over time as certain majors become more demanded due to labor market changes. Notably, if the fundamentals of the markets in times t and $t+1$ are "close" for all t such that for any stable matching μ_t in the market in time t there is a stable matching μ_{t+1} in the market in time $t+1$ such that $||\lambda(\mu_t)-\lambda(\mu_{t+1})||_\infty$ is small, then the convergence of the ability distribution over time is still necessary and sufficient for approximate stability.

For intuition regarding Theorem 2 and Proposition 2, we provide two examples that show the convergence of the TIM process to a stable matching is sensitive to the functional form of peer preferences. In both examples, there is a unique stable matching, but convergence occurs in only one.

Example 1. *Let $N=1$. The lone program c_1 has $q^{c_1} \leq \frac{1}{3}$ measure of seats, and let $r^{\theta,c_1}=r^{\theta,c_0}$ for all θ. Let $s^{c_1}(\lambda(\alpha))$ represent the mean of abilities of enrolled students at c_1 in assignment α, that is,*

$$s^{c_1}(\lambda(\alpha))=\frac{1}{\lambda^{c_1,1}(\alpha)}\int_0^1 y d\lambda^{c_1,y}(\alpha).$$

Let $k \in [\frac{4}{5}, 1-\frac{q^{c_1}}{2-q^{c_1}}]$. Each θ receives utility $u^\theta(c_1|\alpha) = v^{\theta,c_1} - f(s^{c_1}(\lambda(\alpha)), r^{\theta,c_1})$ from attending program c_1 given α, where

$$f(s^{c_1}(\lambda(\alpha)),r^{\theta,c_1})=\begin{cases} 0 \text{ if } r^{\theta,c_1} \geq s^{c_1}(\lambda(\alpha)) \\ k \text{ if } r^{\theta,c_1} < s^{c_1}(\lambda(\alpha)) \end{cases}.$$

The peer preference term $f(\cdot,\cdot)$ reflects that students want to be a "big fish in a little pond" and suffer loss k if their ability is below average at the program.[18] *Let each v^{θ,c_1} be distributed independently of r^{θ,c_1} and uniformly over $[0,1]$: for any intervals on the unit line (b_1,b_2) and (b_3,b_4), $\eta(\{\theta|r^{\theta,c_1} \in (b_1,b_2) \text{ and } v^{\theta,c_1} \in (b_3,b_4)\} = (b_2-b_1)\cdot(b_4-b_3)$. Each θ is better off enrolling at c_1 if and only if $v^{\theta,c_1} - f(s^{c_1}(\lambda(\alpha)),r^{\theta,c_1}) \geq 0$, where we break ties in favor of the student attending c_1.*

Let $s_t^{c_1} := s^{c_1}(\lambda(\mu_t))$ for $t \geq 0$, and initialize the TIM process with μ_0 such that $s_0^{c_1} \leq 1-q^{c_1}$. Then $(p_1^{c_1}, s_1^{c_1}) = (1-q^{c_1}, 1-\frac{q^{c_1}}{2})$, as $\mu_1(c_1) = \{\theta|r^{\theta,c_1} \geq 1-q^{c_1}\}$, that is, the top q^{c_1} scoring measure of students enrolls at c_1 in $t=1$ because they expect (mistakenly for some) to face no peer loss from doing so.

What about $(p_2^{c_1}, s_2^{c_1})$? Given $s_1^{c_1}$, only the $1-k$ fraction of students with $r^{\theta,c_1} < s_1^{c_1}$ for whom $v^{\theta,c_1} \geq k$ prefer c_1 to the outside option, and all students with $r^{\theta,c_1} \geq s_1^{c_1}$ prefer c_1 to the outside option.

[18] Note that $f(\cdot, r^{\theta,c_1})$ is not continuous in its first argument, therefore failing assumption **A4**, however, the qualitative results of this example would be unchanged if we replaced $f(\cdot,r^{\theta,c_1})$ with a function that is continuous, but with a steep slope around the point where $s^{c_1}(\lambda(\alpha))=r^{\theta,c_1}$. The current formulation leads to cleaner calculations.

The program fills all of its seats in μ_2, i.e. $p_2^{c_1} = 1 - \frac{q^{c_1}}{2} - \frac{q^{c_1}}{2(1-k)} \geq 0$.[19] *Therefore, the average ability of the "top half" of the students enrolled in the program is $1 - \frac{q^{c_1}}{4}$ while the average ability of the "bottom half" of the students enrolled is $\frac{1}{2}(1 - \frac{q^{c_1}}{2} + p_2^{c_1})$. This tells us that $s_2^{c_1} = \frac{1}{2}\left[1 - \frac{q^{c_1}}{4} + \frac{1}{2}(1 - \frac{q^{c_1}}{2} + p_2^{c_1})\right]$. But note then that $(p_3^{c_1}, \lambda_3) = (p_1^{c_1}, \lambda_1)$, as now all students with abilities $r^{\theta,c_1} > 1 - q^{c_1}$ wish to attend c_1 given $s_2^{c_1}$. By Assumption **A1**, the market therefore cycles wherein all even periods yield the same matching, while odd periods yield another (note that $p_2^{c_1} < p_1^{c_1}$, as $k > 0$). Therefore, the TIM process may not converge, and by Proposition 2, never achieves stability.*

In the appendix, we show that for sufficiently small q there is a unique stable matching in the matching market. Therefore, this example demonstrates that the TIM does not converge to a stable matching even when there is a unique such matching.

We now consider an example that is nearly identical to Example 1, and differs only in that $s^{c_1}(\lambda)$ represents the *median* ability of enrolled students instead of the *mean* ability.

Example 2. *Consider Example 1 but where $s^{c_1}(\lambda(\alpha))$ represents the median of abilities r^{θ,c_1} of enrolled students at the program given α, that is, $s^{c_1}(\lambda(\alpha)) = \sup\{r | \frac{\lambda^{c_1,r}(\alpha)}{\lambda^{c_1,1}(\alpha)} \leq \frac{1}{2}\}$.*

*The cutoff and median ability at $t = 1$ remains the same as in Example 1, given an upper bound on $s^{c_1}(\lambda_0)$: for $s^{c_1}(\lambda_0) \leq 1 - q^{c_1}$ it is the case that $(p_1^{c_1}, s_1^{c_1}) = (1 - q^{c_1}, 1 - \frac{q^{c_1}}{2})$. Additionally, $p_2^{c_1} = 1 - \frac{q^{c_1}}{2} - \frac{q^{c_1}}{2(1-k)}$. Note however that $s_2^{c_1} = s_1^{c_1} = 1 - \frac{q^{c_1}}{2}$; all of the students with abilities $r^{\theta,c_1} \geq 1 - \frac{q^{c_1}}{2}$ "return" to the program, and while the set of students who attend the program with abilities $r^{\theta,c_1} < 1 - \frac{q^{c_1}}{2}$ differs in periods 1 and 2, there are the same measure of them (filling exactly half of the seats), meaning that they do not affect the median. Therefore, $\succeq^{\theta|\mu_1} = \succeq^{\theta|\mu_2}$ for all $\theta \in \Theta$. By Assumption **A1** it must be that $\mu_2(\theta) = \mu_3(\theta)$ for almost all $\theta \in \Theta$. Therefore, $\lambda(\mu_2) = \lambda(\mu_3)$ and by Proposition 2, the TIM process produces a stable matching for all $t > 2$.*

Ex-ante, it may not have been intuitive that the change from "mean" to "median" would affect convergence to stability, so we turn to a comparison with tâtonnement processes in exchange economies for understanding. In Example 1, the market exhibits an analogue of the *individual* gross substitutes condition: students disprefer having lower ability than the mean ability of their peers, and therefore, the mean ability plays the role of the price in exchange economies. As the "price" of a program rises, each student's utility from attending it decreases. However, we have a failure of *aggregate* gross substitutes; due to capacity constraints and student unit demand, as the "price" rises, low scoring students are now able to attend the program, taking over seats from

[19] Because $p_2^{c_1} \in [0,1)$ by construction and $q^{c_1} \leq \frac{1}{3}$, c_1 fills all of its seats if $k \in [\frac{4}{5}, 1 - \frac{q^{c_1}}{2 - q^{c_1}}]$.

higher scoring students with weaker intrinsic preferences who decline to enroll due to the high "price." The increased demand of these low scoring students drives down the "price," thus leading to a cycle. The same complication does not occur in Example 2; because the median is not affected by outliers (and because the program fills all of its seats in each period). Therefore, (for the given starting condition μ_0) aggregate gross substitutes obtains and the "price" converges.

Our proof of Theorem 2 shows that a failure of aggregate gross substitutes can occur when peer preferences admit a *negative externality group*—informally, a set of students who reduce one another's utilities. When a negative externality group exists, student types can cycle in and out of programs, leading to a cycle in the TIM process as in Scarf (1960), which by Proposition 2 implies that the TIM process never finds a stable matching. The proof also shows that the existence of a negative externality group is likely from a topological perspective; a negative externality group exists generically in the space of peer preference functions. This can be understood as follows: for almost all collection of peer preferences, there are some students who are undesirable to others as peers.

In Fonseca et al. (2025), we present sufficient conditions for convergence of the TIM process to a stable matching. As discussed above, such conditions are joint restrictions on peer preferences and the market structure. First, we study a class of markets with positive network effects which do not admit negative externality groups.[20] Second, we pin down additional components of a market E that guarantee convergence of the TIM process. Intuitively, these conditions ensure the market exhibits a form of aggregate gross substitutes, even though negative externality groups exist, ensuring that our tâtonnement-like process convergences for any starting condition.

IV An Approximately Stable Mechanism

Given the theoretical finding in previous section that the status quo can lead to instability, we consider a mechanism design approach to find a stable matching. One challenge is that the standard approach would require soliciting student preferences as functions of the sets of students attending each program.[21] We instead present a constrained mechanism that does not rely on detailed information about the functional form of peer preferences and only requires students to

[20] Specifically, we study peer preferences that are monotonically increasing in the measure of students matched to a program. These preferences may represent peer preferences of some public school students in Japan (see https://web.archive.org/web/20230324050348/https://news.yahoo.co.jp/articles/124f47d03ca41a70512b5b39e2f04df8718f2c1a, accessed 7/30/2023). We show that these preferences do not admit a negative externality group, and in any market with $N < 3$ these preferences guarantee convergence of the TIM process for any starting condition μ_0. However, for $N \geq 3$ we show again that for any collection of peer preferences within this class, there exist markets and starting conditions for which the TIM process will not converge.

[21] Budish and Kessler (2021) suggest that students may be incapable of accurately stating functional preferences, and Carroll (2018) suggests that any such mechanism may be outside the consideration of centralized clearinghouses.

submit ROLs as in the status quo TIM process. This mechanism does not run across years, and instead attempts to find or approximate a stable matching for each cohort of students. Unlike the TIM process, it suffers neither from instability before reaching steady state, nor instability caused by changes in the market over time. Moreover, as we show, it yields or approximates a stable matching even when the TIM process does not converge. Reporting costs associated with this mechanism are low, in that the the vast majority of students need only report preferences one time.

Students in each cohort are assigned to one of many smaller submarkets, and students in each submarket submit ROLs sequentially. Students in each submarket are given different information regarding the ability distribution of students in each program. We use the subscript "t" to refer to a generic submarket below to be evocative of the time index in the TIM process.

We formalize the fundamentals of each submarket $E_t = [\eta_t, N, q_t, \Theta]$, $t \in \{1,...,T\}$. First, we specify the measure over students η_t. Let $\sum_{t=1}^{T} \eta_t(\Theta) = 1$, where each η_t is constructed "uniformly at random," that is, for any measurable set $\Theta^o \subset \Theta$, it is the case that $\eta_t(\Theta^o) = \eta(\Theta^o) \cdot \eta_t(\Theta)$. We require $\eta_t(\Theta) \to 0$ for all t as $T \to \infty$.

Second, we specify the programs. Each program $c \in C$ is active in each submarket, and has a submarket t specific capacity constraint $q_t^c = q^c \cdot \eta_t(\Theta)$. The capacity vector in submarket t is q_t.

Third, we define the ability distribution. Let $\mathcal{A}_t$ be the set of all assignments in market E_t. For each $x \in [0,1]$, $c \in C$, and $\alpha \in \mathcal{A}_t$ let $\lambda_t^{c,x}(\alpha) := \frac{\eta(\{\theta \in \alpha(c) | r^{\theta,c_0} \leq x\}}{\eta_t(\Theta)}$. The ability distribution in submarket t, $\lambda_t^c(\alpha)$, is the resulting non-decreasing function from $[0,1]$ to $[0,1]$, and let Λ_t be the set of all such functions, which is by construction equal to Λ. Let $\lambda_t(\alpha) := (\lambda_t^{c_1}(\alpha),...,\lambda_t^{c_N}(\alpha),\lambda_t^{c_0}(\alpha))$.

We now describe the proposed *Tâtonnement with Final Matching (TFM)* mechanism. The mechanism is initialized with a (finite) grid over the set $\Lambda(\mathcal{M}) := \{\lambda' \in \Lambda | \exists \mu \in \mathcal{M} \text{ s.t. } \lambda(\mu) = \lambda'\}$ i.e. the set of ability distributions associated with matchings in market E.[22] Each submarket is shown an ability distribution from this grid without replacement. The designer solicits ROLs over programs given this ability distribution in a (one-shot) stable matching mechanism. If the ability distribution from this resulting matching is far from the ability distribution shown to students in this submarket, the mechanism discards this ability distribution from the grid and repeats the process with the next submarket. Otherwise, the mechanism shows the same initial ability distribution to all students and constructs the final matching by soliciting ROLs over programs in a (one-shot) stable matching mechanism. The formal definition is given below.

Definition 4. *The* Tâtonnement with Final Matching (TFM) *mechanism is defined as follows:*

[22] Lemma A.4 in the appendix shows that Λ is compact, implying that such a grid always exists.

step 0: *Initialize the mechanism with $\delta > 0$, $\gamma > 0$, $T > 0$, and a finite subset $\Lambda_0^\gamma \subset \Lambda(\mathcal{M})$ where for each $\lambda \in \Lambda(\mathcal{M})$ there exists some $\lambda' \in \Lambda_0^\gamma$ such that $||\lambda - \lambda'||_\infty < \gamma$.*

step $\tau = K \cdot T + t$, $K \geq 0$, $t \in \{1,...,T\}$**:** *Report to students in submarket E_t some $\lambda_\tau \in \Lambda_{\tau-1}^\gamma$ and, via a one-shot stable mechanism, solicit ROLs over programs to create matching μ_τ. If $||\lambda_\tau - \lambda(\mu_\tau)||_\infty \geq \delta$ then let $\Lambda_\tau^\gamma = \Lambda_{\tau-1}^\gamma \setminus \{\lambda_\tau\}$ and go to step $\tau + 1$.*

At the first step τ such that $\|\lambda_\tau - \lambda(\mu_\tau)\|_\infty < \delta$, terminate the process above. Show all students in market E distribution vector λ_τ and, via a one-shot stable mechanism, solicit ROLs over programs to create final matching μ^{TFM} in aggregate market E. Otherwise, at the conclusion of step $\tau = |\Lambda_0^\gamma|$, assign all students to the outside option as the final matching.

The TFM mechanism depends on the following parameters: δ which defines the stopping criterion, γ which constructs the grid size, and T which determines how many times each subcohort of students is asked to report ROLs over programs (but does not affect the final matching generated).

For any $\delta > 0$ and any $T > 0$, there exists a grid size γ for which the TFM mechanism terminates when each student reports $\succeq^{\theta|\lambda_\tau}$ at step τ. Moreover, for any $\epsilon > 0$, we show that there exists a $\delta^* > 0$ such that for any positive $\delta < \delta^*$, the TFM mechanism terminates in an $\epsilon-$stable matching. The TFM mechanism also has desirable incentive properties. For any $\epsilon > 0$, we show that there exists a $\delta^* > 0$ such that for any positive $\delta < \delta^*$, it is an ϵ-Nash equilibrium for each student θ to reveal her "true" preferences $\succeq^{\theta|\lambda_\tau}$ whenever she is called upon to report an ROL. Because of this, the TFM mechanism potentially keeps the playingfield level between "sophisticated" students who submit ROLs best responding to the strategies of others, and "sincere" students who report truthfully. Finally, we show that there is sufficiently large T such that no student is asked to report an ROL more than twice, and an arbitrarily large share of students are asked only once. Recalling that T does not affect the final matching generated, this implies that for large enough T there are small additional reporting costs associated with this mechanism over canonical, one-shot mechanisms.[23]

Proposition 3. *Let $\epsilon > 0$, and suppose that each student θ reports $\succeq^{\theta|\lambda_\tau}$ at each step τ.*

1. *For any $\delta > 0$ and any $T > 0$ there exists $\gamma^* > 0$ such that for all $\gamma < \gamma^*$ and any associated grid Λ_0^γ, the TFM mechanism terminates (at or before step $|\Lambda_0^\gamma|$).*

[23] Small reporting costs arise straightforwardly in the TFM mechanism—i.e. a small fraction of market participants report their preferences twice—due to our large market assumption. This is because a continuum of students can be subdivided into any finite number of submarkets, each with a continuum of students. A market with a large but finite number of students cannot be similarly subdivided without bound, but qualitatively similar subdivisions, each with a large number of students but also with a small fraction of the overall student body, can easily be constructed if the number of students is sufficiently large.

2. *For any $\epsilon > 0$, there exists $\delta^* > 0$ such that for any positive $\delta < \delta^*$, any $T > 0$, and any Λ_0^γ for which the TFM mechanism terminates, μ^{TFM} is an $\epsilon-$stable matching.*

3. *For any $\epsilon > 0$, there exists $\delta^* > 0$ such that for any positive $\delta < \delta^*$, any $T > 0$, and any Λ_0^γ for which the TFM mechanism terminates, the measure of students who can improve their utility by reporting an ROL other than $\succeq^{\theta|\lambda_\tau}$ at any step τ is strictly less than ϵ, and no student θ can improve her payoff by more than ϵ by reporting an ROL other than $\succeq^{\theta|\lambda_\tau}$ at any step τ.*

4. *For any $\epsilon > 0$, any $\delta > 0$, and any Λ_0^γ for which the TFM mechanism terminates, there exists $T^* > 0$ such that for all $T > T^*$ the measure of students asked to report ROLs strictly more than twice is zero and the measure of students who are asked to report ROLs strictly more than once is strictly less than ϵ.*

Remark 5. *An alternative mechanism mimics the TIM process, but runs "within year," just as the TFM mechanism does. Instead of initializing with a grid Λ_0^γ, the alternative mechanism is initialized with an arbitrary μ_0, and each submarket is shown the ability distribution from the matching created in the prior submarket. The mechanism terminates when the ability distributions in subsequent submarkets are sufficiently close. Other details of the mechanism are as in the TFM mechanism.*

This alternative mechanism does not necessarily terminate. However, it terminates whenever the TIM process converges, and does so even in cases when the TIM process does not converge.[24] *As with the TFM mechanism, termination implies $\epsilon-$stability based on the magnitude of the stopping parameter δ (see point 2 of Proposition 3). Moreover, it inherits the incentive properties and low reporting costs associated with the TFM mechanism (see points 3 and 4 of Proposition 3).*

V Empirical Evidence of Peer Preferences

In this section, we describe details of the New South Wales (NSW) college admissions system and our administrative data from this market. We then discuss how we leverage this context and data to empirically identify preferences over the relative abilities of peer classmates, and present our main findings. Throughout, we let θ represent a generic student, and c a generic program.

[24] To see this point, consider Example 3 in the appendix. For any μ_0 where the mean ability of students assigned to the program is in the interval $(\frac{1}{2}-\delta, \frac{1}{2}+\delta)$ the alternative mechanism will immediately terminate, however, the TIM process converges only for starting conditions μ_0 such that the mean ability of students assigned to the program is exactly $\frac{1}{2}$.

V.A The New South Wales Tertiary Education Admissions System

We study college admissions in NSW (and the Australian Capital Territory) from 2003 to 2016.[25] Students apply for admission at the university-field level (for example, Economics at University of Sydney) through the Universities Admissions Centre (UAC), the centralized clearing-house which processes applications to all major universities in NSW. We refer to the university-field pairs as "programs."

Students receive a score known as the *Australian Tertiary Admission Rank (ATAR)* which measures the student's academic percentile rank, over a re-normalized scale of 30-99.95. The ATAR score is primarily determined from standardized testing, and students do not know their ATAR score at the onset of the application process. The ATAR score is a good predictor of academic performance during undergraduate studies (Manny et al., 2019). Therefore, we view the ATAR score as a proxy for student ability.

Each year, over 20,000 new high school graduates apply to programs where the ATAR score serves as the central admission criterion. To apply for admission, prospective students submit an ROL of up to nine programs to the UAC.[26] We refer to a student as an *applicant* to a program if she lists that program on her ROL. Students initially submit their ROLs before receiving their ATAR scores but can costlessly revise these lists after their ATAR scores become available. Students are incentivized to submit initial ROLs before ATAR scores are revealed, as doing so later incurs monetary penalties.

Students and programs are matched using the student-proposing deferred acceptance mechanism, which takes as inputs student ROLs, program rankings, and program capacities (Guillen et al., 2020).[27] Program rankings over students are determined by the sum of their ATAR score and program-specific bonus points awarded at each program's discretion. Importantly, the bonus points awarded to each student typically differ across programs, are not known in advance, and even the criteria for bonus points are typically not known to students.[28] The variability in bonus points can be

[25] A number of changes to the matching process have occurred since 2016. Namely, students are now only able to list five programs on their ROL, and there is now a "guaranteed entry" option for students with ATAR above a particular threshold (Guillen et al., 2020).

[26] A minority of students, such as adult learners who do not have ATAR scores, apply directly to programs.

[27] Admissions take place in multiple rounds. We describe and analyze the process of the main round that takes place in early January, when the majority of offers are made. There are initial rounds, where offers are made to some programs that do not admit based on the ATAR scores of students, and there are subsequent rounds for students that remain unmatched. As programs may elect not to enter subsequent rounds, there is a strong incentive for students to be matched to a desired program in the main round.

[28] Students are informed that "[a]t the request of our participating institutions, UAC does not release specific details of selection rank adjustments. Each institution has its own policy and will apply adjustment factors in accordance

large. Using data for two years, 2015-2016, we calculate that the average number of bonus points is 5.71 with a standard deviation of 6.54.[29] Therefore, bonus points introduce significant uncertainty into the admissions process, and significantly expand the set of programs for which a student has positive admissions chances. Moreover, and as we explore in our identification strategies, the matching mechanism used makes truthful reporting of preferences a weakly dominant strategy for many students (in the absence of peer preferences).

The resulting matching mechanically creates a minimum ATAR score above which students are "clearly in" (i.e. all students with ATARs above this level are admitted to the program regardless of the number of bonus points they receive if they are not admitted to a more preferred program) at the program level every year. We refer to the clearly-in statistic from the prior year's admitted cohort as the *Previous Year's Statistic (PYS)* and the analogous measure for the current cohort as the *Current Year's Statistic (CYS)*. In the absence of bonus points, the CYS would exactly equal the minimum required ATAR for program entry; in practice, however, the CYS closely aligns with the median ATAR of enrolled students (Bagshaw and Ting, 2016). Consequently, students typically understand that admission to a program is possible even if their ATAR is below the CYS.[30]

When creating their ROLs, students do not know the CYS at any program. However, students can consult each program's PYS as a guide; this information is prominently displayed, by law, on the clearinghouse website. For example, students applying for admission in 2016 are told the following:

> *[CYS] for 2015–16 admissions won't be known until selection is actually made during the offer rounds. Use [PYS] as a guide when deciding on your preferences."*[31]

Between 2003 and 2016—our analysis period—the PYS was the sole measure of peer ability from the prior cohort made available to applicants.

V.B Data

We use data from the UAC clearinghouse for our analysis. Our data contain the universe of applications from graduating high school students processed by UAC for 2003-2016. We identify

with its own schemes," see https://www.uac.edu.au/future-applicants/faqs-and-forms/educational-access-schemes, accessed 9/15/2023.

[29] See Appendix C for details.

[30] UAC reports that, "[m]ost Year 12 students are also aware that the [CYS] is inclusive of bonus points, and therefore does not necessarily represent the lowest ATAR required for the course...the selection rank is made up of more than just ATAR for most applicants," see https://www.uac.edu.au/assets/documents/submissions/transparency-of-higher-education-admissions-processes.pdf, accessed 9/15/2023.

[31] See https://web.archive.org/web/20150911225257/http://www.uac.edu.au/atar/cut-offs.shtml, accessed 9/6/2021.

each student via a unique student identification number. During this period, there are on average 19 universities active per year, each offering numerous programs. We identify and track programs over time using a unique course code.[32] For a subset of years (2010-2016) we observe students' ROLs at two points in time: immediately before they receive their ATAR score (which we call the pre-ROL), and the final list submitted to the clearinghouse after learning their score (which we call the post-ROL). Roughly one month separates our observation of these two ROLs. We observe the post-ROL for all years in our sample (2003-2016). In addition, we observe the students' ATAR scores, detailed information about each program they applied to (field of study, university, and location), and the CYS of each program.

We observe the final assigned program of each student in two years, 2015-2016. With this data, we can estimate both admission likelihoods and the distribution of bonus points. Appendix C shows how we calculate the distribution of bonus points, and Section V.C.2 uses this data to understand rank-order list formation and admission likelihoods.

Finally, we collect data on attrition rates for students commencing studies in a given year, calculated separately by university and each of 12 broad fields of study. The data are provided by the Australian Government Department of Education.[33] To link these attrition records with our UAC records, we aggregate UAC program measures to the year-university-broad field of study. The broad field of study is available for all programs in our dataset.

V.C Empirical analysis of peer preferences

V.C.1 Parameter of interest

In this section, we seek to identify applicant preferences over observable peer ability, as measured by the program PYS, at the time students submit ROLs. Recall that we allow these peer preferences to stem from various sources; they could reflect direct preferences over the peer population itself, such as for social connections or study partners, or more indirectly through downstream considerations, such as in zero-sum grading or competition for recommendations. This definition of peer preferences tracks closely to the assumptions employed in Section II. Our definition explicitly excludes preference for peers as a signal of university characteristics themselves, such as fixed teacher

[32] Prior to 2008, the same program could be listed twice according to its funding structure. The course code allows us to separately identify Commonwealth Supported Place (CSP) programs, which are subsidized, from Domestic Fee Paying (DFEE) programs. In 2008, all fee structures were standardized and all courses became CSP. In what follows, we treat DFEE courses as separate programs, but all of our results are robust to dropping DFEE programs.

[33] The data are publicly available via the Australian Government Department of Education's interactive portal: https://app.powerbi.com/view?r=eyJrIjoiNTA4MTZjZmMtZjRjNS00NzcxLWEzZTktODZmNDZkNGEwM2Y4IiwidCI6ImRkMGNmZDE1LTQ1NTgtNGIxMi04YmFkLWVhMjY5ODRmYzQxNyJ9.

quality or prestige, or student-program "fit" (Rothstein and Yoon, 2008), which we address later.

V.C.2 Motivating Evidence and Non-Peer Preferences

Table A.1 displays summary statistics of the data, including student ATAR scores, the PYS of the programs listed on student ROLs, and "score gaps," defined as the difference between the PYS of a ranked program and the student's ATAR. Panel A shows the sample containing pre- and post-ROLs (2010–2016) and Panel B shows the sample containing post-ROLs only (2003–2016). The summary statistics are consistent across samples and highlight several notable patterns. First, students tend to include programs with PYS scores higher than their own; for instance, the average score gap is 5.8 in the post-ROL sample. Second, students demonstrate an even stronger preference for "reach"-type programs at the top of their lists, with an average score gap of 8 in the same sample. The vast majority of students list at least one program with a positive score gap.

The fact that students are willing to apply to "reach" schools is suggestive evidence that they may not weigh admissions probabilities when forming ROLs. Agnosticism toward admissions probabilities is entirely consistent with the incentive properties of the deferred acceptance mechanism in matching markets without peer preferences; students who prefer weakly fewer than nine programs to their outside option have a weakly dominant strategy to list their true preferences (Haeringer and Klijn, 2009). Indeed, the clearinghouse clearly informs students it is in their best interest to submit truthful ROLs:

> *"Your chance of being selected for a particular course is not decreased because you placed a course as a lower order preference. Similarly, you won't be selected for a course just because you entered that course as a higher order preference. Place the course you would like to do most at the top, your next most preferred second and so on down the list...If you're interested in several courses, enter the course codes in order of preference up to the maximum of nine course preferences."*[34]

Nevertheless, a key alternative mechanism we initially consider is that students use observable peer information as a proxy for admissions chances, and not because of preferences over peers. A growing literature presents theoretical reasons why students may directly consider admissions chances. For instance, students may fail to fully understand the strategic properties of the matching mechanism (Li, 2017), may exhibit non-classical preferences such as loss aversion (Dreyfuss et al.,

[34] See https://web.archive.org/web/20150918170643/http://www.uac.edu.au/undergraduate/apply/course-preferences.shtml, accessed 9/6/2021.

2021; Meisner, 2022; Meisner and von Wangenheim, 2023), or may optimally acquire costly information about programs where they have higher admission probabilities due to uncertainty about their preferences (Immorlica et al., 2020; Grenet et al., 2022; Hakimov et al., 2023). Appendix D highlights a shared prediction of these proposed mechanisms: when the admission probability to a program exhibits a discontinuous decrease, the likelihood of ranking that program will also discontinuously decline.

We investigate this alternative mechanism by testing whether students respond to discontinuities in admission probabilities by altering their program choices.[35] We are able to do so because the distribution of bonus points, which affects admission probabilities in NSW, is not smooth. As we show in Appendix C, programs disproportionately award bonus points at "round" numbers (e.g., 0 or 5) rather than at other values (e.g., 1 or 6), creating admission discontinuities among students with small ATAR score differences. If student stated preferences reflect admission probability concerns, we should observe corresponding discontinuities in program choices at these thresholds, assuming a continuous distribution of student preferences over programs with similar PYSs.

We operationalize this hypothesis by conducting McCrary (2008)-type tests for bunching at these thresholds, estimating equations of the form:

$$Y = \alpha + \sum_{k \in \{0,5\}} \gamma_k \mathbb{1}_{\{X>k\}} + \sum_{k \in \{0,5\}} \delta_k \mathbb{1}_{\{X>k\}}(X - k - 0.5) + \phi X + \eta, \tag{1}$$

where Y is the proportion of applicants who rank a program with a given score difference at the top of their ROL. The running variable, X, is defined as the score difference between the program's PYS and the student's ATAR.

In this specification, the model includes three linear slope terms, captured by coefficients ϕ, δ_0, and δ_5, and two discrete level shifts at thresholds $k \in \{0,5\}$, captured by coefficients γ_0 and γ_5. The linear slope terms, centered around $(X - k - 0.5)$, ensure that the level shifts correspond precisely to midpoints between adjacent slopes. Given the discrete nature of the running variable, this model forms a parametrized regression discontinuity design (Clark and Del Bono, 2016; Rose and Shem-Tov, 2021).

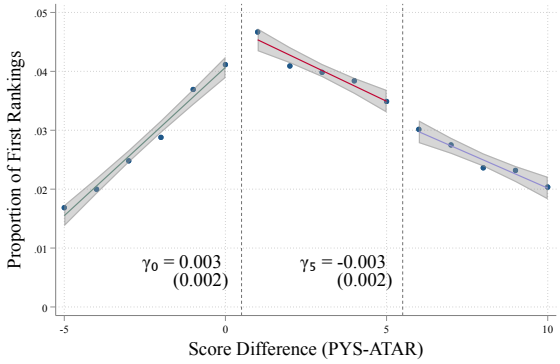
Panel A of Figure 1 presents the results. The linear slopes closely match the observed data, and there is minimal visual or statistical evidence of discontinuities at the thresholds connecting these linear segments. At the first discontinuity (score gap = 0), we estimate a small but marginally

[35] A number of papers make use of residential discontinuities to perform similar tests (e.g., Agarwal and Somaini, 2018).

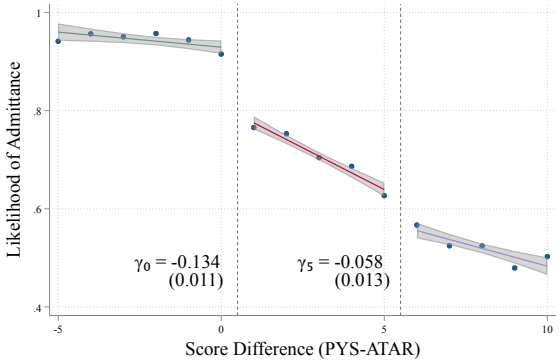
statistically significant *positive* effect on the proportion of students ranking a program just above the threshold ($\gamma_0 = 0.003,\ p = 0.05$). Models suggesting preferences over admissions concerns would predict a negative coefficient. At the second discontinuity (score gap = 5), we estimate an insignificant negative effect ($\gamma_5 = -0.003,\ p = 0.13$).

Figure 1: RD relationship between admission probability and choice

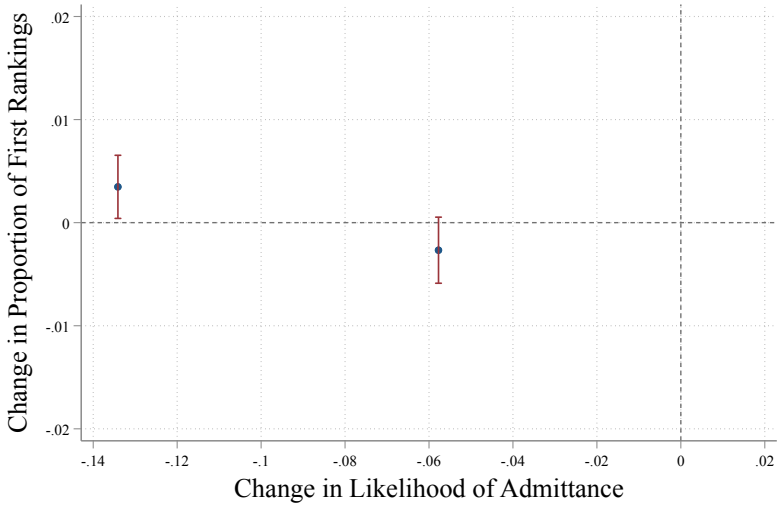
(a) Choice



(b) Admission



(c) Visual IV



Panel A plots the relationship between the proportion of students ranking a program first on their ROL and the program-student score difference, linearly fitted. Panel B plots a similar relationship with the outcome of admittance to this program, estimated at the student level. The program-student score difference is the gap between the program's PYS and the student's ATAR score. Each dot represents the average share of students being admitted or choosing a corresponding score bin. Panel C plots the estimated change in program choice and corresponding 95% confidence intervals (Panel A) by the estimated change in admission probability (Panel B) for the two score difference thresholds. The sample in all panels is restricted to students who first-rank a program within this range.

To relate these findings more directly to admission concerns, Panel B estimates changes in admission probabilities at score gaps of 0 and 5 using a student-level regression. That is, the

outcome variable is an indicator for whether the student was admitted. At a score gap of 0, the probability of admission decreases by 13 percentage points, whereas at a score gap of 5, the probability of admission decreases by a smaller 6 percentage points. Both effects are statistically significant at the 1% level. Panel C presents a Visual Instrumental Variables (VIV)-type plot linking the estimated changes in admission probabilities (from Panel B) to the corresponding changes in program choices (from Panel A). If students were selecting away from programs where they have low admissions probabilities, we would expect the VIV plot to 1) show negative coefficients on the vertical axis, and 2) that the vertical values would be more negative the further away from 0. Neither of these patterns are consistent with the data.

Our analysis above finds little support for the hypothesis that students incorporate admission probabilities when forming their ROLs. Students may not alter program choices in response to admissions discontinuities due to the incentive properties of the mechanism or due to unawareness of the admission discontinuities. Both of these reasons would support an analysis that assumes students do not alter choices in response to admissions probabilities.

Consequently, we proceed in the following section to investigate the presence of peer preferences. Our main identification strategy exploits aspects of the institutional context and incentive properties of the matching mechanism to isolate variation in which admissions probabilities cannot confound our analysis unless students submit weakly dominated ROLs.

V.C.3 Estimating Peer Preferences

By law, when making application decisions, students must be shown each program's PYS. We seek to estimate preferences over the PYS using student application decisions, under the assumption that students anticipate the PYS reflects their future peers.[36] Interpreting student responses as revealing their preferences requires careful consideration of students' incentives for truthfully reporting their preferences, as induced by the matching mechanism. Our data and institutional context enable two distinct research designs to do so.

In Appendix E, we use variation in programs' PYSs over time to study applicant demand for peers using a transparent event-study design. We find that increasing overall peer ability induces changes in application behavior, including a decrease in demand from students with lower ATAR scores but no change for students with higher scores. Two assumptions are necessary. First, to interpret these results as causal requires a parallel trends assumption: demand must have evolved

[36] Recall that the matching clearinghouse explicitly instructs students to use the PYS as a signal of their future peers when reporting their preferences (see page 24), and we make a similar assumption in our theoretical framework. We formalize this assumption when discussing our identification strategies below.

similarly for programs with and without an increase in observable peer ability. This assumption is plausible given evidence presented in Appendix E demonstrating that past student demand is unlikely to be a confounder. Second, to interpret our results as reflecting student preferences, we assume student ROLs reflect their true rankings over programs (given peer ability as revealed by the PYS). Recall that the matching mechanism used by the clearinghouse makes truthful reporting (i.e., in descending order of preference) a *weakly dominant* strategy for students whose number of acceptable programs does not exceed the cap on listed choices (Haeringer and Klijn, 2009). We assume that students play weakly dominant strategies when available to them and restrict the sample to students unconstrained by the cap on the number of programs, a common approach in the literature (e.g. Hastings et al., 2009; Abdulkadiroğlu et al., 2017; Luflade, 2019).

For our second and primary design, we investigate peer preferences by exploiting a unique feature of the NSW market and data: we observe students' ROLs at two points in time—both before (pre-ROL) and after (post-ROL) they learn their own ATAR score. Students are incentivized to submit preferences early through lower application fees, before learning their ATAR score, and the overwhelming majority (99.5%) of students do so. Subsequently, they can update their ROL without financial cost after learning their score.

Pairwise information contained in the two ROLs can be used to relax the assumption that students play their weakly dominant strategy to perfectly report their true preferences, which a recent literature has challenged.[37] For example, Fack et al. (2019) argue students face a cost of submitting longer ROLs. Therefore, if a student has low admission likelihood for a program (i.e., a "reach" program) then she may optimally omit that program from her ROL even if she prefers it to being unmatched. In this case, an important result from Haeringer and Klijn (2009) still applies: the relative ranking of any two programs c and c' on a student's ROL reflects her true ordinal preferences over c and c' in any *weakly undominated* strategy. We assume students play weakly undominated strategies, which allows us to interpret "switches" in the relative rankings between two programs as indicative of a change in ordinal preferences upon learning their ability relative to those of their peers. As we describe below, we further restrict our sample to students who neither add nor remove programs from their pre-ROLs to avoid confounds, focusing exclusively on students who "only switch" the relative rankings of programs from their pre- to post-ROLs. Note that the underlying basis for this identification strategy, the assumption that students play *weakly undominated* strategies, is weaker than the assumption that students play *weakly dominant* strategies,

[37] See Chen and Sönmez (2006); Li (2017); Rees-Jones (2018); Sóvágó and Shorrer (2018); Chen and Pereyra (2019); Larroucau and Rios (2020a); Hassidim et al. (2021); Artemov et al. (2023).

an advantage of the current identification strategy over the approach we explore in Appendix E.

To develop our approach for using changes to ROLs, we first define notation. Let $U^{\pi}_{\theta c}$ denote student θ's (perceived) utility from enrolling in program $c \in C$ in stage $\pi \in \{0,1\}$, where $\pi = 0$ corresponds to the stage prior to the revelation of students' ATAR scores, and $\pi = 1$ corresponds to the stage after. Throughout, we assume preferences are strict. Let R^{π}_{θ} denote student θ's ROL in stage π. The first-ranked program is

$$R^{\pi}_{\theta 1} = \underset{c \in C}{\operatorname{argmax}}\, U^{\pi}_{\theta c}$$

and subsequent choices ℓ are constructed recursively:

$$R^{\pi}_{\theta \ell} = \underset{c \in C \setminus \{R^{\pi}_{\theta m} | m < \ell\}}{\operatorname{argmax}} U^{\pi}_{\theta c}.$$

We primarily focus on post-ROL choices, R^{1}_{θ}, while accounting for the patterns established by the pre-ROL, R^{0}_{θ}. Denote the pre-ROL ranking of program c by student θ as $r_{\theta}(c)$. Our baseline assumption is that utility has the following form:

$$U^{1}_{\theta c} = \psi_{r_{\theta}(c)} + \sum_{j=1}^{J} \alpha_j (PYS_c - S_{\theta})^j + \varepsilon_{\theta c}. \tag{2}$$

This utility function is additively composed of multiple components. First, students have mean preferences for unobserved program characteristics, which we estimate as a function of their past ranking, $\psi_{r_{\theta}(c)}$. To ensure well-defined measures of past rankings and to implement the pairwise test outlined above, we restrict the sample to students listing identical sets of programs in both their pre- and post-ROLs (but potentially in a different order).[38] That is, letting $C^{\pi}_{\theta} := \{c | c \in R^{\pi}_{\theta}\}$ denote the set of programs listed on θ's stage-π ROL, we focus exclusively on students with $C^{0}_{\theta} = C^{1}_{\theta}$. Thus, the variation in our outcomes arises from switches in the relative ranking of programs.

The second component of utility pertains to preferences for relative peer ability. We allow students to have flexible preferences for program c's PYS, denoted PYS_c, relative to their own ability measured by their ATAR score, S_{θ}. The polynomial terms allow us to flexibly represent any functional form of student preferences over these fundamentals for sufficiently large degree J.

The new information we exploit between stage 0 and stage 1 is the revelation of student θ's

[38] We keep more than 50% of the sample after imposing this restriction. The samples appear to be relatively similar with the average ATAR in the full sample being 72 as compared to the 70.4 in the no add-drop sample.

ATAR score. For this score to affect ROLs there must be a change in students' beliefs about their own ability. This behavior can be formally motivated by a simple belief-updating model. We consider mutually-exclusive groups g as comprising students θ with similar preferences. Alternative definitions of these groups are explored empirically. Letting $g(\theta)$ represent the group θ belongs to, we assume stage-0 utility takes the form

$$U^0_{\theta c} = \phi^0_{g(\theta)c} + \sum_{j=1}^{J} \alpha_j (PYS_c - B^0_{g(\theta)})^j + \varepsilon^0_{\theta c}.$$

This specification defines both mean preferences for non-peer characteristics at program c, $\phi^0_{g(\theta)c}$, and (degenerate) beliefs about own score, $B^0_{g(\theta)}$, as homogeneous within group g. Between stages 0 and 1, we assume students update their beliefs perfectly to their realized ATAR, S_θ, and their preferences may shift systematically due to factors common to programs that occupy the same rank r in their initial (stage-0) ROL. We thus represent deterministic preference heterogeneity at stage-1 as $\phi^0_{g(\theta)c} + d_{g(\theta)r_\theta(c)}$, encompassing both program-specific amenities and systematic shifts shared within group-by-rank cells.

We account for this preference heterogeneity empirically using observed choices from stage-0. Our key identifying assumption, formally stated and detailed in Appendix F, is that conditioning on group-by-rank fixed effects, which we denote $\psi_{g(\theta)r_\theta(c)}$, effectively eliminates deterministic preference variation across programs. Intuitively, conditioning on both group membership and the initial rank significantly reduces systematic within-cell variation in amenities and peer expectations, making residual variation plausibly small and ignorable. To understand this assumption more clearly, consider the limiting scenario in which we define students as being in the same group if and only if they have exactly the same pre-ROL. Then, all deterministic variation is absorbed by these fine-grained fixed effects, leaving no residual variation. Our practical approach, using broader group-by-rank cells, approximates this idealized scenario. Formally, we write our estimating equation as:

$$U^1_{\theta c} = \psi_{g(\theta)r_\theta(c)} + \sum_{j=1}^{J} \alpha_j (PYS_c - S_\theta)^j + \varepsilon^1_{\theta c}.$$

In our most parsimonious model, corresponding to Equation 2, we assume $\psi_{g(\theta)r_\theta(c)} = \psi_{r_\theta(c)}$; that is, we account for stage-0 preference heterogeneity by assuming all students are in the same group and therefore condition only on the stage-0 ROL ranking of program c. To assess whether our partition of students into groups adequately accounts for preference heterogeneity, we progressively

create more fine-grained group definitions. As we move toward finer group definitions in our empirical analysis, we approximate the limiting scenario of complete initial rank conditioning and can empirically assess whether residual taste heterogeneity is negligible in practice. This design and empirical approach is similar to the analysis of updating neighborhood choice in Ferreira and Wong (2023), though their derivation of information updating relies on the Index Sufficiency assumption introduced by Dahl (2002).

Assuming the errors are distributed extreme-value type I, this choice problem takes the form of an exploded logit (Beggs et al., 1981). We estimate this model using maximum likelihood, setting the polynomial degree to $J=3$ in Equation 2. As discussed, we restrict our sample to applicants who neither add nor remove programs from their ROLs before and after ATAR score revelation, focusing exclusively on relative ranking decisions. Further, we limit the analysis to students who listed at least three program choices. To interpret the results, we present predicted probabilities of a program being ranked first, second, or third as a function of the score difference, $PYS_c - S_\theta$.

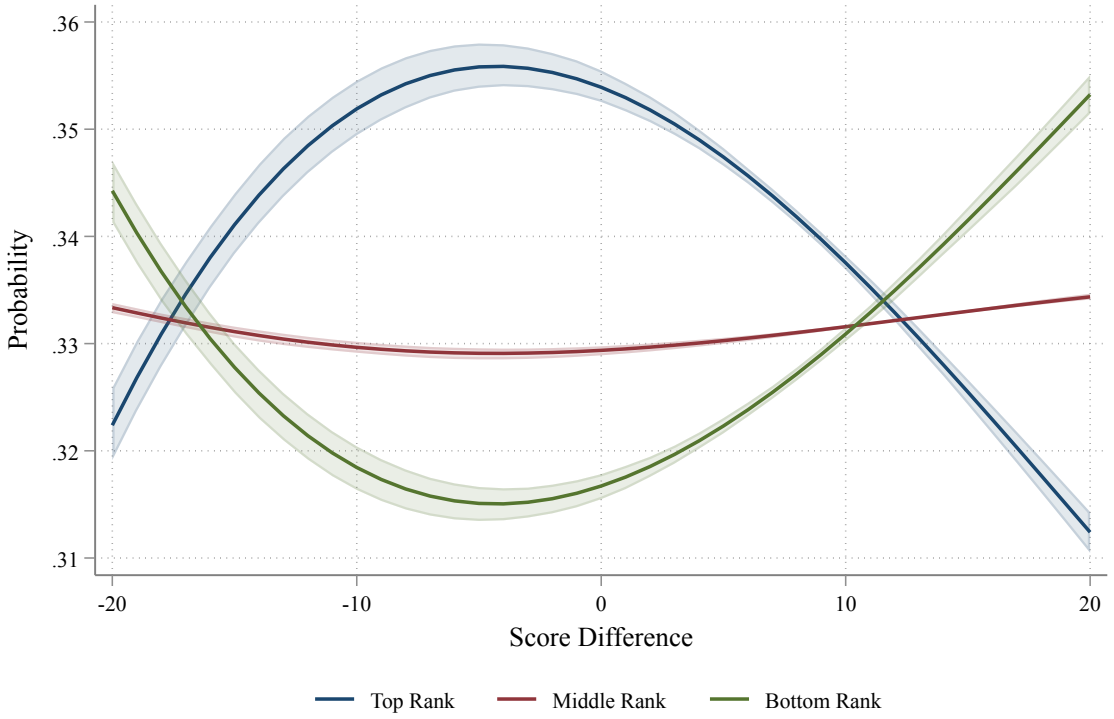
Figure 2 presents the estimated preferences based on our baseline specification, controlling for students' initial (stage-0) rankings as outlined in Equation 2. Our findings reveal that, on average, students exhibit a clear preference for programs where their own ATAR score slightly exceeds the PYS. Specifically, the most preferred relative position corresponds to a situation where a student's score is approximately 5 points above the program's PYS, suggesting a strong "big fish" preference—students prefer being among the higher-performing peers within their chosen program. Interestingly, on average students typically list programs with PYSs around 6 points higher than their own ATAR on their ROL (see Table A.1). This pattern suggests that factors beyond peer comparisons—such as program prestige or additional amenities—significantly influence their choices and mask the component of relative peer preferences identified here.

The magnitude of the most preferred relative score difference (approximately -5) is notably distant from score thresholds where strategic ranking behavior due to admission probabilities would typically occur. Students do not merely prefer marginally safer choices to ensure admission, but rather exhibit a preference for programs where their relative peer ranking is comfortably higher. This behavior strongly suggests intrinsic preferences for peer status rather than strategic admissions considerations.

We also confirm the robustness of these results by examining polynomial flexibility in our utility specification. Figure A.2 demonstrates nearly identical choice predictions when extending the model to include a fourth-degree polynomial term in the score difference. Overall, these empirical results provide compelling evidence that peer-related considerations meaningfully shape

student program choices in school choice markets.

Figure 2: Updating program choices on relative scores



This figure plots the probabilities for a program to be rank 1, rank 2, or rank 3 on the post-ROL by the score difference between the program PYS and the student ATAR. We compute predicted probabilities for being ranked first, second, or third by evaluating the estimated rank-order logit on a grid of covariate values $x \in [-20,20]$. For each grid point, we replace the alternative's covariate with x, compute the sequential choice probabilities implied by the model, and then average those probabilities across all observations. Rank 3 is obtained as one minus the sum of the rank 1 and rank 2 probabilities. The rank-order logit estimates come from column (2) in Table A.2. Bootstrapped 95% CIs calculated from 200 draws.

Table A.2 presents exploded-logit coefficients across various specifications. Column (1), which does not include stage-0 ROL rank dummies, indicates a strong preference for higher relative peers. However, this result likely reflects correlated program-specific unobservables. Column (2), following the utility specification from Equation 2, instead, most notably, yields a negative coefficient on the linear relative score difference term, underscoring the necessity of accounting for past ranking information.

Columns (3-7) introduce further heterogeneity in preferences. Column (3) adds variables capturing the share of the population who added each program at each stage-0 ROL rank, thus controlling for aggregate demand effects. Columns (4-6) allow further differentiation of stage-0 ranks: by individual stage-0 ROL length (column 4), by university (column 5), and by program field-of-study (column 6). Finally, column (7) provides the most flexible specification, allowing

stage-0 ranks to vary based on the field-university combination of a student's top-ranked program at stage 0. This approach accommodates significant heterogeneity, with eight stage-0 ranks allowed to vary with over 180 potential top-choice university-field combinations.

Returning to the empirical design question—whether group-by-rank covariates sufficiently capture stage-0 preference heterogeneity—the notable robustness of our peer-preference results across various group and program definitions suggests minimal residual heterogeneity. Across all specifications, we consistently observe similar patterns in relative peer preferences.

Two key assumptions are required to interpret these results as evidence of relative peer preferences. First, as discussed in Section V.C.1, we assume students use the PYS to update about peers rather than other program characteristics such as academic rigor or difficulty.[39] A straightforward prediction of the program-learning mechanism is that preferences will be less elastic to changes in the PYS for programs that students are initially more informed about. Column (8) in Table A.2 re-estimates our base specification allowing differential effects for programs that are newer as opposed to older, based on the notion that students have more information about older programs. The absence of differential responsiveness supports our assumption, indicating students perceive the PYS predominantly as reflecting peer composition rather than other unobserved program attributes.

Second, we assume pre-ROLs reflect relative student preferences over any two ranked programs prior to learning their ATAR score, instead of being mere placeholders. Recalling that the matching mechanism incentivizes truthtelling on the post-ROL, we believe this assumption is justified as there is a high correlation between students' pre- and post-ROLs.[40] Specifically, Table A.1 shows that the average program PYS and average score gap are comparable across pre- and post-ROLs, suggesting students likely construct pre-ROLs using a similar process to their post-ROLs. Moreover, we would not expect to observe systematic patterns in the changes from the pre- to the post-ROL if students were forming the pre-ROL "at random."

V.D Assessing market stability

Peer preferences affect stability of the matching if, as our theoretical model explains, characteristics of realized peers differ from those of expected peers. Following Proposition 2, a natural empirical test of market stability is to examine whether the difference between each program's CYS and PYS equals or converges to zero.

Figure 3 plots the average absolute difference between the CYS and PYS over time. The figure

[39] Rothstein and Yoon (2008) discusses how students may be wary of their fit or mismatch with a program.

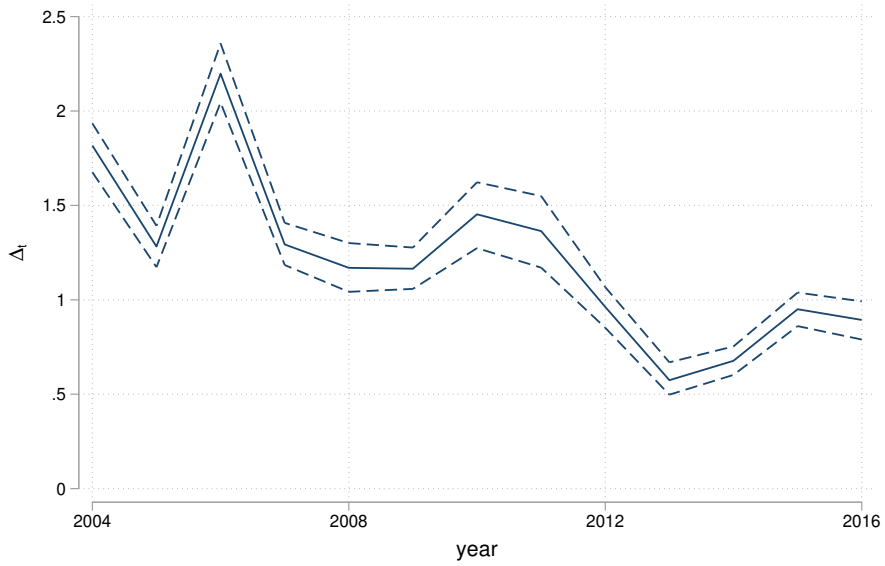
[40] 50% of students submit different pre- and post-ROLs, with 33% of rankings changing between the two ROLs on average.

shows a substantial difference between these measures throughout the study period. The absolute difference hovers between 1.5 and 2 points in the earliest available years, gradually declines, and then rises again toward the end of our sample period. Although this initial decrease suggests a reduction in market instability, the data decisively reject the null hypothesis of zero difference in all years. This pattern suggests persistent instability in the market.

Instability can affect student welfare and market functioning through multiple channels, many of which are difficult to measure directly. We focus on one important and observable consequence of instability: program attrition. Attrition provides a particularly clear measure of the welfare impacts of instability, since failing to complete a program directly affects human capital accumulation and labor market outcomes. In our theoretical framework, whenever a blocking pair is consummated—either with another program or a student's outside option—attrition occurs. Consequently, we expect attrition to be higher at programs with larger gaps between observable and actual peers, measured as the difference between CYS and PYS. However, since some blocking pairs remain unconsummated due to fixed costs (e.g., moving), our measure predominantly captures pairs yielding substantial utility gains, thus undercounting total welfare costs from instability.

We define attrition as occurring when a student who commences in year t does not com-

Figure 3: Absolute Difference between CYS and PYS over time



This figure presents the absolute difference between the current year statistic (CYS) and the past year statistic (PYS) for all years. This difference is calculated as $\Delta_t = \frac{\sum_c |CYS_{t,c} - PYS_{t,c}|}{|C_t|}$ where $|C_t|$ is the number of programs in year t. The dotted line shows bootstrapped 95% CIs taken from 1000 draws over the parameter estimates of Δ_t.

plete the program by the end of year $t+4$; otherwise, we consider the student to have completed the program.[41] For privacy reasons, we observe completion at the university-year-broad-field level.[42] To match this data aggregation, we add together the university-year-program specific values $CYS_{t,c} - PYS_{t,c}$ to calculate a university-year-broad field of study level sum, $CYS_{t,b(c)} - PYS_{t,b(c)}$, where $b(c)$ represents the broad field of study for program c.

Table 1: Relationship Between Completion Rate and (PYS-CYS)

	Completion rate (%)					
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. PYS-CYS	-.15*	-.12*	-.14**			
	(.078)	(.064)	(.069)			
Avg. PYS-CYS, positive				-.25***	-.18*	-.19**
				(.094)	(.092)	(.093)
Linear field-of-study trends	No	Yes	No	No	Yes	No
Year by field-of-study fixed effects	No	No	Yes	No	No	Yes
N	1399	1399	1399	1399	1399	1399

This table presents the relationship between the 4-year completion rate of commencing students and average PYS-CYS. The university-program-year-specific PYS-CYS is estimated estimated as a weighted-average. All columns control for year and university-program fixed effects in a two-way fixed effects specification, with columns (2) and (3) and (5) and (6) additionally controlling for linear field of study year trends and field of study-year fixed effects, respectively. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 1 examines the relationship between completion rates and differences between observed and realized peer ability over time. We conduct this analysis in two ways. First, we examine all changes in observed peer differences. Second, we focus exclusively on positive increases in peer differences, capturing students' asymmetric responses to peers with relatively higher scores.[43] The latter approach aligns with our findings in Section V.C.3, which indicate greater disutility for programs with higher-scoring peers.

Using a two-way fixed effects design that controls for year and university-by-broad-field-of-study fixed effects, column (1) shows marginally statistically significant evidence that a one unit

[41] We cannot distinguish between a student who permanently attrits and a student who "temporarily" attrits and completes the program after the 4-year period. However, as either outcome likely leads to efficiency costs, we believe our upcoming analysis sheds an important light on the role of blocking pairs caused by peer preferences on ex-post market outcomes.

[42] As discussed in V.B, there are 12 broad fields of study, including fields such as "Management and Commerce" and "Natural and Physical Sciences."

[43] We construct this treatment variable using the formula $\max\{0, CYS_{t,c} - PYS_{t,c}\}$ before aggregating to the broad-field-of-study level.

increase in the estimated $CYS_{t,b(c)} - PYS_{t,b(c)}$ decreases completion by 0.15 percent ($p = 0.05$).[44] More intuitively, a one standard deviation increase in $CYS_{t,b(c)} - PYS_{t,b(c)}$ corresponds to a 0.067 standard deviation decrease in completion rates after accounting for fixed effects. Columns (2-3) demonstrate robustness to differential trends, either through linear field-specific time trends ($p = 0.07$) or through field-year fixed effects ($p < 0.05$). These specifications mitigate concerns that gradual, field-specific changes or differential labor market shocks might influence our findings.

Columns (4-6) further investigate how completion rates vary specifically to positive increases in the average score difference. Section V.C.3 indicates that students prefer programs with peers scoring below themselves rather than substantially above. Consequently, attrition should respond more strongly to positive increases in the score difference. We find supportive evidence of this hypothesis: all coefficients are larger for positive score changes. These findings highlight substantial impacts of peer-driven instability on student attrition.

Notably, the attrition outcome studied here reflects only one dimension of the broader impacts of peer-driven instability. Benchmarking our findings provides context for interpreting this component. Dawson et al. (2021) reviews comprehensive interventions targeting college completion, highlighting substantial variability in impacts. Highly intensive interventions can significantly increase graduation rates (Weiss et al., 2019), while less intensive ones, such as informational advising or minimal financial incentives (e.g. Scrivener and Weiss, 2009; Alamuddin et al., 2018), typically have small or negligible effects. Our analysis, which does not explicitly target completion rates, finds modest but discernible impacts, aligning more closely with lighter-touch interventions rather than intensive comprehensive programs. Thus, the estimated effects in our study importantly signal meaningful consequences of instability, with attrition as one manifestation of the broader impacts.

An important caveat to our findings is the potential for compositional and academic rigor effects. Specifically, higher peer-score differences might attract better-prepared students who are less likely to attrit, which would bias our negative effect associated with peer-driven instability towards zero. Conversely, programs with higher peer scores may also become academically more demanding, independently increasing attrition. Hence, our estimated effects could either understate or overstate the true impact of peer-driven instability on attrition. In practice, we observe a positive correlation between completion rates and average university-field cutoffs ($R^2 = 0.15$), suggesting that, if anything, our estimates likely undercount impacts on attrition.

Together, the robustness and magnitude of our results indicate that peer-driven instability significantly affects students' educational trajectories. Our findings underscore the potential economic

[44] We cluster at the university-field level.

importance of reducing mismatches between expected and actual peer composition.

VI Conclusion

Our analysis provides new evidence on the presence of peer preferences and the barriers they pose to constructing a stable matching in status quo school-choice markets. We show that the status quo process used in markets around the world of revealing peer information from the previous entering class and then instructing students to "rank their true preferences" is not a reliable method for ensuring stability.

Using data from the NSW college admissions market, we show the empirical importance of peer preferences. Students exhibit preferences over relative peer comparisons. We develop and reject a simple test of whether the matching generated for any given cohort is stable and show this instability is associated with the observable consequence of greater attrition.

Historically, matching markets have responded to instability by redesigning the matching mechanism (Roth, 2002). Most relatedly, in a market where participants have preferences over spouses—a form of peer preferences distinct from the one we study—instability led to a redesign of the matching mechanism in use (Roth and Peranson, 1999). We similarly propose a new mechanism for use in school choice markets. This mechanism is a relatively small modification to iterative mechanisms already in use in higher education markets in China, Brazil, Germany, and Tunisia (see Luflade, 2019; Bo and Hakimov, 2022) and finds a (approximately) stable matching in the presence of peer preferences. Moreover, it incentivizes truthful reporting, thus leveling the playing field between strategically sophisticated and unsophisticated students, and imposes low reporting costs.

Our analysis, both theoretical and empirical, focuses solely on one peer characteristic: ability. As discussed, our theoretical framework extends easily to include preferences over other characteristics, and in such settings, incorporating information about all payoff-relevant dimensions of peer characteristics ensures a stable matching. In our empirical setting, however, we are restricted by the institutional context to studying preferences only over ability. In principle, one concern could be that our finding of "big fish in a little pond" preferences over ability is misattributed, and students care about other peer characteristics. However, our attrition results provide ex-post evidence that students indeed disprefer matching to programs where they are overmatched by their peers in terms of ability.

Our work also suggests caution ought to be applied when considering the impacts of policy changes in school choice markets, even those not directly targeted at accounting for peer preferences. Empirical papers frequently estimate student preferences for use in counterfactual analyses (e.g., preferences are estimated prior to a proposed policy change aimed at increasing

representation of specific groups of students). While inferring the impact of any policy change is difficult due to omitted variables in the estimation of preferences, any counterfactual policy which affects the matching will necessarily change student peers, potentially changing student preference rankings over programs. As a result, a full understanding of the equilibrium effects of a policy change requires consideration of how student preferences over programs will be affected by the corresponding change in peers.

References

Abdulkadiroğlu, A., N. Agarwal, and P. A. Pathak (2017). The welfare effects of coordinated assignment: Evidence from the new york city high school match. *American Economic Review 107*(12), 3635–3689.

Abdulkadiroğlu, A., Y.-K. Che, and Y. Yasuda (2015). Expanding "choice" in school choice. *American Economic Journal: Microeconomics 7*(1), 1–42.

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2020). Do parents value school effectiveness? *American Economic Review 110*(5), 1502–39.

Abdulkadiroğlu, A. and T. Sönmez (2003). School choice: A mechanism design approach. *American Economic Review 93*(3), 729–747.

Agarwal, N. and P. Somaini (2018). Demand analysis using strategic reports: An application to a school choice mechanism. *Econometrica 86*(2), 391–444.

Ainsworth, R., R. Dehejia, C. Pop-Eleches, and M. Urquiola (2023). Why do households leave school value added on the table? the roles of information and preferences. *American Economic Review 113*(4), 1049–1082.

Alamuddin, R., D. Rossman, and M. Kurzweil (2018). Monitoring advising analytics to promote success (MAAPS): Evaluation findings from the first year of implementation. Technical report, ITHAKA S+ R.

Allende, C. (2020). Competition under social interactions and the design of education policies. mimeo.

Artemov, G., Y.-K. Che, and Y. He (2023). Stable Matching with Mistaken Agents. *Journal of Political Economy Microeconomics 1*(2), 270–320.

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

Azmat, G. and N. Iriberri (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics 94*(7), 435–452.

Bagshaw, E. and I. Ting (2016). Nsw universities taking students with atars as low as 30. *The Sydney Morning Herald*.

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

Barlow, R. E. and H. D. Brunk (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association 67*(337), 140–147.

Beggs, S., S. Cardell, and J. Hausman (1981). Assessing the potential demand for electric cars. *Journal of Econometrics 17*(1), 1–19.

Berger, U. (2007). Brown's original fictitious play. *Journal of Economic Theory 135*(1), 572–578.

Beuermann, D. W. and C. K. Jackson (2022). The Short- and Long-Run Effects of Attending The Schools that Parents Prefer. *Journal of Human Resources 57*(3), 725–746.

Beuermann, D. W., C. K. Jackson, L. Navarro-Sola, and F. Pardo (2023). What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output. *Review of Economic Studies 90*(1), 65–101.

Bo, I. and R. Hakimov (2022). The iterative deferred acceptance mechanism. *Games and Economic Behavior 135*, 411–433.

Brown, G. W. (1951). Iterative solutions of games by fictitious play. In T. C. Koopmans (Ed.), *Activity Analysis of Production and Allocation*, pp. 374–376. Wiley.

Budish, E. and J. B. Kessler (2021). Can market participants report their preferences accurately (enough)? *Management Science, Forthcoming*.

Bykhovskaya, A. (2020). Stability in matching markets with peer effects. *Games and Economic Behavior 122*, 28–54.

Campos, C. (2024). Social interactions, information, and preferences for schools: Experimental evidence from los angeles. Technical report, National Bureau of Economic Research.

Card, D., A. Mas, E. Moretti, and E. Saez (2012). Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *American Economic Review 102*(6), 2981–3003.

Carroll, G. (2018). On mechanisms eliciting ordinal preferences. *Theoretical Economics 13*(3), 1275–1318.

Çelebi, O. (2022). Best-response dynamics in the boston mechanism. mimeo.

Che, Y.-K., D. W. Hahm, J. Kim, S.-J. Kim, and O. Tercieux (2022). Prestige seeking in college application and major choice. mimeo.

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

Chen, Y. and T. Sönmez (2006). School Choice: An Experimental Study. *Journal of Economic Theory 127*(1), 202–231.

Clark, D. and E. Del Bono (2016). The long-run effects of attending an elite school: Evidence from the united kingdom. *American Economic Journal: Applied Economics 8*(1), 150–176.

Dahl, G. B. (2002). Mobility and the return to education: Testing a roy model with multiple markets. *Econometrica 70*(6), 2367–2420.

Dawson, R. F., M. S. Kearney, and J. X. Sullivan (2021). Why expanded student supports can improve community college outcomes and boost skill attainment. *Brookings Institution*.

Dreyfuss, B., O. Heffetz, and M. Rabin (2021). Expectations-based loss aversion may help explain seemingly dominated choices in strategy-proof mechanisms. *American Economic Journal: Microeconomics, Forthcoming*.

Dur, U., R. G. Hammond, and O. Kesten (2021). Sequential school choice: Theory and evidence from the field and lab. *Journal of Economic Theory 198*, 105344.

Echenique, F. and M. B. Yenmez (2007). A solution to matching with preferences over colleagues. *Games and Economic Behavior 59*(1), 46–71.

Ellickson, B., B. Grodal, S. Scotchmer, and W. R. Zame (1999). Clubs and the market. *Econometrica 67*(5), 1185–1217.

Ergin, H. and T. Sönmez (2006). Games of school choice under the boston mechanism. *Journal of Public Economics 90*, 215–237.

Esponda, I. and D. Pouzo (2016). Berk-nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica 84*(3), 1093–1130.

Fack, G., J. Grenet, and Y. He (2019). Beyond truth-telling: Preference estimation with centralized school choice and college admissions. *American Economic Review 109*(4), 1486–1529.

Ferreira, F. and M. Wong (2023). Estimating preferences for neighborhood amenities under imperfect information. NBER WP 28165.

Fonseca, R., B. Pakzad-Hurson, and M. Pecenco (2025). Entry and exit in school choice markets. mimeo.

Frank, R. H. (1985). *Choosing the Right Pond: Human Behavior and the Quest for Status*. Oxford University Press.

Gale, D. and L. S. Shapley (1962). College admissions and the stability of marriage. *The American Mathematical Monthly 69*(1), 9–15.

Gentile Passaro, D., F. Kojima, and B. Pakzad-Hurson (2023). Equal pay for *Similar* work. mimeo.

Greinecker, M. and C. Kah (2021). Pairwise stable matching in large economies. *Econometrica 89*(6), 2929–2974.

Grenet, J., Y. He, and D. Kübler (2022). Preference discovery in university admissions: The case for dynamic multi-offer mechanisms. *Journal of Political Economy 130*(6), 1427–1716.

Grigoryan, A. (2021). School choice and the housing market. mimeo.

Grigoryan, A. (2022). On the convergence of deferred acceptance in large matching markets. mimeo.

Guillen, P., O. Kesten, A. Kiefer, and M. Melatos (2020). A field evaluation of a matching mechanism: University applicant behaviour in australia. *The University of Sidney Economics Working paper Series*.

Haeringer, G. and F. Klijn (2009). Constrained school choice. *Journal of Economic Theory 144*(5), 1921–47.

Hakimov, R., D. Kübler, and S. Pan (2023). Costly information acquisition in centralized matching markets. *Quantitative Economics 14*(4), 1447–1490.

Hassidim, A., A. Romm, and R. I. Shorrer (2021). The limits of incentives in economic matching procedures. *Management Science 67*(2), 951–963.

Hastings, J. S., T. J. Kane, and D. O. Staiger (2009). Heterogeneous preferences and the efficacy of public school choice. mimeo.

Immorlica, N. S., J. D. Leshno, I. Y. Lo, and B. J. Lucier (2020). Information acquisition in matching markets: The role of price discovery. mimeo.

Kojima, F., P. A. Pathak, and A. E. Roth (2013). Matching with couples: Stability and incentives in large markets. *The Quarterly Journal of Economics 128*(4), 1585–1632.

Larroucau, T. and I. Rios (2020a). Do “short-list" students report truthfully? strategic behavior in the chilean college admissions problem. mimeo.

Larroucau, T. and I. Rios (2020b). Dynamic college admissions. mimeo.

Leshno, J. D. (2022). Stable Matching with Peer-Dependent Preferences in Large Markets: Existence and Cutoff Characterization. mimeo.

Li, S. (2017). Obviously strategy-proof mechanisms. *American Economic Review 107*(11), 3257–87.

Luflade, M. (2019). The value of information in centralized school choice systems. mimeo.

Manny, A., H. Yam, and R. Lipka (2019). The usefulness of the atar as a measure of academic achievement and potential. *https://www.uac.edu.au/assets/documents/submissions/usefulness-of-the-atar-report.pdf*.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics 142*(2), 698–714.

Meisner, V. (2022). Report-dependent utility and strategy-proofness. *Management Science 69*(5), 2547–2155.

Meisner, V. and J. von Wangenheim (2023). Loss aversion in strategy-proof school-choice mechanisms. *Journal of Economic Theory 207*.

Moschovakis, Y. (2006). *Notes on Set Theory, Second Edition*. Springer.

Narita, Y. (2018). Match or mismatch? learning and inertia in school choice. mimeo.

Neilson, C. (2019). The rise of centralized choice and assignment mechanisms in education markets around the world. mimeo.

Nguyen, T. and R. Vohra (2018). Near-feasible stable matchings with couples. *American Economic Review 108*(11), 3154–69.

Pathak, P. A. and T. Sönmez (2008). Leveling the playing field: Sincere and sophisticated players in the boston mechanism. *American Economic Review 98*(4), 1636–1652.

Pop-Eleches, C. and M. Urquiola (2013). Going to a better school: Effects and behavioral responses. *American Economic Review 103*(4), 1289–1324.

Pycia, M. (2012). Stability and preference alignment in matching and coalition formation. *Econometrica 80*(1), 323–362.

Pycia, M. and M. B. Yenmez (2023). Matching with externalities. *The Review of Economic Studies 90*(2), 948–974.

Qiu, J. and R. Zhao (2007). *College admissions cutoffs and application guide: 2007-2008, Second Edition*. Science Press: Beijing.

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

Rose, E. K. and Y. Shem-Tov (2021). How does incarceration affect reoffending? estimating the dose-response function. *Journal of Political Economy 129*(12), 3302–3356.

Roth, A. E. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica 70*(4), 1341–1378.

Roth, A. E. and E. Peranson (1999). The redesign of the matching market for american physicians: Some engineering aspects of economic design. *American Economic Review 89*(4), 748–780.

Rothstein, J. and A. Yoon (2008). Mismatch in law school. NBER WP 14275.

Rothstein, J. M. (2006). Good principals or good peers? parental valuation of school characteristics, tiebout equilibrium, and the incentive effects of competition among jurisdictions. *American Economic Review 96*(4), 1333–1350.

Royden, H. (1988). *Real Analysis (Third Edition)*. Collier Macmillan.

Sasaki, H. and M. Toda (1996). Two-sided matching problems with externalities. *Journal of Economic Theory 70*(1), 93–108.

Scarf, H. (1960). Some examples of global instability of the competitive equilibrium. *International Economic Review 1*(3), 157–172.

Scrivener, S. and M. J. Weiss (2009). More guidance, better results? three-year effects of an enhanced student services program at two community colleges. *Three-Year Effects of an Enhanced Student Services Program at Two Community Colleges (August 1, 2009). New York: MDRC*.

Song, Y., K. Tomoeda, and X. Xia (2020). Sophistication and cautiousness in college applications. mimeo.

Sóvágó, S. and R. I. Shorrer (2018). Obvious mistakes in a strategically simple college-admissions environment. mimeo.

Tincani, M. M. (2018). Heterogeneous peer effects in the classroom. mimeo.

Weiss, M. J., A. Ratledge, C. Sommo, and H. Gupta (2019). Supporting community college students from start to degree completion: Long-term evidence from a randomized trial of cuny's asap. *American Economic Journal: Applied Economics 11*(3), 253–297.

APPENDIX

This document presents examples and proofs omitted in the main text, additional theoretical results, and additional empirical evidence.

A Proofs

Theorem 1

Before proving this result, we present the following condition which requires that the ordinal preferences of only a small measure of students change when the assignment changes slightly. Intuitively, it can be viewed as an ordinal version of **A4**.

A4' Peer preferences are *aggregate unresponsive*: for any $\epsilon > 0$ there exists some $\delta > 0$ such that if for any two assignments $\alpha, \alpha' \in \mathcal{A}$ we have that $\sup_{c,x} |\lambda^{c,x}(\alpha) - \lambda^{c,x}(\alpha')| := ||\lambda(\alpha) - \lambda(\alpha')||_\infty < \delta$, then $\eta(\{\theta \in \Theta | \succeq^{\theta|\alpha} \neq \succeq^{\theta|\alpha'}\}) < \epsilon$.

Lemma A.1. *A4' is satisfied in any market E satisfying A1- A4.*

Proof. Consider any market $E = [\eta, q, N, \Theta]$. Consider any ability distribution λ, which by **A2** is sufficient for the description of preferences. By **A1** almost all students have strict preferences induced by λ, that is, for any two programs c, c', $c \succeq^{\theta|\lambda} c'$ and $c' \succeq^{\theta|\lambda} c$ for almost no students. Fix any $\epsilon > 0$. By the uniform continuity of $f^{\theta,c}$ for all θ and all $c \in C \setminus \{c_0\}$ (**A4**), there exists some $\delta > 0$ such that for any ability distribution λ' with $||\lambda - \lambda'||_\infty < \delta$ we have that $\eta(\{\theta | \succeq^{\theta|\lambda} = \succeq^{\theta|\lambda'}\}) > 1 - \epsilon$. Then E satisfies **A4'**: $\eta(\{\theta | \succeq^{\theta|\lambda} \neq \succeq^{\theta|\lambda'}) < \epsilon$ for any λ' with $||\lambda - \lambda'||_\infty < \delta$. □

We now proceed with the proof of Theorem 1.

Proof of Theorem 1. By Lemma 2, it suffices to show the existence of a rational expectations, market clearing cutoff-distribution vector pair (p, λ). Define $Z(p, \lambda) = Z^d(p, \lambda) \times Z^\lambda(p, \lambda)$, with the first factor defined as a vector with entries for each $c \in C$ given by:

$$Z^{d,c}(p, \lambda) = \begin{cases} \frac{p^c}{1 + q^c - D^c(p,\lambda)} \text{ if } D^c(p, \lambda) \leq q^c \\ p^c + D^c(p, \lambda) - q^c \text{ if } D^c(p, \lambda) > q^c \end{cases} \tag{A.1}$$

and the second given by:

$$Z^\lambda(p, \lambda) = \lambda(A(p, \lambda)). \tag{A.2}$$

Z^λ is a mapping from $[0,1]^{N+1} \times \Lambda^{N+1}$ to Λ^{N+1}. So, Z is a mapping from $K := [0,1]^{N+1} \times \Lambda^{N+1} \to K$. We endow K with the metric induced by the sup norm; all notions of compactness and continuity will be relative to this metric. Our proof involves the following steps:

Step 1 If (p,λ) is a fixed point of Z, then (p,λ) satisfies rational expectations and is market clearing,

Step 2 K is a convex, compact, non-empty Hausdorff topological vector space, and

Step 3 Z is continuous.

Steps 2 and 3 imply by Schauder's fixed-point theorem that Z has a fixed point, which by Step 1 yields the desired result.

Proof of Step 1: To see that a fixed point (p,λ) of Z implies that (p,λ) satisfies rational expectations and are market clearing note that $Z^{\lambda}(p,\lambda)=\lambda$ implies that $\lambda=\lambda(A(p,\lambda))$. Therefore, (p,λ) satisfies rational expectations. $Z^{d}(p,\lambda)=p$ implies $D^{c}(p,\lambda)\leq q^{c}$ for all $c\in C$. Moreover, for any $c\in C$, if $D^{c}(p,\lambda)<q^{c}$ then it must be that $p^{c}=0$. Therefore, (p,λ) is market clearing. □

Proof of Step 2: Before proceeding to show the desired properties, we first offer a useful characterization of Λ. Let $\psi:[0,1]\to[0,1]$ be an absolutely continuous function such that $\psi(x)=\int_0^x\psi'(y)dy$ for all $x\in[0,1]$, where $\psi'(y)\in[0,1]$ for almost all $y\in[0,1]$. Let Ψ be the set of all such functions.

Lemma A.2. $\Psi=\Lambda$.

Proof. That $\Psi\subset\Lambda$ is established in Proposition 7 of Gentile Passaro et al. (2023). It remains to show that $\Lambda\subset\Psi$. Throughout the proof of this lemma, we forgo indexing α and λ terms by c to avoid unnecessary notation, as the arguments hold for any program c.

For any measurable subset (assignment) $\alpha\subset\Theta$, we define a measure $\bar{\eta}^{\alpha}$ over $[0,1]$ such that for any (Lebesgue) measurable set $A\subset[0,1]$, $\bar{\eta}^{\alpha}(A):=\eta(\{\theta\in\alpha|r^{\theta,c_0}\in A\})$. Two observations are in order. First, $\bar{\eta}^{\Theta}(A)=|A|$ because r^{θ,c_0} is uniformly distributed i.e. $\bar{\eta}^{\Theta}$ corresponds to the Lebesgue measure. Second, for any $\alpha\in\mathcal{A}$, $\bar{\eta}^{\alpha}$ is absolutely continuous by construction with respect to $\bar{\eta}^{\Theta}$. Absolute continuity holds because for any A such that $\bar{\eta}^{\Theta}(A)=0$, $0\leq\bar{\eta}^{\alpha}(A)=\eta(\{\theta\in\alpha|r^{\theta,c_0}\in A\})\leq\eta(\{\theta\in\Theta|r^{\theta,c_0}\in A\})=\bar{\eta}^{\Theta}(A)=0$, where the first inequality follows because $\bar{\eta}^{\alpha}$ is a measure and the second inequality follows because $\alpha\subset\Theta$.

Let $\alpha\in\mathcal{A}$, i.e. α is a measurable subset of Θ. Then $\lambda^{x}(\alpha)$ is Lipschitz continuous in x with constant 1. To see this, for any $x,x'\in[0,1]$ where $x'\geq x$ without loss of generality,

$$\begin{aligned}\lambda^{x'}(\alpha)-\lambda^{x}(\alpha)&=\eta^{\alpha}(\{\theta\in\alpha|r^{\theta,c_0}\leq x'\})-\eta^{\alpha}(\{\theta\in\alpha|r^{\theta,c_0}\leq x\})\\&=\eta^{\alpha}(\{\theta\in\alpha|r^{\theta,c_0}\in(x,x']\})\\&\leq\eta(\{\theta\in\Theta|r^{\theta,c_0}\in(x,x']\})\\&=x'-x,\end{aligned}\tag{A.3}$$

where the inequality follows because $\alpha\subset\Theta$ and the final equality follows from the assumption that r^{θ,c_0} is uniformly distributed over $[0,1]$. Lipschitz continuity implies that $\lambda^{x}(\alpha)$ is absolutely

continuous in x, which in turn implies that for almost all $x \in [0,1]$, $\frac{d\lambda^x(\alpha)}{dx} := \lambda'^x(\alpha)$ exists. Therefore, for any $x \in [0,1]$ we can write

$$\lambda^x(\alpha) = \lambda^0(\alpha) + \int_0^x \lambda'^y(\alpha) dy. \tag{A.4}$$

By construction, $\lambda^x(\alpha) = \bar{\eta}^\alpha([0,x])$ for all $x \in [0,1]$. Absolute continuity of $\lambda(\alpha)$ in x implies that the Radon-Nikodym derivative of measure $\bar{\eta}^\alpha$ is almost everywhere equal to $\lambda'^x(\alpha)$ (Royden, 1988, page 303). Furthermore, $\lambda'^y(d) \in [0,1]$ for almost all $y \in [0,1]$. To see this, note that

$$\begin{aligned}
\lambda'^y(\alpha) &= \lim_{\Delta y \to 0} \frac{\lambda^{y+\Delta y}(\alpha) - \lambda^y(\alpha)}{\Delta y} \\
&= \lim_{\Delta y \to 0} \frac{\bar{\eta}^\alpha([0, y+\Delta y]) - \bar{\eta}^\alpha([0,y])}{\Delta y} \\
&= \lim_{\Delta y \to 0} \frac{\eta(\{\theta \in \alpha | r^{\theta,c_0} \in (y, \Delta y]\})}{\Delta y}.
\end{aligned} \tag{A.5}$$

for almost all $y \in [0,1]$ by absolute continuity. The final line in Equation A.5 is weakly greater than 0 because η is a measure, which establishes that $\lambda'^y(\alpha) \geq 0$ for almost all y. Also, the final line in Equation A.5 is weakly smaller than $\lim_{\Delta y \to 0} \frac{\eta(\{\theta \in \Theta | r^{\theta,c_0} \in (y,\Delta y]\})}{\Delta y} = 1$, because $\alpha \subset \Theta$ and where the equality follows from the assumption that r^{θ,c_0} is uniformly distributed over $[0,1]$. Moreover, $\lambda^0(\alpha) = 0$ because $0 = \eta(\{\theta \in \Theta | r^{\theta,c_0} \leq 0\}) \geq \eta(\{\theta \in \alpha | r^{\theta,c_0} \leq 0\}) \geq 0$ where the equality by construction of Θ, the first inequality follows because $\alpha \subset \Theta$, and the final inequality follows because α is measurable. Therefore, for any $x \in [0,1]$ we can rewrite Equation A.4 as

$$\lambda^x(\alpha) = \int_0^x \lambda'^y(\alpha) dy,$$

where $\lambda'^y(\alpha) \in [0,1]$ for almost all $y \in [0,1]$. Therefore, there is some $\psi \in \Psi$ such that $\lambda(\alpha) = \psi$. □

We now return to showing K has the desired properties. It is clear that K is a Hausdorff topological vector space as it is a metric space (i.e. we endow it with the metric induced by the sup norm). We show that Λ is convex, compact, and non-empty, which demonstrates that K satisfies these properties as the product of convex, compact, and non-empty sets.

It is clear that Λ is nonempty. For example, $\alpha = \Theta$ corresponds to $\lambda^x(\alpha) = x$ for all $x \in [0,1]$.

Lemma A.3. *Λ is convex.*

Proof. Take any $\lambda_1, \lambda_2 \in \Lambda$ and any $\beta \in (0,1)$. We must show $\beta\lambda_1 + (1-\beta)\lambda_2 \in \Lambda$. For any $x \in [0,1]$,

$$\begin{aligned}
\beta\lambda_1^x + (1-\beta)\lambda_2^x &= \beta \int_0^x \lambda_1'^y dy + (1-\beta) \int_0^x \lambda_2'^y dy \\
&= \int_0^x [\beta\lambda_1'^y + (1-\beta)\lambda_2'^y] dy
\end{aligned}$$

where the first equality follows by Lemma A.2. Note also that $\beta\lambda_1'^y+(1-\beta)\lambda_2'^y\in[0,1]$ for almost all $y\in[0,1]$ because $\beta\in(0,1)$ and $\lambda_1'^y,\lambda_2'^y\in[0,1]$ for almost all $y\in[0,1]$. Therefore, by Lemma A.2, $\beta\lambda_1^x+(1-\beta)\lambda_2^x\in\Lambda$. □

Lemma A.4. *Λ is compact.*

Proof. Each $\lambda\in\Lambda$ is uniformly bounded ($\lambda^x(\alpha)\in[0,1]$ by construction for all $x\in[0,1]$ and all $\alpha\in\mathcal{A}$) and uniformly equicontinuous (which follows from the fact that each $\lambda\in\Lambda$ is Lipschitz continuous in $x\in[0,1]$ with constant 1). By the Arzelà-Ascoli Theorem, the closure of Λ is therefore compact. To show that Λ is compact, it remains only to show that Λ is closed.

To this end, consider a sequence of functions $(\lambda_\ell)_{\ell=1}^\infty$ with $\lambda_\ell\in\Lambda$ for all ℓ that converges to λ_* with respect to the sup norm, that is, for any $\epsilon>0$ there exists $L\geq 0$ such that $||\lambda_\ell-\lambda_*||_\infty<\epsilon$ for all $\ell>L$. We show the following properties:

$\lambda_*^0=0$. Suppose not, for the sake of contradiction. In particular, suppose $\lambda_*^0=\delta$ for some $\delta\neq 0$. For each ℓ, $\lambda_\ell^0=0$ because $\lambda_\ell\in\Lambda$ for all ℓ. Therefore, for $\epsilon\leq|\delta|$ and any ℓ, $||\lambda_\ell-\lambda_*||_\infty\geq|\lambda_\ell^0-\lambda_*^0|=|\delta|\geq\epsilon$. Contradiction with the assumption that $(\lambda_\ell)_{\ell=1}^\infty$ converges to λ_* with respect to the sup norm.

λ_* is non-decreasing. Suppose not, for the sake of contradiction. In particular, suppose that there exists $x,x'\in[0,1]$ with $x<x'$ such that $\lambda_*^x-\lambda_*^{x'}=\delta>0$. Let $\epsilon=\frac{\delta}{2}$. Then there exists $L\geq 0$ such that $|\lambda_\ell^x-\lambda_*^x|<\epsilon$ and $|\lambda_\ell^{x'}-\lambda_*^{x'}|<\epsilon$ for all $\ell>L$. Consider any $\ell'>L$. Then by the preceding argument, $\lambda_{\ell'}^x>\lambda_*^x-\epsilon$ and $\lambda_{\ell'}^{x'}<\lambda_*^{x'}+\epsilon$. Therefore, $\lambda_{\ell'}^x-\lambda_{\ell'}^{x'}>\lambda_*^x-\lambda_*^{x'}-2\epsilon=\delta-2\epsilon=0$, which implies that $\lambda_{\ell'}$ is not non-decreasing. Contradiction with $\lambda_{\ell'}\in\Lambda$. The non-decreasing property of λ_* establishes that $\lambda_*'^x$ exists for almost all $x\in[0,1]$ and is weakly positive for any x where it exists.

$\lambda_*^x\in[0,1]$ for all $x\in[0,1]$. The preceding two arguments imply that $\lambda_*^x\geq 0$ for all $x\in[0,1]$. It therefore remains to show that $\lambda_*^x\leq 1$ for all $x\in[0,1]$. By the non-decreasing property of λ^*, it suffices to show that $\lambda_*^1\leq 1$. Suppose for contradiction that this is not the case, in particular, suppose $\lambda_*^1=1+\delta$ for some $\delta\geq 0$. For each ℓ, $\lambda_\ell^1\leq 1$ because $\lambda_\ell\in\Lambda$ for all ℓ. Therefore, for $0<\epsilon\leq\delta$ and any ℓ, $||\lambda_\ell-\lambda_*||_\infty\geq|\lambda_\ell^1-\lambda_*^1|\geq 1+\delta-\lambda_\ell^1\geq\delta\geq\epsilon$. Contradiction with the assumption that $(\lambda_\ell)_{\ell=1}^\infty$ converges to λ_* with respect to the sup norm.

λ_* is Lipschitz continuous with constant 1. Recall that λ_* is non-decreasing by our earlier arguments. Suppose for the sake of contradiction that λ_* is not Lipschitz continuous with constant 1. In particular, suppose that for some $x,x'\in[0,1]$ with $x'>x$ it is the case that $\lambda_*^{x'}-\lambda_*^x=x'-x+\delta$ for some $\delta>0$. Let $\epsilon=\frac{\delta}{2}$. Then $|\lambda_\ell^x-\lambda_*^x|<\epsilon$ and $|\lambda_\ell^{x'}-\lambda_*^{x'}|<\epsilon$ for all $\ell>L$. Consider any $\ell'>L$. Then by the preceding argument, $\lambda_{\ell'}^x<\lambda_*^x+\epsilon$ and $\lambda_{\ell'}^{x'}>\lambda_*^{x'}-\epsilon$. Therefore, $\lambda_{\ell'}^{x'}-\lambda_{\ell'}^x>\lambda_*^{x'}-\lambda_*^x-2\epsilon=x'-x+\delta-2\epsilon=x'-x$, which implies that $\lambda_{\ell'}$ is not Lipschitz continuous with constant 1. Contradiction with $\lambda_{\ell'}\in\Lambda$.

Lipschitz continuity of λ_* implies that λ_* is absolutely continuous, i.e. $\lambda_*^x = \lambda_*^0 + \int_0^x \lambda_*'^y dy$. We have established that $\lambda_*^0 = 0$, $\lambda_*^x \in [0,1]$ for all $x \in [0,1]$, and that $\lambda_*'^x \in [0,1]$ for almost all $x \in [0,1]$. By Lemma A.2 it is therefore the case that $\lambda_* \in \Lambda$, establishing closedness, and therefore compactness, of Λ, as desired. □

□

Proof of Step 3: Consider any pairs $(p,\lambda) \in [0,1]^{N+1} \times \Lambda^{N+1}$ and $(p',\lambda') \in [0,1]^{N+1} \times \Lambda^{N+1}$ where we write $\alpha = A(p,\lambda)$ and $\alpha' = A(p',\lambda')$. We must show that for any $\epsilon > 0$ there exists $\delta > 0$ such that if $||(p,\lambda) - (p',\lambda')||_\infty < \delta$ then $||Z(p,\lambda) - Z(p',\lambda')||_\infty < \epsilon$. Note that by construction, $Z^{d,c}(p,\lambda)$ is continuous in $D^c(p,\lambda)$ for all $c \in C$ (this follows from Equation A.1 and noting that $D^c(\cdot,\cdot) \leq 1 < 1 + q^c$). Also, $Z^\lambda(p,\lambda) = \lambda(\alpha)$ and $Z^\lambda(p',\lambda') = \lambda(\alpha')$ by Equation A.2. Therefore, it suffices to show that for any $\epsilon > 0$ there exists $\delta > 0$ such that if $||(p,\lambda) - (p',\lambda')||_\infty < \delta$ then both $|D^c(p,\lambda) - D^c(p',\lambda')| < \epsilon$ for all $c \in C$ and $||\lambda(\alpha) - \lambda(\alpha')||_\infty < \epsilon$.

By Assumption **A1**, for almost all $\theta \in \Theta$ we have that $\alpha(\theta) \neq \alpha'(\theta)$ if and only if $D^\theta(p,\lambda) \neq D^\theta(p',\lambda')$. Denote the set of students for whom $D^\theta(p,\lambda) \neq D^\theta(p',\lambda')$ as $\Theta(\alpha,\alpha') := \{\theta | D^\theta(p,\lambda) \neq D^\theta(p',\lambda')\}$. We first argue that for sufficiently small δ, $\eta(\Theta(\alpha,\alpha')) < \epsilon$. Note that $\theta \in \Theta(\alpha,\alpha')$ only if either $\{c | p^c \leq r^{\theta,c}\} \neq \{c | p'^c \leq r^{\theta,c}\}$ (different choice sets), or $\succeq^{\theta|\lambda} \neq \succeq^{\theta|\lambda'}$ (different ordinal rankings), or both.

Let the students with different choice sets be denoted $\Theta^1(\alpha,\alpha') := \{\theta | \{c | p^c \leq r^{\theta,c}\} \neq \{c | p'^c \leq r^{\theta,c}\}\}$, and the students with different ordinal preferences $\Theta^2(\alpha,\alpha') := \{\theta | \succeq^{\theta|\lambda} \neq \succeq^{\theta|\lambda'}\}$.

For any $\delta < 1$, when $||(p,\lambda) - (p',\lambda')||_\infty < \delta$ the measure of students with different choice sets $\eta(\Theta^1(\alpha,\alpha')) < (N+1)\delta$ by construction. This is due to the fact that $|p^c - p'^c| < \delta$ for all programs $c \in C$ and the ongoing assumption of a uniform distribution of student scores within program. Let $\epsilon' = \frac{\epsilon}{N+2}$. By **A4'**, there exists $\delta^1 > 0$ such that when $||(p,\lambda) - (p',\lambda')||_\infty < \delta^1$ the measure of students with different ordinal rankings $\eta(\Theta^2(\alpha,\alpha')) < \epsilon'$. Let $\delta = \min\{\frac{\epsilon}{N+2}, \delta^1\}$. Therefore, if $||(p,\lambda) - (p',\lambda')||_\infty < \delta$ it must be the case that

$$\begin{aligned}
\eta(\Theta(\alpha,\alpha')) &\leq \eta(\Theta^1(\alpha,\alpha') \cup \Theta^2(\alpha,\alpha')) \\
&\leq \eta(\Theta^1(\alpha,\alpha')) + \eta(\Theta^2(\alpha,\alpha')) \\
&< (N+1)\delta + \epsilon' \\
&\leq (N+1)\frac{\epsilon}{N+2} + \frac{\epsilon}{N+2} \\
&= \epsilon
\end{aligned} \tag{A.6}$$

where the first inequality holds because $\theta \in \Theta(\alpha,\alpha')$ only if θ is an element of at least one of

$\Theta^1(\alpha,\alpha')$ and $\Theta^2(\alpha,\alpha')$. Therefore, the proof is complete if we can show that

$$\eta(\Theta(\alpha,\alpha')) \geq |D^c(p,\lambda) - D^c(p',\lambda')| \text{ for all } c \in C \tag{A.7}$$

and

$$\eta(\Theta(\alpha,\alpha')) \geq ||\lambda(\alpha) - \lambda(\alpha')||_\infty. \tag{A.8}$$

To see that Inequality A.7 holds, note that for any $c \in C$ we have that

$$\begin{aligned}
\eta(\Theta(\alpha,\alpha')) &= \frac{1}{2}\sum_{\tilde{c}\in C}[\eta(\alpha(\tilde{c})\backslash\alpha'(\tilde{c})) + \eta(\alpha'(\tilde{c})\backslash\alpha(\tilde{c})] \\
&\geq \eta(\alpha(c)\backslash\alpha'(c)) + \eta(\alpha'(c)\backslash\alpha(c)) \\
&= \eta(\alpha(c)) + \eta(\alpha'(c)) - 2\eta(\alpha(c)\cap\alpha'(c)) \\
&= \max\{\eta(\alpha(c)),\eta(\alpha'(c))\} + \min\{\eta(\alpha(c)),\eta(\alpha'(c))\} - 2\eta(\alpha(c)\cap\alpha'(c)) \\
&\geq \max\{\eta(\alpha(c)),\eta(\alpha'(c))\} - \min\{\eta(\alpha(c)),\eta(\alpha'(c))\} \\
&= |\eta(\alpha(c)) - \eta(\alpha'(c))| \\
&= |D^c(p,\lambda) - D^c(p',\lambda')|
\end{aligned} \tag{A.9}$$

The first equality follows because each student $\theta \in \Theta(\alpha,\alpha')$ is double counted in the right-hand side of the top line.[1] The first inequality follows because the total measure of students with different assignments with respect to α and α' is weakly greater than the measure of students who are assigned to program c in exactly one of the two assignments. The second inequality follows because $\min\{\eta(\alpha(c)),\eta(\alpha'(c))\} \geq \eta(\alpha(c)\cap\alpha'(c))$.

To see that Inequality A.8 holds, note that for any $c \in C$ and any $x \in [0,1]^{N+1}$,

$$\eta(\Theta(\alpha,\alpha')) \geq |\eta(\alpha(c)) - \eta(\alpha'(c))| \geq |\lambda^{c,x}(\alpha) - \lambda^{c,x}(\alpha')|$$

where the first inequality follows from Inequalities A.9 and the second inequality follows because the difference in the measure of students with scores below x assigned to c at α and α' cannot be larger than the total measure of students who are assigned to c in only one of α and α'. □

The completion of the proofs of the three steps completes the proof of the theorem. □

Proposition 1

Proof of Part 1: Let μ_* be a stable matching. As we argue in Remark 3, letting $\succ$ represent a profile of ROLs such that $\succ^\theta = \succeq^{\theta|\mu_*}$ for all θ, $\varphi(\succ) = \mu_*$ for any stable mechanism φ.[2] For each θ, let $\tilde{\succ}^\theta$ be the submitted preferences for θ such that $\mu_*(\theta)$ is the unique acceptable program, and let $\tilde{\succ}$ be the profile of such reports for all $\theta \in \Theta$. Because φ is stable, $\varphi(\tilde{\succ}) = \varphi(\succ) = \mu_*$. To see

[1] That is, if $\theta \in \alpha(c_1)\cap\alpha'(c_2)$ then θ contributes to the sum on the right-hand side for both c_1 and c_2.

[2] Remark 3 follows straightforwardly from **A3** and (Grigoryan, 2022).

that this is a Bayes Nash equilibrium, note that for any θ and any program $c \succ^{\theta|\mu_*} \mu_*(\theta)$, stability of μ_* implies that there is no deviating report $\succ^{\theta} \neq \tilde{\succeq}^{\theta}$ that will result in θ matching with c.

Suppose for contradiction that $\tilde{\succeq}$ is a Bayes Nash equilibrium of φ but that $\mu = \varphi(\tilde{\succeq})$ is not a stable matching. Then there exists some $\theta \in \Theta$ and some $c \in C$ such that (θ, c) form a blocking pair (with respect to $\succeq^{\theta|\mu}$). By Remark 3 and the fact that φ is a stable mechanism, μ is the unique stable matching with respect to $\tilde{\succeq}$. Let p be the associated cutoff vector. Now consider reported preferences $\hat{\succeq}$ where $\hat{\succeq}^{\theta'} = \tilde{\succeq}^{\theta'}$ for all $\theta' \neq \theta$ and $\hat{\succeq}^{\theta}$ lists only program c as acceptable. There is similarly a unique stable matching μ' with respect to these preferences, but the cutoff vector for this stable matching must also be p, due to the reported preferences of a zero measure set of students differing between $\hat{\succeq}$ and $\tilde{\succeq}$. Since (θ, c) block μ it must be that $r^{\theta,c} \geq p^c$. But then $\varphi^{\theta}(\hat{\succeq}) = c$ since c is a stable mechanism. Contradiction with $\tilde{\succeq}$ being a Bayes Nash equilibrium. □

Proof of Part 2: Let $\tilde{\succeq}$ be a Bayes Nash equilibrium, and suppose for contradiction that $\varphi(\tilde{\succeq}) = \mu_*$. By the assumption that μ_* is stable, it must be that μ_* is associated with some cutoff vector p satisfying $p \leq \bar{p}$ (Lemma 2).

Consider any student θ such that $r^{\theta} \geq \tilde{p}$. By Assumption **A1**, $\succeq^{\theta|\mu_*}$ is strict for almost all such θ, and we proceed assuming $\succeq^{\theta|\mu_*}$ is strict. By the stability of μ_* and the fact that θ's score $r^{\theta,c}$ at each program c exceeds c's cutoff, it must be the case that $\mu_*(\theta)$ is the $\succeq^{\theta|\mu_*}$-maximal program.

Moreover, it follows from Assumption **A3** that for each program $c \in C$ there exists a set Θ^c of positive measure such that for each $\theta^c \in \Theta^c$: $\bar{p}^c < r^{\theta^c,c} < \tilde{p}^{\theta,c}$ and c is the unique $\succeq^{\theta^c|\mu_*}$-maximal program. By the stability hypothesis, $\mu_*(\theta^c) = c$ for each $\theta^c \in \Theta^c$. Because φ respects rankings, this implies that θ, who recall satisfies $r^{\theta} \geq \tilde{p}$, is admitted to her top-ranked program according to her submitted preferences $\tilde{\succeq}^{\theta}$. Therefore, stability implies that $\mu_*(\theta)$ is θ's top-ranked program according to $\tilde{\succeq}^{\theta}$. By the equilibrium hypothesis, it must be that the $\tilde{\succeq}^{\theta}$-maximal program is the same as the $\succeq^{\theta|\mu(\sigma^{\theta},\tilde{\succeq})}$-maximal program. That is, it must be that, for equilibrium profile $\tilde{\succeq}$, student θ realizes that she will receive her top-ranked program, and therefore, her top-ranked program must coincide with the top-ranked program according to her true preferences (given her beliefs over the distribution of types).

The logic of the previous two paragraphs implies that the top-ranked program according to $\succeq^{\theta|\mu_*}$ coincides with the top-ranked program according to $\succeq^{\theta|\mu(\sigma^{\theta},\tilde{\succeq})}$ for almost all θ with $r^{\theta} \geq \tilde{p}$. But this contradicts the ongoing assumption that $\eta(L_{\tilde{\succeq},\varphi,\tilde{p}}) > 0$. □

Before proceeding, we provide a lemma which is useful in several upcoming proofs.

Lemma A.5. *For any $\lambda,\ \lambda' \in \Lambda^{N+1}$, define $p, p' \in [0,1]^{N+1}$ to be the unique respective cutoff vectors such that (p,λ) and (p',λ') are market clearing. Let $\mu = A(p,\lambda)$, and $\mu' = A(p',\lambda')$. For*

any $\epsilon > 0$ *there exists* $\delta > 0$ *such that if* $|D^c(p,\lambda) - D^c(p,\lambda')| < \delta$ *for all* $c \in C$ *then* $||p - p'||_\infty < \epsilon$ *and* $||\lambda(\mu) - \lambda(\mu')||_\infty < \epsilon$.

Proof. Fix $\epsilon > 0$, and let $\omega > 0$ define the bound on the support of student types in Assumption **A3**. We first argue that there exists $\delta > 0$ such that $||p - p'||_\infty < \epsilon$ when $|D^c(p,\lambda) - D^c(p,\lambda')| < \delta$ for all $c \in C$. If $p = p'$ then we are done. In the complementary case, assume without loss of generality that $p^c > p'^c$ for some $c \in C$.

Let $\delta = \epsilon\omega$. Then for such λ, λ',

$$\epsilon\omega > D^c(p,\lambda) - D^c(p,\lambda') = q^c - D^c(p,\lambda') \geq 0 \tag{A.10}$$

where the equality follows because the assumption that $p^c > p'^c$ implies that $p^c > 0$ which therefore implies that $D^c(p,\lambda) = \eta(\mu(c)) = q^c$.

In order to respect c's capacity constraint, Inequality A.10 implies that there is at most a $\epsilon\omega$ measure of students matched to c in μ' with scores below p^c, $\eta\{\theta \in \mu'(c) | r^{\theta,c} < p^c\} \leq \epsilon\omega$. By bound ω from Assumption **A3**, it must be that $p'^c \in (p^c - \epsilon, p^c)$. Applying this argument across all programs $c \in C$ implies that $||p - p'||_\infty < \epsilon$.

That $||\lambda(\mu) - \lambda(\mu')||_\infty < \epsilon$ follows from the above argument and Inequality A.8. □

Proposition 2

Proof of Part 1:

"If" part Suppose $\lambda_t = \lambda_{t-1}$. Then $\lambda_{t-1} = \lambda_t = \lambda(A(p_t, \lambda_{t-1}))$, that is, (p_t, λ_{t-1}) satisfies rational expectations. By construction, (p_t, λ_{t-1}) is market clearing. Therefore, by Lemma 2, $\mu_t = A(p_t, \lambda_{t-1})$ is stable.

"Only if" part If μ_t is stable then $(\hat{p}, \lambda_t)$ satisfies rational expectations and is market clearing by Lemma 2, where $\hat{p}^c := \inf\{r^{\theta,c} | \theta \in \mu_t(c)\}$ for each $c \in C$. Therefore, $\mu_t = A(\hat{p}, \lambda_t)$ by Lemma 2. Because $\mu_t = A(p_t, \lambda_{t-1})$ by construction, it must therefore be that $p_t = \hat{p}$, for otherwise Assumption **A3** would imply $\mu_t = A(p_t, \lambda_{t-1}) \neq A(\hat{p}, \lambda_t) = \mu_t$, which is a contradiction. As previously argued, Remark 3 implies there is a unique $p \in [0,1]^{N+1}$ such that (p, λ_t) is market clearing. Since $\mu_{t+1} = A(p_{t+1}, \lambda_t)$ is market clearing by construction, it must be that $p_t = p_{t+1}$. Then $\lambda_{t+1} = \lambda(\mu_{t+1}) = \lambda(A(p_{t+1}, \lambda_t)) = \lambda(A(p_t, \lambda_t)) = \lambda(A(\hat{p}, \lambda_t)) = \lambda(\mu_t) = \lambda_t$, where the third equality follows from $p_{t+1} = p_t$ and the fourth equality follows from $p_t = \hat{p}$.

□

Proof of Part 2:

"If" part Fix any $\epsilon > 0$. We want to show that there exists $\delta > 0$ such that if $||\lambda_t - \lambda_{t-1}||_\infty < \delta$ then μ_t is $\epsilon-$stable. Let B denote the set of students who block μ_t, that is

$B := \{\theta | (\theta, c)$ block μ_t for some $c \in C\}$. Let $B^{\lambda_t,\lambda_{t-1}} := \{\theta | D^\theta(p_t,\lambda_t) \neq D^\theta(p_t,\lambda_{t-1})\}$. The following result states that almost surely $\theta \in B$ if and only if $\theta \in B^{\lambda_t,\lambda_{t-1}}$.

Lemma A.6. $\eta(\{B \setminus B^{\lambda_t,\lambda_{t-1}}\} \cup \{B^{\lambda_t,\lambda_{t-1}} \setminus B\}) = 0$.

Proof. We prove this result by showing that $\eta(B \setminus B^{\lambda_t,\lambda_{t-1}}) = 0$ and $\eta(B^{\lambda_t,\lambda_{t-1}} \setminus B) = 0$. This implies that $\eta(\{B \setminus B^{\lambda_t,\lambda_{t-1}}\} \cup \{B^{\lambda_t,\lambda_{t-1}} \setminus B\}) \leq \eta(B \setminus B^{\lambda_t,\lambda_{t-1}}) + \eta(B^{\lambda_t,\lambda_{t-1}} \setminus B) = 0$. For each $\theta \in B$ there exists some $c^\theta \in C$ such that (θ, c^θ) block μ_t. By construction, the cutoff vector p_t satisfies $r^{\theta,c^\theta} \geq p_t^{c^\theta}$ and $c^\theta \succ^{\theta|\mu_t} \mu_t(\theta)$, which implies that $D^\theta(p_t,\lambda_t) \neq D^\theta(p_t,\lambda_{t-1})$. Therefore, $\eta(B \setminus B^{\lambda_t,\lambda_{t-1}}) = 0$. By Assumption **A1**, for almost all $\theta \in B^{\lambda_t,\lambda_{t-1}}$ there exists a unique $c^\theta = D^\theta(p_t,\lambda_t)$. If $c^\theta \neq D^\theta(p_t,\lambda_{t-1})$ then (θ, c^θ) form a blocking pair at μ_t for almost all $\theta \in B^{\lambda_t,\lambda_{t-1}}$. Therefore, $\eta(B^{\lambda_t,\lambda_{t-1}} \setminus B) = 0$. □

Returning to the proof of the proposition, by **A4'** there exists $\delta > 0$ such that if $\|\lambda_{t-1} - \lambda_t\|_\infty < \delta$, then $\eta(\{\theta | \succeq^{\theta|\mu_{t-1}} \neq \succeq^{\theta|\mu_t}\}) < \epsilon$. For almost all $\theta \in B^{\lambda_t,\lambda_{t-1}}$ it is the case that $\succeq^{\theta|\mu_t - 1} \neq \succeq^{\theta|\mu_t}$, i.e. some subset of students whose ordinal rankings change demand a different program given p_t. Therefore, if $\|\lambda_{t-1} - \lambda_t\|_\infty < \delta$,

$$\eta(B) = \eta(B^{\lambda_t,\lambda_{t-1}}) \leq \eta(\{\theta | \succeq^{\theta|\mu_{t-1}} \neq \succeq^{\theta|\mu_t}\}) < \epsilon$$

where the equality follows from Lemma A.6. Thus, for $\|\lambda_{t-1} - \lambda_t\|_\infty < \delta$, $\eta(B) < \epsilon$ as desired.

"Only if" part Fix any $\delta > 0$ and let B continue to represent the set of students involved in at least one blocking pair at μ_t. We must show that there exists $\epsilon > 0$ such that if $\eta(B) < \epsilon$ then $\|\lambda_t - \lambda_{t+1}\|_\infty < \delta$, and this follows directly from Lemmas A.5 and A.6. □

Example 1 (continued)**.** *We now find conditions under which the market has a unique stable matching. Assume, subject to later verification, that there exists a stable matching $\mu_* = A(p_*,\lambda_*)$ in which c_1 fills all of its seats. Let $s_*^{c_1} = s^{c_1}(\lambda_*)$. As all students θ with $r^{\theta,c_1} \geq s_*^{c_1}$ will attend c_1, $1 - s_*^{c_1}$ measure of seats are occupied by students who face no peer costs. In order for $p_*^{c_1}$ to satisfy market clearing, it must be that $(s_*^{c_1} - p_*^{c_1})(1-k) = q^{c_1} - (1 - s_*^{c_1})$. As s^{c_1} is a function of λ, a necessary condition for rational expectations of $(p_*^{c_1},\lambda_*)$ is that $\frac{1+s_*^{c_1}}{2}(1 - s_*^{c_1}) + \frac{p_*^{c_1}+s_*^{c_1}}{2}(q^{c_1} - (1 - s_*^{c_1})) = s_*^{c_1}$. Solving these equations yields:*

$$p_*^{c_1} = \frac{1 - q^{c_1} - k s_*^{c_1}}{1-k}, \quad s_*^{c_1} = \frac{k - kq^{c_1} - 2 \pm \sqrt{4 + k^2(q^{c_1}-1)^2 - 4k((q^{c_1})^2 - 3q^{c_1} + 1)}}{2k}.$$

Noting that $k - kq^{c_1} - 2 < 0$, only the "plus" solution is viable. In order for the "plus" solution to satisfy the necessary condition, it must be that $(k - kq^{c_1} - 2)^2 \leq 4 + k^2(q^{c_1}-1)^2 - 4k((q^{c_1})^2 - 3q^{c_1} + 1)$, which is shown, following a standard calculation, to hold with a strict inequality whenever $q^{c_1} < 1$.

The above demonstrates that there is at most one stable matching in which c_1 fills all of its seats. We argue that when q^{c_1} is sufficiently small any *stable matching must involve c_1 filling all of its seats by showing that for sufficiently small q^{c_1}, it must be that $p_*^{c_1} > 0$. To see this, note that all students θ with $r^{\theta,c_1} > s_*^{c_1}$ will enroll in c_1. Therefore, $s_*^{c_1} > 1 - q^{c_1}$. For any fixed $k < 1$, $s_*^{c_1} \to 1$ as $q^{c_1} \to 0$. This implies that as $q^{c_1} \to 0$, $p_*^{c_1} = 0$ implies that $\eta(\mu_*(c_1)) \to 1-k$, which violates the definition of matching as too large a measure of students is assigned to c_1.*

There exists at least one stable matching by Theorem 1,[3] and our above arguments pin down the corresponding cutoffs $p_*^{c_1}$ and average abilities $s_*^{c_1} = s^{c_1}(\lambda(\mu_*))$ that must be identical in any two stable matchings for sufficiently small q^{c_1}. But if there exist two stable matchings, μ_* and $\mu\prime$, note that by our assumption that student preferences depend on $s^{c_1}(\lambda)$, $\succeq^{\theta|\mu_*} = \succeq^{\theta|\mu\prime}$ for all $\theta \in \Theta$. By Remark 3 it must be that $\mu_*(\theta) = \mu\prime(\theta)$ for all $\theta \in \Theta$. Therefore, there is a unique stable matching for sufficiently small q^{c_1}.

Theorem 2

In the main body, we informally described Theorem 2. Here, we formalize the result. We say that market $E = [\eta, q, N, \Theta]$ admits a *negative externality group* if there exists a $c \in C \setminus \{c_0\}$, a measurable set $\alpha(c)$, and measurable sets $\Theta^I \subset \alpha(c)$ and $\Theta^O \subset \Theta \setminus \alpha(c)$ with $\eta(\Theta^I) > \eta(\Theta^O)$ such that $f^{\theta,c}(\lambda^c(\Theta^O \cup \alpha(c) \setminus \Theta^I)) > f^{\theta,c}(\lambda^c(\alpha(c)))$ for all $\theta \in \Theta^I$. In words, a negative externality group at α requires a set of students Θ^I to prefer a program c when a (possibly empty) smaller set of students Θ^O replaces them.

The existence of a negative externality group depends on peer preference functions $f^\theta = (f^{\theta,c_1}, ..., f^{\theta,c_N})$ and scores r^{θ,c_0} (which together define peer preferences). Therefore, we develop additional notation to describe markets with "similar" peer preferences. Let $E = [\eta, q, N, \Theta]$. We say that $\mathbf{f} \mapsto \Theta$ when $f \in \mathbf{f}$ if and only if there is some $(\theta, c) \in \Theta \times C \setminus \{c_0\}$ such that $f^{\theta,c} = f$. Let $g^c(\cdot) : \Lambda \to [a', b']$ be a function mapping ability distributions into an interval of the real numbers $[a', b']$ for each $c \in C \setminus \{c_0\}$, and let $\mathbf{g} = (g^{c_1}, ..., g^{c_N})$ be a collection of such functions. If $E = [\eta, q, N, \Theta]$ is such that $\mathbf{f} \mapsto \Theta$ and $\tilde{E} = [\tilde{\eta}, q, N, \tilde{\Theta}]$ is such that for all $\tilde{\theta} \in \tilde{\Theta}$ there exists $\theta \in \Theta$ such that $r^{\tilde{\theta},c_0} = r^{\theta,c_0}$ and $f^{\tilde{\theta},c} = f^{\theta,c} + g^c$ for all $c \in C \setminus \{c_0\}$, and for every $\theta \in \Theta$ such that there exists $\tilde{\theta} \in \tilde{\Theta}$ for which $r^{\tilde{\theta},c_0} = r^{\theta,c_0}$ or $f^{\tilde{\theta},c} = f^{\theta,c} + g^c$ for all $c \in C \setminus \{c_0\}$ then we write $\mathbf{f} + \mathbf{g} \mapsto \tilde{\Theta}$. We define norm $||\cdot||_{\mathbf{f}}$ such that for any $E = [\eta, \cdot, N, \Theta]$ and $\tilde{E} = [\tilde{\eta}, \cdot, N, \tilde{\Theta}]$, $||E - \tilde{E}||_{\mathbf{f}} = \epsilon$ if

1. there exist collections of functions $\mathbf{f}$ and $\mathbf{g} = (g^{c_1}, ..., g^{c_N})$ such that $\mathbf{f} \mapsto \Theta$, $\mathbf{f} + \mathbf{g} \mapsto \tilde{\Theta}$, and $\sup_{c,\lambda} |g^c(\lambda)| = \epsilon$, and

[3] As our proof of Theorem 1 shows, existence of a stable matching arises when condition **A4** is replaced with **A4'**, a weaker condition that the current example satisfies.

2. for any $R \subset [0,1]$ and $\mathbf{f}' \subset \mathbf{f}$, let $\alpha^{R,\mathbf{f}'} := \{\theta \in \Theta | (r^{\theta,c_0}, f^\theta) \in R \times \mathbf{f}'^N\}$ and $\alpha^{R,\mathbf{f}'+\mathbf{g}} := \{\tilde{\theta} \in \tilde{\Theta} | (r^{\tilde{\theta},c_0}, f^{\tilde{\theta}} - (g^{c_1},...,g^{c_N})) \in R \times \mathbf{f}'^N\}$. Then any such $\alpha^{R,\mathbf{f}'}$ is η measurable if and only if $\alpha^{R,\mathbf{f}'+\mathbf{g}}$ is $\tilde{\eta}$ measurable, and for all measurable sets $\eta(\alpha^{R,\mathbf{f}'}) = \tilde{\eta}(\alpha^{R,\mathbf{f}'+\mathbf{g}})$.

In words, two markets are within ϵ of one another with respect to the $||\cdot||_{\mathbf{f}}$ norm if (1) the peer preferences of each student differs by at most ϵ in the two markets and (2) the set of students with particular abilities and peer preferences (net of perturbations $\mathbf{g}$) has the same measure in both markets.

We also formalize what it means for two markets to have the same "program-side" structure, that is, the set of programs, capacities, and distribution of scores at each of the programs is identical. We say that two markets $E = [\eta,q,N,\Theta]$ and $\hat{E} = [\hat{\eta},q,N,\hat{\Theta}]$ are *program-side identical* if for any collection of open intervals $\{(a_n,b_n)\}_{n\in\{1,2,...,N\}}$ it is the case that:

$$\eta\{\theta \in \Theta | (r^{\theta,c_1},...,r^{\theta,c_N}) \in (a_1,b_1) \times ... \times (a_N,b_N)\} = \hat{\eta}\{\hat{\theta} \in \hat{\Theta} | (r^{\hat{\theta},c_1},...,r^{\hat{\theta},c_N}) \in (a_1,b_1) \times ... \times (a_N,b_N)\}.$$

Theorem 2 (Formal)**.** *Let* $N \geq 1$.

1. *The set of markets that admit a negative externality group is open and dense in the set of all markets with respect to the* $||\cdot||_{\mathbf{f}}$ *norm.*
2. *Suppose* $E = [\eta,\cdot,N,\Theta]$ *admits a negative externality group. Then there exists a market* $\tilde{E} = [\tilde{\eta},q,N,\tilde{\Theta}]$ *such that* $||E - \tilde{E}||_{\mathbf{f}} = 0$ *and a starting condition* μ_0 *such that the TIM process does not converge in market* $\tilde{E}$.
3. *Let* $\hat{E} = [\hat{\eta},q,N,\hat{\Theta}]$ *be an arbitrary market. There exists a program-side identical market* $E = [\eta,q,N,\Theta]$ *in which there is a unique score distribution* λ_* *such that the TIM process converges if and only if the initial assignment* μ_0 *satisfies* $\lambda(\mu_0) = \lambda_*$.

Proof of Part 1:

Openness: We first argue that the set of markets that admit a negative externality group is open with respect to the $||\cdot||_{\mathbf{f}}$ norm. To do so, consider a market $E = [\eta,\cdot,N,\Theta]$ where $\mathbf{f} \mapsto \Theta$ that admits a negative externality group at program $c' \in C \setminus \{c_0\}$ and assignment $\alpha(c')$, $\Theta^I \subset \alpha(c')$ and $\Theta^O \subset \Theta \setminus \alpha(c')$. Therefore, $\eta(\Theta^I) > \eta(\Theta^O)$. We wish to show that there exists $\epsilon > 0$ such that if $\tilde{E} = [\tilde{\eta},\cdot,N,\tilde{\Theta}]$ such that $\mathbf{f} + \mathbf{g} \mapsto \Theta$ satisfies $||E - \tilde{E}||_{\mathbf{f}} < \epsilon$, then $\tilde{E}$ admits a negative externality group. By uniform continuity (see **A4**) for any sufficiently small $\delta > 0$ there exists a set $\Theta^i \subset \Theta^I$ with $\eta(\Theta^i) > \eta(\Theta^O)$ such that $f^{\theta,c'}(\lambda^{c'}(\Theta^O \cup \alpha(c') \setminus \Theta^i)) - f^{\theta,c'}(\lambda^{c'}(\alpha(c'))) > \delta$ for all $\theta \in \tilde{\Theta}^I$. Fix any such δ and consider and $\tilde{E}$ such that $||E - \tilde{E}||_{\mathbf{f}} < \frac{\delta}{4}$, that is, let $\epsilon = \frac{\delta}{4}$.

We now construct sets of students in market $\tilde{E}$ and then argue that they form a negative externality group. To do so, begin by letting $L \in \mathbb{N}$ and for each $\ell \in \{1,...,L\}$ let $\alpha(c')_{\ell,L} = \{\theta \in \alpha(c') | r^{\theta,c_0} \in [\frac{\ell-1}{L}, \frac{\ell}{L})\}$. Let $\mathbf{f}_{\ell,L}^{\alpha(c')}$ be the set of peer preference functions for all pairs (θ,c) such that $\theta \in \alpha(c')_{\ell,L}$. For each $\ell \in \{1,...,L\}$ let $\tilde{\alpha}(c')_{\ell,L} \subset \{\tilde{\theta} \in \tilde{\Theta} | f^{\tilde{\theta},c'} - g^{c'} \in \mathbf{f}_{\ell,L}^{\alpha(c')}$ and $r^{\tilde{\theta},c_0} \in [\frac{\ell-1}{L}, \frac{\ell}{L})\}$

subject to $\eta(\alpha(c')_{\ell,L}) = \tilde{\eta}(\tilde{\alpha}(c')_{\ell,L})$ for all $\ell \in \{1,...,L\}$. Note that by construction of market $\tilde{E}$ (in particular, $\tilde{\Theta}$ and $\tilde{\eta}$) such sets $\tilde{\alpha}(c')_{\ell,L}$ exist. Consider set $\tilde{\alpha}(c')_L := \bigcup_{\ell=1}^{L} \tilde{\alpha}(c')_{\ell,L}$, which represents a set of students with similar peer preferences and abilities as those in $\alpha(c')$. It is the case that

$$\tilde{\eta}(\tilde{\alpha}(c')_L) = \tilde{\eta}(\bigcup_{\ell=1}^{L} \tilde{\alpha}(c')_{\ell,L}) = \sum_{\ell=1}^{L} \tilde{\eta}(\tilde{\alpha}(c')_{\ell,L}) = \sum_{\ell=1}^{L} \eta(\alpha(c')_{\ell,L}) = \eta(\bigcup_{\ell=1}^{L} \alpha(c')_{\ell,L}) = \eta(\alpha(c')), \quad \text{(A.11)}$$

where the second and fourth equalities follow from the countable additivity of measures. Similarly construct sets $\tilde{\Theta}^i_L \subset \tilde{\Theta}^I_L \subset \tilde{\alpha}(c')_L$ and $\tilde{\Theta}^O_L \subset \tilde{\Theta} \setminus \tilde{\alpha}(c')_L$. Note that by the logic of Equation A.11, $\tilde{\eta}(\tilde{\Theta}^i_L) > \tilde{\eta}(\tilde{\Theta}^O_L)$.

Let $\tilde{\mathcal{A}}$ be the set of all assignments in market $\tilde{E}$. For each $x \in [0,1]$, $c \in C$, and $\tilde{\beta} \in \tilde{\mathcal{A}}$, let $\tilde{\lambda}^{c,x}(\tilde{\beta}) := \tilde{\eta}(\{\tilde{\theta} \in \tilde{\beta}(c) | r^{\theta,c_0} \leq x\})$, and let $\tilde{\lambda}^c(\beta)$ be the resulting non-decreasing function from $[0,1]$ to $[0,1]$. We claim that for large L, $f^{\tilde{\theta},c'}(\tilde{\lambda}^{c'}_L(\tilde{\Theta}^O_L \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}^i_L)) + g^{c'}(\tilde{\lambda}^{c'}_L(\tilde{\Theta}^O_L \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}^i_L)) > f^{\tilde{\theta},c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L)) + g^{c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L))$ for all $\tilde{\theta} \in \tilde{\Theta}^i_L$, which completes the construction of a negative externality group since we have already argued that $\tilde{\eta}(\tilde{\Theta}^i_L) > \tilde{\eta}(\tilde{\Theta}^O_L)$. To see this, first note that by construction $||\tilde{\lambda}^{c'}(\tilde{\Theta}^O_L \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}^i_L) - \lambda^{c'}(\Theta^O \cup \alpha(c') \setminus \Theta^i)||_\infty \overset{L\to\infty}{\to} 0$ and $||\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L) - \lambda^{c'}(\alpha(c'))||_\infty \overset{L\to\infty}{\to} 0$ by the absolute continuity of the measure over student abilities induced by η. Therefore, by uniform continuity of peer preferences (see **A4**), for sufficiently large L and for all $\tilde{\theta} \in \tilde{\Theta}^i_L$,

$$f^{\tilde{\theta},c'}(\tilde{\lambda}^{c'}_L(\tilde{\Theta}^O_L \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}^i_L)) - f^{\tilde{\theta},c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L)) > f^{\theta,c'}(\lambda^{c'}(\Theta^O \cup \alpha(c') \setminus \Theta^i)) - f^{\theta,c'}(\lambda^{c'}(\alpha(c'))) - \frac{\delta}{2} > \delta - \frac{\delta}{2} = \frac{\delta}{2}, \quad \text{(A.12)}$$

where the second inequality follows from the construction of Θ^i and the selection of δ. Because $||E - \tilde{E}||_{\mathbf{f}} < \frac{\delta}{4}$, for any L and all $\tilde{\theta} \in \tilde{\Theta}^i_L$,

$$g^{c'}(\tilde{\lambda}^{c'}_L(\tilde{\Theta}^O_L \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}^i_L)) - g^{c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L)) > -\frac{\delta}{2}. \quad \text{(A.13)}$$

Combining Equations A.12 and A.13 implies that for sufficiently large L and for all $\tilde{\theta} \in \tilde{\Theta}^i_L$, $f^{\tilde{\theta},c'}(\tilde{\lambda}^{c'}_L(\tilde{\Theta}^O_L \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}^i_L)) + g^{c'}(\tilde{\lambda}^{c'}_L(\tilde{\Theta}^O_L \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}^i_L)) > f^{\tilde{\theta},c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L)) + g^{c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L))$. This establishes openness, as desired.

Denseness: We now argue that the set of measures that admit a negative externality group is dense with respect to the $||\cdot||_{\mathbf{f}}$ norm. To do so, it suffices to consider a market $E = [\eta, \cdot, N, \Theta]$ where $\mathbf{f} \mapsto \Theta$ that does not admit any negative externality groups. Fix $\epsilon > 0$. We wish to show there exists a market $\tilde{E} = [\tilde{\eta}, \cdot, N, \tilde{\Theta}]$ such that $\mathbf{f} + \mathbf{g} \mapsto \Theta$ and $||E - \tilde{E}||_{\mathbf{f}} < \epsilon$, such that $\tilde{E}$ admits a negative externality group.

Consider market E and any assignment α such that there exists a program $c' \in C \setminus \{c_0\}$ with $\eta(\alpha(c')) > 0$. Consider any measurable subset $\Theta^I \subset \alpha(c')$ with $\eta(\Theta^I) = \delta > 0$, and let $\Theta^O = \emptyset$.

Define $\gamma(\delta) := \sup_{\theta \in \Theta^I} f^{\theta,c'}(\lambda^{c'}(\alpha(c'))) - f^{\theta,c'}(\lambda^{c'}(\alpha(c') \setminus \Theta^I)) + \frac{1}{\delta}$. Because E admits no negative externality groups, it must be that for some $\theta \in \Theta^I$, $f^{\theta,c'}(\lambda^{c'}(\alpha(c'))) \geq f^{\theta,c'}(\lambda^{c'}(\alpha(c') \setminus \Theta^I \cup \Theta^O)) = f^{\theta,c'}(\lambda^{c'}(\alpha(c') \setminus \Theta^I))$, where the last equality follows because $\Theta^O = \emptyset$. Therefore, $\gamma(\delta) > 0$. By uniform continuity (see **A4**), for sufficiently small δ, $\gamma(\delta) < \epsilon$ because as $\eta(\Theta^I) = \delta \to 0$, $\sup_{x \in [0,1]} |\lambda^{c',x}(\alpha(c')) - \lambda^{c',x}(\alpha(c') \setminus \Theta^I))| \to 0$. Take any such δ for which $\gamma(\delta) + \frac{1}{\delta} < \epsilon$.

We now construct market $\tilde{E}$. Construct sets $\tilde{\alpha}_L(c')$ and $\tilde{\Theta}^I_L$ as in the openness proof, and let L be sufficiently large such that $\gamma(\delta) + \frac{1}{\delta} > \sup_{\tilde{\theta} \in \tilde{\Theta}^I_L} f^{\tilde{\theta},c'}(\tilde{\lambda}(\tilde{\alpha}_L(c'))) - f^{\tilde{\theta},c'}(\tilde{\lambda}(\tilde{\alpha}_L(c') \setminus \tilde{\Theta}^I_L)) + \frac{1}{\delta}$.[4] For all $c \neq c'$ let $g^c(\cdot) = 0$, and for any measurable set $\beta \subset \tilde{\mathcal{A}}$ let

$$g^{c'}(\tilde{\lambda}(\beta)) = (\gamma(\delta) + \frac{1}{\delta}) \cdot \max \left\{ 0, 1 - \frac{\sup_{x \in [0,1]} |\tilde{\lambda}^{x,c'}(\tilde{\alpha}_L(c') \setminus \tilde{\Theta}^I_L) - \tilde{\lambda}^{x,c'}(\beta)|}{\sup_{x \in [0,1]} |\tilde{\lambda}^{x,c'}(\tilde{\alpha}_L(c') \setminus \tilde{\Theta}^I_L) - \tilde{\lambda}^{x,c'}(\alpha_L(c'))|} \right\} \tag{A.14}$$

By construction, $\tilde{\alpha}_L(c')$, $\tilde{\Theta}^I_L$, and $\tilde{\Theta}^O := \emptyset$ form a negative externality group because $\tilde{\eta}(\tilde{\Theta}^I_L) > 0 = \tilde{\eta}(\tilde{\Theta}^O)$ and

$$\begin{aligned} f^{\tilde{\theta},c'}(\tilde{\lambda}(\tilde{\alpha}_L(c'))) + g^{c'}(\tilde{\lambda}(\tilde{\alpha}_L(c'))) &= f^{\tilde{\theta},c'}(\tilde{\lambda}(\tilde{\alpha}_L(c'))) \\ &< \gamma(\delta) + f^{\tilde{\theta},c'}(\tilde{\lambda}(\tilde{\alpha}_L(c') \setminus \tilde{\Theta}^I_L)) \\ &< f^{\tilde{\theta},c'}(\tilde{\lambda}(\tilde{\alpha}_L(c') \setminus \tilde{\Theta}^I_L)) + g^{c'}(\tilde{\lambda}(\tilde{\alpha}_L(c') \setminus \tilde{\Theta}^I_L)), \end{aligned}$$

where the equality and final inequality follow from Equation A.14 and the first inequality follows from the construction of $\gamma(\delta)$ and selection of sufficiently large L as previously argued. Because $\gamma(\delta) < \epsilon, ||E - \tilde{E}||_{\mathbf{f}} < \epsilon$.

We complete this proof by showing that $\tilde{E}$ satisfies assumption **A4**. Uniform boundedness is satisfied because $g^c(\cdot) \in [0, \gamma(\delta) + \frac{1}{\delta}]$. Uniform continuity is satisfied due to the construction of $\mathbf{g}$: for any $\beta, \beta' \in \tilde{\mathcal{A}}$,

[4] Such an L exists because $\gamma(\delta) > \sup_{\theta \in \Theta^I} f^{\theta,c'}(\lambda^{c'}(\alpha(c'))) - f^{\theta,c'}(\lambda^{c'}(\alpha(c') \setminus \Theta^I)) \geq \sup_{\tilde{\theta} \in \tilde{\Theta}^I_L} f^{\tilde{\theta},c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c'))) - f^{\theta,c'}(\lambda^{c'}(\alpha(c') \setminus \Theta^I))$ by construction.

$$|g^{c'}(\tilde{\lambda}^{c'}(\beta))-g^{c'}(\tilde{\lambda}^{c'}(\beta'))| \leq |g^{c'}(\tilde{\lambda}^{c'}(\beta))-g^{c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}_L(c')\setminus\tilde{\Theta}^I_L))|+|g^{c'}(\tilde{\lambda}^{c'}(\beta'))-g^{c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}_L(c')\setminus\tilde{\Theta}^I_L))|$$

$$=(\gamma(\delta)+\frac{1}{\delta})\cdot\max\left\{0,1-\frac{\sup_{x\in[0,1]}|\tilde{\lambda}^{x,c'}(\tilde{\alpha}_L(c')\setminus\tilde{\Theta}^I_L)-\tilde{\lambda}^{x,c'}(\beta)|}{\sup_{x\in[0,1]}|\tilde{\lambda}^{x,c'}(\tilde{\alpha}_L(c')\setminus\tilde{\Theta}^I_L)-\tilde{\lambda}^{x,c'}(\alpha_L(c'))|}\right\}$$

$$+(\gamma(\delta)+\frac{1}{\delta})\cdot\max\left\{0,1-\frac{\sup_{x\in[0,1]}|\tilde{\lambda}^{x,c'}(\tilde{\alpha}_L(c')\setminus\tilde{\Theta}^I_L)-\tilde{\lambda}^{x,c'}(\beta')|}{\sup_{x\in[0,1]}|\tilde{\lambda}^{x,c'}(\tilde{\alpha}_L(c')\setminus\tilde{\Theta}^I_L)-\tilde{\lambda}^{x,c'}(\alpha_L(c'))|}\right\},$$

where the inequality follows from the triangle inequality, and the equality follows from Equation A.14. □

Proof of Part 2: We prove this result for $N=1$ and then discuss how it easily extends to the case in which $N>1$. Suppose in market $E=[\eta,\cdot,1,\Theta]$ with $\mathbf{f}\mapsto\Theta$ there exists a negative externality group. Following the logic of the argument in the previous part of Theorem 2, for any $\tilde{E}=[\tilde{\eta},\cdot,1,\tilde{\Theta}]$ such that $\mathbf{f}\mapsto\tilde{\Theta}$ and $||E-\tilde{E}||_{\mathbf{f}}=0$, then there exists a negative externality group, which we denote by sets $\tilde{\alpha}(c_1)$, $\tilde{\Theta}^I$, and $\tilde{\Theta}^O$.

We proceed to construct the desired such market $\tilde{E}$ for which the TIM process does not converge. Fix $\epsilon>0$ and let $\omega>0$ be such that for all $\tilde{\theta}$ and any $\beta,\beta'\in\tilde{\mathcal{A}}$, $|f^{\tilde{\theta},c_1}(\tilde{\lambda}^c_1(\beta))-f^{\tilde{\theta},c_1}(\tilde{\lambda}^c_1(\beta'))|<\epsilon$ if $||\tilde{\lambda}^{c_1}(\beta)-\tilde{\lambda}^{c_1}(\beta')||_\infty<\omega$. In what follows, ω will serve as the relevant lower bound on the support of student types in Assumption **A3**.

- There are five disjoint sets of students: $\tilde{\Omega}$ such that $\tilde{\eta}(\tilde{\Omega})=\omega$, $\tilde{\Theta}^i\subset\tilde{\Theta}^I$ where $\tilde{\eta}(\tilde{\Theta}^i)=(1-\omega)\tilde{\eta}(\tilde{\Theta}^I)$, $\tilde{\Theta}^o\subset\tilde{\Theta}^O$ where $\tilde{\eta}(\tilde{\Theta}^o)=(1-\omega)\tilde{\eta}(\tilde{\Theta}^O)$, $\tilde{\Theta}^u\subset\tilde{\alpha}(c_1)\setminus\tilde{\Theta}^I$ where $\tilde{\eta}(\tilde{\Theta}^u)=(1-\omega)\tilde{\eta}(\tilde{\alpha}(c_1)\setminus\tilde{\Theta}^I)$, and $\tilde{\Theta}^\ell$ where $\tilde{\eta}(\tilde{\Theta}^\ell)=(1-\omega)[1-\tilde{\eta}(\tilde{\Theta}^o)-\tilde{\eta}(\tilde{\Theta}^i)-\tilde{\eta}(\tilde{\Theta}^U)]$. Because these sets are disjoint, it follows that $\tilde{\eta}(\tilde{\Omega}\cup\tilde{\Theta}^i\cup\tilde{\Theta}^\ell\cup\tilde{\Theta}^o\cup\tilde{\Theta}^u)=1$.
- Let $v^{\tilde{\theta},c_1}$ be such that: $u^{\tilde{\theta}}(c_1|\tilde{\alpha})+\epsilon<u^{\tilde{\theta}}(c_0|\tilde{\alpha})$ for all $\tilde{\theta}\in\tilde{\Theta}^i$, $u^{\tilde{\theta}}(c_1|\beta)>u^{\tilde{\theta}}(c_0|\beta)$ for all $\beta\in\tilde{\mathcal{A}}$ and for all $\tilde{\theta}\in\tilde{\Theta}^o$, $u^{\tilde{\theta}}(c_1|\beta)>u^{\tilde{\theta}}(c_0|\beta)$ for all $\beta\in\tilde{\mathcal{A}}$ and for all $\tilde{\theta}\in\tilde{\Theta}^u$, $u^{\tilde{\theta}}(c_1|\beta)>u^{\tilde{\theta}}(c_0|\beta)$ for all $\beta\in\tilde{\mathcal{A}}$ and for all $\tilde{\theta}\in\tilde{\Omega}$, and $u^{\tilde{\theta}}(c_0|\beta)>u^{\tilde{\theta}}(c_1|\beta)$ for all $\beta\in\tilde{\mathcal{A}}$ and for all $\tilde{\theta}\in\tilde{\Theta}^\ell$.
- Scores at c_1 satisfy: $\tilde{\eta}(\theta\in\tilde{\Omega}|r^{\theta,c_1}<x)=\omega x$ for any $x\in[0,1]$, and $r^{\tilde{\theta}^\ell,c_1}<r^{\tilde{\theta}^o,c_1}<r^{\tilde{\theta}^i,c_1}<r^{\tilde{\theta}^u,c_1}$ for any $\tilde{\theta}^\ell\in\tilde{\Theta}^\ell$, any $\tilde{\theta}^o\in\tilde{\Theta}^o$, any $\tilde{\theta}^i\in\tilde{\Theta}^i$, and any $\tilde{\theta}^u\in\tilde{\Theta}^u$.
- $q^{c_1}=(1+\omega)\tilde{\eta}(\tilde{\Theta}^i\cup\tilde{\Theta}^u)$.

Let $\mu_0(c_1)=\tilde{\Theta}^i\cup\tilde{\Theta}^u\cup\{\tilde{\theta}\in\tilde{\Omega}|r^{\tilde{\theta},c_1}\geq 1-\tilde{\eta}(\tilde{\Theta}^i\cup\tilde{\Theta}^u)\}$. Therefore, by construction of q, $\tilde{\eta}(\mu_0(c_1))=q^{c_1}$.

Note that $||\tilde{\lambda}^{c_1}(\mu_0(c_1))-\tilde{\lambda}^{c_1}(\tilde{\alpha}(c_1))||_\infty<\omega$, and so by the construction of preferences and the uniform continuity of $f^{\theta,c_1}(\cdot)$, all students $\tilde{\theta}\in\tilde{\Theta}^u\cup\tilde{\Theta}^o\cup\tilde{\Omega}$ have preferences $u^{\tilde{\theta}}(c_1|\mu_0)>u^{\tilde{\theta}}(c_0|\mu_0)$,

and all students $\tilde{\theta} \in \tilde{\Theta}^i \cup \tilde{\Theta}^\ell$ have preferences $u^{\tilde{\theta}}(c_0|\mu_0) > u^{\tilde{\theta}}(c_1|\mu_0)$. Given these preferences and the student scores defined above, $\mu_1(c_1) = \tilde{\Theta}^u \cup \tilde{\Theta}^o \cup \{\tilde{\theta} \in \tilde{\Omega} | r^{\tilde{\theta},c_1} \geq \tau\}$, where τ is defined implicitly by the infimum value of $x \geq 0$ such that $\tilde{\eta}(\tilde{\Theta}^u) + \tilde{\eta}(\tilde{\Theta}^o) + \tilde{\eta}(\{\tilde{\theta} \in \tilde{\Omega} | r^{\tilde{\theta},c_1} \geq x\}) \leq q^{c_1}$. Note that $||\tilde{\lambda}^{c_1}(\mu_1(c_1)) - \tilde{\lambda}^{c_1}(\tilde{\Theta}^o \cup \tilde{\alpha}(c_1) \setminus \tilde{\Theta}^i)||_\infty < \omega$, and so by the construction of preferences and the uniform continuity of $f^{\tilde{\theta},c_1}(\cdot)$, all students $\tilde{\theta} \in \tilde{\Theta}^u \cup \tilde{\Theta}^o \cup \tilde{\Theta}^i \cup \tilde{\Omega}$ have preferences $u^{\tilde{\theta}}(c_1|\mu_1) > u^{\tilde{\theta}}(c_0|\mu_1)$, and all students $\tilde{\theta} \in \tilde{\Theta}^\ell$ have preferences $u^{\tilde{\theta}}(c_0|\mu_1) > u^{\tilde{\theta}}(c_1|\mu_1)$. Given these preferences and the student scores defined above, $\mu_2(c_1) = \mu_0(c_1)$. Therefore, the TIM process cycles, and does not converge.

A similar construction is possible for any N. The preceding logic can be modified such that students $\tilde{\theta} \in \tilde{\Omega}$ are "uniformly at random" likely to most prefer any program $c \in C \setminus \{c_0\}$ for all assignments, and that no student $\tilde{\theta} \notin \tilde{\Omega}$ finds any program $c \neq c_1$ preferable to c_0 for any assignment. □

Proof of Part 3. We prove the claim constructively; we present the following example which demonstrates the claim for $N = 1$ and $q^{c_1} \geq 1$, and then discuss how the example easily extends to arbitrary N and q.

Example 3. *Let $N = 1$. The lone program c_1 has $q^{c_1} \geq 1$ measure of seats, and let $r^{\theta,c} = r^{\theta,c_0}$ for all $\theta \in \Theta$. Let $s^{c_1}(\lambda(\alpha))$ be the mean ability of students assigned to c_1 in assignment α, that is*

$$s^{c_1}(\lambda(\alpha)) = \frac{1}{\lambda^{c_1,1}(\alpha)} \int_0^1 y d\lambda^{c_1,y}(\alpha).$$

$\gamma < 1$ measure of students have weak peer preferences and receive strictly positive utility from attending c_1 regardless of the assignment.[5] *Students with weak peer preferences have scores r^{θ,c_1} that are "uniformly distributed" over $[0,1]$. The remaining $1-\gamma$ measure of students have strong peer preferences and receive utility $v^{\theta,c_1} - f(s^{c_1}(\lambda), r^{\theta,c_1})$ from matching with the program at assignment α, where*

$$f(s^{c_1}(\lambda(\alpha)), r^{\theta,c_1}) = \begin{cases} 0 \text{ if } r^{\theta,c_1} \geq \frac{1}{2} \text{ and } s^{c_1}(\lambda(\alpha)) \leq \frac{1}{2} \\ 0 \text{ if } r^{\theta,c_1} < \frac{1}{2} \text{ and } s^{c_1}(\lambda(\alpha)) > \frac{1}{2} \\ k|\frac{1}{2} - s^{c_1}(\lambda(\alpha))| \text{ otherwise} \end{cases}$$

for some $k > \frac{8}{1-\gamma}$. Students with strong peer preferences have intrinsic match values v^{θ,c_1} and scores r^{θ,c_1} that are independently and uniformly distributed over $[0,1]$. The peer preference term $f(\cdot,\cdot)$ reflects that students want their own score to be different from the average scores of their peers, and suffer loss proportional to the average score of students if they are in the "majority half" of the class.

[5] We include these students with weak peer preferences so as to satisfy Assumption **A3**.

We claim that the matching μ_* such that $\mu_*(\theta)=c_1$ for all $\theta\in\Theta$ is the unique stable matching.[6] *Clearly, μ_* is a matching since $q^{c_1}\geq 1$. Then $\lambda_*=\lambda(\mu_*)$ has the property that $\lambda_*^{c_1,(y,y)}=y$ for all $y\in[0,1]$. Note that $\mu_*=A(0,\lambda_*)$ is stable: it is market clearing (i.e. $p_*^{c_1}=0$) and satisfies rational expectations, i.e. $s^{c_1}(\lambda_*)=\frac{1}{2}$ and so all students prefer to attend c_1 to the outside option. Furthermore, it is easy to see that this is the unique stable matching. Any market clearing matching $\mu_\prime$ must satisfy $p_\prime^{c_1}=0$. If $s_\prime^{c_1}=s^{c_1}(\lambda(\mu_\prime))<\frac{1}{2}$ all the students with scores $r^{\theta,c_1}\geq\frac{1}{2}$ prefer to be matched to c_1 while a non-negligible set of students with scores $r^{\theta,c_1}<\frac{1}{2}$ prefer not to be matched to c_1. This implies that $s^{c_1}(\lambda(A(p_\prime,s_\prime)))>\frac{1}{2}>s_\prime^{c_1}$. Therefore, $(p_\prime,\lambda(\mu_\prime))$ does not satisfy rational expectations, and so $\mu_\prime$ is not stable. A similar argument follows if $s_\prime^{c_1}>\frac{1}{2}$.*

We further claim that the TIM process does not converge for any $s_0^{c_1}=s^{c_1}(\lambda(\mu_0))\neq\frac{1}{2}$. Recalling that $s(\cdot)$ is a function of λ, if the sequence $s_1,s_2,...$ does not converge, then neither does the sequence $\lambda_1,\lambda_2,...$. In words, if the market does not initialize with a distribution that induces an average score exactly equal to that in the unique stable matching, the TIM process will not generate a stable matching in any $t\geq 1$.

To show this claim, let $s_0^{c_1}=\frac{1}{2}-\delta$ for some $\delta\in(0,\frac{1}{2}]$; due to symmetry of market fundamentals, similar logic holds if $\delta\in[-\frac{1}{2},0)$ instead. We present two exhaustive cases. First, suppose that $k\delta\geq 1$. Then in μ_1 none of the students with $r^{\theta,c_1}<\frac{1}{2}$ who have strong peer preferences will enroll in c_1, and all other students will. Therefore,

$$s^{c_1}(\lambda(\mu_1))=\frac{\frac{1}{4}(\frac{1}{2}\gamma)+\frac{3}{4}\frac{1}{2}}{\frac{1}{2}(1+\gamma)}=\frac{3+\gamma}{4(1+\gamma)}\quad\text{and}\quad s^{c_1}(\lambda(\mu_2))=\frac{1+3\gamma}{4(1+\gamma)}.$$

From here, a cycle forms: for any odd $t>1$, $s^{c_1}(\lambda(\mu_t))=s^{c_1}(\lambda(\mu_1))$ and $s^{c_1}(\lambda(\mu_{t+1}))=s^{c_1}(\lambda(\mu_2))$, meaning that the TIM process does not converge to the unique stable matching.

Next, suppose $k\delta<1$. By a similar calculation, we have that

$$s^{c_1}(\lambda(\mu_1))=\frac{\gamma+(1-\gamma)(1-k\delta)+3}{4(1+\gamma+(1-\gamma)(1-k\delta))}$$

For $k\geq\frac{8}{1-\gamma}$, as we have assumed, we claim that $s^{c_1}(\lambda(\mu_1))\geq\frac{1}{2}+\delta$. To see this, note that $\frac{\gamma+(1-\gamma)(1-k\delta)+3}{4(1+\gamma+(1-\gamma)(1-k\delta))}-\frac{1}{2}-\delta\geq 0$ if and only if $k-\gamma k-8+4k\delta-4\gamma k\delta\geq 0$. Since $\gamma<1$, $k-\gamma k-8\geq 0$ implies the desired condition. Noting the symmetry of the market, it is the case that for odd t, the sequence $s_t,s_{t+2},s_{t+4}...$ is non-decreasing where each element is strictly larger than $\frac{1}{2}$ and $s_{t+1},s_{t+3},s_{t+5},...$ is non-increasing where each element is strictly smaller than $\frac{1}{2}$. Therefore, the TIM process does not converge to the unique stable matching.

This example can easily be extended to $N>1$ and arbitrary q (similarly to how we extended the argument in the *Proof of Part 2* of the current Theorem). Specifically, we can embed the example

[6] We ignore indeterminacy caused by a zero measure set of students who are indifferent between being matched to the program and not, which does not affect our analysis.

above into any market with $N>1$, and consider a set of students $\Theta' \subset \Theta$ with $\eta(\Theta') \leq q^{c_1}$ whose scores are uniformly distributed over $[0,1]$ and whose utility for attending all programs other than c_1 is strictly negative given any assignment. In the example above $\Theta' = \Theta$ and $\eta(\Theta') = \eta(\Theta) = 1 \leq q^{c_1}$. □

Proposition 3

Proof of Part 1: This follows from the "Only if" part of the proof of part 2 of Proposition 2. □

Proof of Part 2: This follows from the "If" part of the proof of part 2 of Proposition 2. □

Proof of Part 3: Suppose the TFM mechanism terminates in period $\tau^* > 0$. Because the final matching is not constructed at any step $\tau < \tau^*$ in which $\lambda(\mu_\tau)$ is being updated, and because each $\lambda(\mu_\tau)$ is unaffected by the submitted preferences of any zero measure set of student, no student affects the final matching by misreporting preferences in any step $\tau < \tau^*$. Therefore, we only regard incentives to misreport at the final step.

Fix $\epsilon > 0$. Termination of the TFM mechanism implies that $||\lambda_{\tau^*} - \lambda(\mu_{\tau^*})||_\infty < \delta$. Assuming (almost) all students $\theta' \in \Theta$ report preferences $\succeq^{\theta'|\lambda_{\tau^*}}$, we have that any $\theta \in \Theta$ can profitably misreport her preferences only if $\succeq^{\theta|\lambda(\mu_{\tau^*})} \neq \succ^{\theta|\lambda_{\tau^*}}$. By **A4'** (which holds by Lemma A.1), there exists δ_1^* such that $\eta(\{\theta | \succeq^{\theta|\lambda(\mu_{\tau^*})} \neq \succ^{\theta|\lambda_{\tau^*}}\}) < \epsilon$ for any stopping rule $\delta < \delta_1^*$. Therefore, the measure of students who can profitably misreport preferences is arbitrarily small for sufficiently small δ, as desired. Also, there exists δ_2^* such that for any $\delta < \delta_2^*$, $|u^\theta(c|\lambda_{\tau^*}) - u^\theta(c|\lambda(\mu_{\tau^*})| < \epsilon$ for all θ and all c by part 2 of Proposition 3 and uniform continuity of peer preferences, (see **A4**). Therefore, the utility gain of misreporting is strictly less than ϵ for all θ. Letting $\delta^* := \min\{\delta_1^*, \delta_2^*\}$ completes the proof. □

Proof of Part 4: Suppose the TFM mechanism terminates at step $\tau^* = K \cdot T + t$. Note that the stopping criterion is independent of K, T, t, i.e. τ^* depends only on δ and Λ_0^γ. There exists T_1 such that for any $T > T_1$, $K = 0$ and $\tau^* = t$. Moreover, for any $\varepsilon > 0$ there exists $T_2 > T_1$ such that $\frac{t}{T} = \frac{\tau^*}{T} < \varepsilon$ for any $T > T_2$. $K = 0$ implies that the measure of students that reports ROLs more than twice is zero, and $t = \tau^*$ implies that the share of submarkets that report ROLs twice is $\frac{\tau^*}{T}$. Recall our assumption that $\eta(\Theta_\ell) \to 0$ for all $\ell \in 1, ..., T$ as $T \to \infty$. Therefore, there exists T^* such that the measure of students asked to report ROLs twice is given by $\sum_{\ell=1}^{\tau^*} \eta(\Theta_\ell) < \epsilon$ for any $T > T^*$. □

B Preferences over the entire distribution of peer ability

We show by construction that certain functional forms of peer preferences cannot be represented via (any finite number of) summary statistics of "ability." Let $E = [\eta, q, N, \Theta]$. For all students

θ and all programs $c \in C \setminus \{c_0\}$ let $u^\theta(c|\alpha) = v^{\theta,c} + f^{\theta,c}(\lambda^c(\alpha))$, where

$$f^{\theta,c}(\lambda^c(\alpha)) = -\int_0^1 |\lambda^{c,y}(\alpha) - \lambda^y_\theta| dy, \qquad \lambda^y_\theta = \begin{cases} 0 \text{ if } y \leq r^{\theta,c_0} \\ 1 \text{ if } y > r^{\theta,c_0} \end{cases}.$$

This functional form represents that each student θ has a "bliss point" and most prefers to attend a program c when her peers at program c all have ability equal to r^{θ,c_0}. For any assignment, the peer cost of attending program c is the difference in area between the actual distribution of abilities at program c and her bliss point distribution.

Let a *summary statistic* of abilities at program c be a function $s^c : \Lambda \to [0,1]$. For $\lambda \in \Lambda^{N+1}$ let $s(\lambda) = \times_{c \in C} s^c(\lambda)$ be the vector of summary statistics. Fix any finite number J of summary statistics $\{s^j(\lambda)\}_{j=1,...,J}$. We claim the peer preferences above cannot be represented via a utility function over $\{s^j(\lambda)\}_{j=1,...,J}$. To see this, fix θ and c, and let θ's utility over program c be characterized as above. First note that the subset of assignments $\hat{\mathcal{A}}$ such that $f^{\theta,c}(\lambda(\alpha)) \neq f^{\theta,c}(\lambda(\alpha'))$ for any distinct $\alpha, \alpha' \in \hat{\mathcal{A}}$ is open and dense in $\mathcal{A}$.[7] Therefore, it suffices to show that there does not exist a function $h^{\theta,c} : [0,1]^J \to \Lambda$ such that $h^{\theta,c}(s^1(\lambda(\alpha)),...,s^J(\lambda(\alpha))) = \lambda^c(\alpha)$ for all α. The set Λ has cardinality equal to that of the continuum (Moschovakis, 2006, page 18). Since $h^{\theta,c}(\cdot)$ has only J arguments, it cannot be surjective, implying that there is some α such that $h^{\theta,c}(s^1(\lambda(\alpha)),...,s^J(\lambda(\alpha))) \neq \lambda^c(\alpha)$.

C Calculating the bonus point distribution

In this section, we calculate the distribution of bonus points allocated to students using data from ROLs and offers in the years 2015-2016. We use observable information on student θ's ATAR score, X_θ, and a binary admissions outcome for student θ to program c included on the ROL, $A_{\theta c}$, to calculate the distribution of bonus points $Z_{\theta c}$. We denote the CYS of program c as ϕ_c. Recall from Section V.A, that student θ is admitted (assuming she has not been admitted to a higher-ranked program) to program c based on whether the sum of her ATAR score and bonus points exceeds c's CYS, $A_{\theta c} = \mathbb{1}[X_\theta + Z_{\theta c} \geq \phi_c]$. We assume bonus points are distributed independently and identically across student-program pairs and denote the distribution by $F_Z(z) = P(Z_{\theta c} \leq z)$. We can therefore identify the distribution of bonus points by one minus the probability of admission

$$F_Z(\phi_c - X_\theta) = 1 - P(A_{\theta c} = 1 | \phi_c - X_\theta). \tag{A.15}$$

Since $(A_{\theta c}, \phi_c, X_\theta)$ are all observable, this CDF is nonparametrically identified.

[7] Any two assignments α, α' that differ among a positive measure set of students will by construction yield $\lambda(\alpha) \neq \lambda(\alpha')$. Recalling that we endow the set Λ with metric induced by the $||\cdot||_\infty$ norm, the subset of ability distributions $\hat{\Lambda}$ such that $f^{\theta,c}(\lambda) \neq f^{\theta,c}(\lambda')$ for any $\lambda, \lambda' \in \hat{\Lambda}$ is open (Take any $\lambda, \lambda' \in \Lambda$ such that without loss of generality $f^{\theta,c}(\lambda) = f^{\theta,c}(\lambda') + \delta$ for some $\delta > 0$. There exists sufficiently small ϵ such that $|f^{\theta,c}(\lambda) - f^{\theta,c}(\lambda'')| < \delta$ for any λ'' such that $||\lambda(\cdot) - \lambda''(\cdot)||_\infty < \epsilon$. Therefore, it must be that $f^{\theta,c}(\lambda) \neq f^{\theta,c}(\lambda'')$.) and dense (Fix $\epsilon > 0$. Take any $\lambda, \lambda' \in \Lambda$ such that $f^{\theta,c}(\lambda) = f^{\theta,c}(\lambda')$. It is easy to see that there exists some λ'' such that $||\lambda(\cdot) - \lambda''(\cdot)||_\infty < \epsilon$ such that $f^{\theta,c}(\lambda) \neq f^{\theta,c}(\lambda'')$).

We estimate this function in the following ways. First, we use an isotonic regression to estimate F_Z, which imposes the shape-constraint that the distribution is non-decreasing (Barlow and Brunk, 1972). We define student-by-program observations as eligible if a student was not admitted to a program higher on her ROL. We exclude a small number of students with abnormal admission records such that they are not admitted to any program despite having at least one program that they should have received admittance without bonus points or being admitted to a program not on their observed ROL.[8]

Figure A.1 shows the distribution estimated across all programs in the dark line and separate estimates for each rank of the ROL in gray. As expected, the probability of "negative bonus points" is close to zero and likely only reflects some amount of measurement error. The empirical CDF shows discontinuities in the distribution around 0 and 5 bonus points. The CDF increases at a slower rate but does not fully converge to one until approximately 30 bonus points. Overall, the distribution of bonus points implies a mean of 5.71 bonus points and standard deviation of 6.51 bonus points. We also find limited heterogeneity when plotting empirical CDFs calculated at each rank of the ROL. This suggests the assumption of a common distribution of bonus points across students may be a reasonable one, as otherwise one would expect the distribution to diverge as students become increasingly selected in the higher ranks.

D Alternative explanations for empirical findings

In this appendix, we describe three alternative models of student behavior proposed in the literature, as discussed in Section V.C.3. Common to all of these alternative models is a de facto cost of rejection from a program for students. We show that all of these models predict a discontinuity in the probability that a student optimally ranks a program when there is an attendant discontinuity in admissions probability. We find no such relationship in our regression discontinuity analysis (see Figure 1), and therefore, we find little evidence to support these alternative models.

Throughout this appendix section, we omit time indices and assume that each student θ draws a value $v^{\theta,c} \sim G_c$ independently for each program c, where each each distribution G_c:

1. has an associated continuous density g_c, where $g_c(x)$ is positive and bounded away from 0 if and only if $x \in [0,1]$,
2. For any $\epsilon > 0$ there exists $\delta > 0$ such that $||G_c - G_{c'}||_\infty < \epsilon$ if $|PYS_c - PYS_{c'}| < \delta$.

1) is a standard assumption generating full support of preferences over programs, and 2) is a continuity condition—students have, in aggregate, similar preferences for programs with similar observables.

[8] Recall that a small number of students are admitted to programs through non-ATAR-related channels.

We assume there is a non-increasing function $p(\cdot)$ that maps the difference between a program's PYS and a student's ATAR score into an expected probability of admission, and we take this probability to be independent (conditional on the score gap) across programs. To match our empirical setting, we assume that $p(0)=1$ and there is a discontinuity at 0—each student perceives a substantially lower probability of admission to a program whose PYS just exceeds her own ATAR score, compared to a program with a PYS just below her ATAR score. This is justified by a non-zero probability of receiving zero bonus points at a program.[9] Let $\Delta = p(0) - \lim_{x \to 0^+} p(x)$.

D.1 Incorrect beliefs

One potential model is that students do not fully understand the deferred acceptance mechanism, with a well-known concern being that students do not realize that rejection from a program does not reduce the probability of matching with a lower-ranked program on their ROL (Li, 2017).

We assume that each student θ perceives that ranking a program on her (post-)ROL and being rejected means there is a $\kappa > 0$ probability that she is then also rejected from all lower-ranked programs on her ROL. (Our conclusions extend if rejection probability at subsequently-ranked programs depends on the identities of higher-ranked programs).

Suppose that student θ has an ATAR score of r^{θ,c_0}, and for some small ϵ, consider programs c_1 and c_2, where $PYS_{c_1} \in [r^{\theta,c_0}-\epsilon, r^{\theta,c_0}]$ and $PYS_{c_2} \in (r^{\theta,c_0}, r^{\theta,c_0}+\epsilon)$ where $v^{\theta,c_1}, v^{\theta,c_2} \geq 0$. Because $p(\cdot)$ is non-increasing, it must be that the student perceives at least $1-p(\epsilon) \geq \Delta$ higher probability of being admitted to program c_1 than c_2. Consider any ROL $\succ$ in which c_1 is ranked, there exists some c ranked below c_1, and either c_2 is not ranked c_2 is ranked below c_1. Also consider ROL $\succ'$ which ranks all programs besides c_1 and c_2 and instead switches the ranking of c_1 and c_2 if both are ranked or replaces c_1 with c_2 otherwise.

It suffices to consider the expected utility difference between $\succ$ and $\succ'$ conditional on being rejected from all programs ranked above c_1 according to $\succ$ and not triggering the perceived κ probability of subsequent rejection from all programs ranked above c_1 according to $\succ$. By the assumption that c_1 is not the lowest-ranked program according to $\succ$, it must be that the student receives a "continuation value" of $\bar{v} \geq 0$ if she is not admitted to program c_1 and also does not trigger automatic rejection at all subsequent programs.

Because of the perceived risk associated with rejection, the student will prefer to submit $\succ$ over $\succ'$ if and only if $v^{\theta,c_2} \cdot p(PYS_{c_2} - r^{\theta,c_0}) + (1-p(PYS_{c_2} - r^{\theta,c_0})) \cdot (1-\kappa) \cdot \bar{v} \geq v^{\theta,c_1}$, which implies $v^{\theta,c_2} - v^{\theta,c_1} \geq (1-p(PYS_{c_2} - r^{\theta,c_0}))\left[(1-\kappa)\bar{v} + v^{\theta,c_2}\right] \geq \Delta\left[(1-\kappa)\bar{v} + v^{\theta,c_2}\right] \geq \Delta \cdot v^{\theta,c_2}$, where the second inequality follows because $1-p(\epsilon) \geq 1-p(PYS_{c_2} - r^{\theta,c_0}) \geq \Delta$ and the final inequality

[9] An identical argument can be made at a score gap of 5 points in addition to a score gap of 0, as Figure 1 finds discontinuities in admissions probabilities at both values.

follows because $\bar{v} \geq 0$ and $\kappa \in (0,1]$.

For sufficiently small ϵ the probability that $v^{\theta,c_2} - v^{\theta,c_1} \geq 0$ is approximately $\frac{1}{2}$. Therefore, because $\Delta > 0$, the probability that $v^{\theta,c_2} - v^{\theta,c_1} \geq \Delta \cdot v^{\theta,c_2}$ is strictly less than $\frac{1}{2}$ if v^{θ,c_2} is bounded away from zero. Clearly, due to our assumptions on the continuity of g_c for all c, the probability that v^{θ,c_2} is (arbitrarily) bounded away from zero is (arbitrarily) large. Therefore, the probability that θ prefers to submit $\succ'$ instead of $\succ$ is strictly greater than, and bounded away from, $\frac{1}{2}$.

Averaging over all students, this logic implies that the share of students who prefer to submit an ROL in which they rank a program with a PYS just exceeding their own ATAR score is discontinuously lower than the share of students who would prefer to switch said program with a different program with a PYS equalling, or just lower than, their ATAR score.

D.2 Non-classical preferences

Non-classical utility functions can also explain some non-standard behavior in matching markets. Dreyfuss et al. (2021) and Meisner and von Wangenheim (2023) study a model in which students have expectations-based loss aversion. As a result, they may fail to rank otherwise desirable options in strategy-proof mechanisms to avoid disappointment from rejection. Meisner (2022) studies a model where students explicitly dislike rejection from programs.

We assume that each student θ perceives a cost for each program she ranks on her ROL that she is rejected from (i.e. any program that the student ranks above her assigned program). We take this cost to be some constant $\kappa > 0$, although our claims apply if we condition this cost on the identity of the program. Following the argument in Section D.1, the student will prefer to submit $\succ$ over $\succ'$ if and only if $v^{\theta,c_2} \cdot p(PYS_{c_2} - r^{\theta,c_0}) + (1 - p(PYS_{c_2} - r^{\theta,c_0})(\bar{v} - \kappa) \geq v^{\theta,c_1}$. One can again see that for sufficiently small ϵ, the probability that θ prefers to submit $\succ'$ instead of $\succ$ is strictly greater than, and bounded away from, $\frac{1}{2}$.

D.3 Optimal information acquisition

One other potential explanation is that student preferences change over time. This could possibly be due to exogenous factors (e.g. news coverage of a scandal at a program just prior to submission of the post-ROL) or strategic choices to acquire information about programs (Immorlica et al., 2020; Grenet et al., 2022; Hakimov et al., 2023), where students have incentives not to "waste" information acquisition costs on programs they will be rejected from.

From a timing standpoint, only one month separates our observation of the pre- and post-ROLs, mitigating opportunities for information from exogenous factors. Moreover, our analysis in Section V.C.3 allows for preference drift over time, further limiting the possible impact of exogenous information in explaining our empirical findings.

We therefore investigate the potential strategic choice of students to acquire information about

programs. Formally, suppose that for each student θ and each program c, $v^{\theta,c}$ represents a signal of θ's value for attending program c. Student θ's value for matching with program c is $\hat{v}^{\theta,c} = v^{\theta,c} + \sigma^{\theta,c}$ where $\sigma^{\theta,c} \sim U(-\kappa,\kappa)$ independently across students and programs, for some $\kappa > 0$. Each student θ can privately learn her draw $\hat{v}^{\theta,c}$ for up to one program prior to matching. (Although we assume an "all-or-nothing" information acquisition framework, our conclusions likely extend to many more nuanced frameworks.) If student θ matches to program c, her utility is $U(\hat{v}^{\theta,c})$ where $U(\cdot)$ is a bounded, weakly increasing, and strictly concave function from $[-\kappa, 1+\kappa] \to [0,1]$. This captures that students prefer programs for which they have high draws, and the concavity ensures risk aversion.

Again, consider the case in which student θ has an ATAR score of r^{θ,c_0}, and for some small ϵ, consider programs c_1 and c_2, where $PYS_{c_1} \in [r^{\theta,c_0} - \epsilon, r^{\theta,c_0}]$ and $PYS_{c_2} \in (r^{\theta,c_0}, r^{\theta,c_0} + \epsilon)$. We make four claims: First, holding fixed the ROLs of other students, each θ is weakly better off if she learns her value for some program c. In the absence of learning her values for any program, she has a weakly dominant strategy to rank programs in descending order of her signals. Upon learning the value for any program, she will optimally alter this order if and only if the learned utility for the selected program rises or falls below the expected utility from another.

Second, consider two potential signal vectors for student θ, $v^\theta = (v^{\theta,c_0}, v^{\theta,c_1}, ..., v^{\theta,c_N})$ and $\tilde{v}^\theta = (\tilde{v}^{\theta,c_0}, \tilde{v}^{\theta,c_1}, ..., \tilde{v}^{\theta,c_N})$, such that $\tilde{v}^{\theta,c} = v^{\theta,c}$ for all $c \notin \{c_1, c_2\}$, $\tilde{v}^{\theta,c_1} = v^{\theta,c_2}$, and $\tilde{v}^{\theta,c_2} = v^{\theta,c_1}$. That is, $\tilde{v}^\theta$ is obtained from v^θ by permuting the signals of c_1 and c_2. Consider the case in which θ optimally learns $\hat{v}^{\theta,c_2}$ upon receiving signal vector v^θ (we ignore non-generic and non-payoff relevant cases in which there are multiple optimal selections). According to her weakly dominant strategy, θ will rank programs in terms of their expected utility (where her expected utility for c_2 is $U(\hat{v}^{\theta,c_2})$).[10] We claim that θ must then optimally learn $\hat{v}^{\theta,c_1}$ upon receiving signal vector $\tilde{v}^\theta$. Recall that σ^{θ,c_1} and σ^{θ,c_2} are independently and identically distributed, and that θ is guaranteed entry to c_1 but not c_2. Conditional on not matching with a program preferred to c_2 upon observing v^θ and learning $\hat{v}^{\theta,c_2}$, there is a probability of at least $\Delta > 0$ that θ is not admitted to c_2 and the information gathered is therefore payoff irrelevant (the first claim establishes that information acquisition improves expected payoffs). The probability of rejection also implies that θ does not always (i.e. for almost every draw of signals) optimally learn $\hat{v}^{\theta,c_2}$ upon receiving signal vector v^θ if she optimally learns $\hat{v}^{\theta,c_1}$ upon receiving signal vector $\tilde{v}^\theta$.

Third, for sufficiently small ϵ, θ is ex-ante more likely to optimally learn v^{θ,c_1} than v^{θ,c_2}. This follows from the second bullet and the assumption that the signal distributions G_{c_1} and G_{c_2} are arbitrarily close for sufficiently small ϵ and therefore, $\tilde{v}^\theta$ and v^θ are nearly equally likely to occur.

[10] Depending on θ's ATAR score, there are payoff equivalent ROLs that omit programs with zero probability of acceptance, or programs that are dispreferred to others which guarantee acceptance. Our conclusions will hold regardless of which of these ROLs is selected.

Fourth, student θ is, for sufficiently small $\epsilon > 0$, discontinuously more likely to optimally submit $\succ'$ than $\succ$. This follows from the third claim, and the fact that $U(\cdot)$ is strictly concave. Therefore, resolution of uncertainty provides student θ an expected utility "boost" from that program.

E Panel-based analysis

This appendix complements the analysis presented in Section V.C.3 by employing an alternative panel-based event-study strategy to reinforce and validate our findings regarding student peer preferences. While Section V.C.3 examines student program choices in response to newly revealed relative peer ability, this research design leverages panel variation in observable program peer ability (PYS) to study demand. Causal identification occurs through counterfactual comparisons between programs matched on similar historical levels and trends, using a transparent event-study framework. A simple example illustrates the empirical design. Consider two programs, c and c', that exhibit identical PYSs for several years prior to year $t = 0$. In year $t = -1$, the CYS for program c happens to increase more than that of program c'. In the subsequent year $t = 0$, the PYS of program c is therefore higher than that of program c'. Our empirical analysis investigates how students respond to this newly observable increase in peer ability at program c.

In keeping with much of the related literature (e.g. Hastings et al., 2009; Abdulkadiroğlu et al., 2017; Luflade, 2019), we restrict the analysis sample to only include students who submitted fewer choices than the maximum allowed (i.e., nine).[11] This restriction ensures that observed ROLs reflect ordinal student preferences. A student who prefers weakly fewer than nine programs to her outside option has a weakly dominant strategy to truthfully include all preferred programs, ordered according to her ordinal preferences (Haeringer and Klijn, 2009). Therefore, limiting our sample removes any incentive for strategic application behavior.

We operationalize the research design using the following regression specification:

$$Y_{cyt} = \sum_{j=-5, j\neq -2}^{3} \beta_j \mathbb{1}[j = t] PYS_{cy} + \delta_{g(c,y),t} + \gamma_{cy} + \varepsilon_{cyt} \tag{A.16}$$

where Y_{cyt} measures applicant characteristics (e.g., log average test score, log number of applicants) to program c in reference to focal year y for relative year $t \in \{-5,...,3\}$. This event-study specification flexibly estimates how available peer information (PYS_{cy}) for program c in year y relates to student outcomes at various relative time periods (t) before and after changes in observable peer ability occur. As such, the panel dataset is structured so that each program-year combination represents a potential event characterized by changing observable peer information, generating

[11] In our data, 60% of students submit final ROLs with strictly fewer than nine programs. Note that even if these students are not representative of those who list nine programs, the existence of peer preferences in a large subsample of the population can significantly affect stability in the market.

relative time periods analogous to traditional event-study designs.

Central to this design, for each program c and year y, we restrict comparisons to other programs with the same PYS in years $t=-1$ to $t=-3$.[12] We do so by constructing explicit groups, denoted $g=g(c,y)$, which match programs within the same focal year based on identical prior-year PYS.[13] The specification includes relative-time-by-group fixed effects, $\delta_{g(c,y),t}$, enforcing that all identification arises exclusively from comparisons between programs within the same matched group at each relative time period. While not strictly necessary, we further include "unit" fixed effects, here being program-focal year fixed effects, γ_{cy}, which causes us to normalize the within-group comparisons relative to the difference in a chosen period (we choose $t=-2$), such that $\beta_{-2}=0$.[14] We cluster our standard errors at the program level to account for arbitrary correlation in errors over time within programs.

The key assumption of this research design is a parallel trends assumption: within group g, programs experiencing an increase in their PYS would have evolved similarly in outcomes to programs which did not experience changes to their PYS. We view this as an ex-ante plausible assumption as the group structure restricts comparisons between programs similar in terms of unmatched applicant observables and entry levels in the pre-period and hence the programs could likely continue on similar paths. Importantly, the group-by-year fixed effects flexibly control for program differences that could confound the estimated relationship, such as fixed or commonly-evolving program prestige and quality.

The event-study and its normalization to period $t=-2$ help assess this assumption in multiple ways. First, as is standard, we can use the event-study to provide an indication of whether there appears to be empirical evidence consistent with the parallel trends assumption prior to students observing the change in information.

Second, the observable peer information, the PYS in period $t=0$, is the admission cutoff without bonus points in year $t=-1$ (i.e., the CYS). Therefore, if changing the CYS has a direct effect on or correlation with our outcomes, this will be evident from β_{-1}. Finding that $\beta_{-1}=0$ addresses the natural concern that observed changes in peer ability may reflect systematic program-specific improvements—such as advertised enhancements to teaching quality—rather than plausibly exogenous fluctuations in the PYS from year to year. Such violations of our identifying assumptions have

[12] We do not match on periods -4 or earlier nor on other outcomes, as such periods and outcomes can be potentially used to test identifying assumptions. We additionally require programs to have at least 15 applicants in the first matching period, -3, to study well-defined programs who consistently receive applicants.

[13] As the group definition is stringent, many observations do not have a natural comparison group and hence are not used in estimation.

[14] If we do not include this unit fixed effect and do not normalize the coefficients, across all outcomes, none of the pre-period coefficients are statistically significant, as expected.

clear empirical implications. Specifically, if changes in PYS reflect unobservable program-specific improvements, such as advertised enhancements, we should detect similar changes in student application behavior in both year $t=-1$ and the focal year $t=0$.

Figure A.3 presents five panels plotting the coefficients of the main outcomes. Panel A shows the relationship between a one unit increase in PYS and the CYS in periods t before and after the focal year $t=0$. The coefficient in $t=-1$ is mechanically 1 since the CYS in this period is the PYS in the following period. In years 0-3, we see that the CYS remains high, averaging approximately 0.7 units higher, indicating that the evolution of other student outcomes may be affected by the continued higher test score statistics. Also evident, mean reversion from the change to the observable PYS in period 0 is not an important feature of the events we study.

Turning towards student application behavior, we focus first on outcomes for aggregate applicant characteristics of the program: the log applicant average ATAR score (Panel B) and log number of applicants (Panel C). Crucially, prior to the revelation of the updated PYS information in period $t=0$, we detect no statistically significant relationship between increased PYS and either the average applicant ATAR or number of applicants, supporting our parallel trends assumption. Notably, the small and insignificant coefficient in period -1 suggests no detectable structural change in program quality or attractiveness coinciding with the prior-year change in CYS, alleviating concerns about endogenous program improvements driving the observed results. An increase in the CYS is uncorrelated with average applicant ATAR score and log number of applicants. Consequently, we interpret the change in the CYS/PYS occuring from a mix of reductions in program capacity and variation in higher moments of the distribution of the application pool.

We find that an increase of one point in the observable PYS leads to economically significant changes to the applicant distribution. In period 0, the average student test score increases by over 0.2%. Simultaneously, the total number of applicants declines by more than 4%. Both effects are highly statistically significant and are relatively stable or increase in magnitude for the next 3 periods in the event window. The observed decrease in total applicants indicates that an improvement in observable peer ability discourages more applicants than it attracts. The combination of fewer overall applicants and higher average applicant scores implies a compositional shift driven by differential peer ability sensitivity: higher-scoring students are relatively more attracted or less discouraged by improvements in observable peer ability compared to lower-scoring peers.

With relative peer preferences, one natural hypothesis is that students will respond asymmetrically if they are above or below the ability of most of their peers. Panels D and E of Figure A.3 visualize this potential heterogeneous response by studying the number of applicants with ATAR scores above and below the PYS in the reference period ($t=-2$), before changes in peer ability

occur. Consistent with the PYS being unrelated to past program differences, we find no evidence of differential effects prior to period 0 for both outcomes. Starting in period 0, we find a large decrease in the number of applicants with ATAR scores below the past PYS value. Table A.3 reports the average of the post-period coefficients, showing reductions of about 7% of applicants in this group.[15] For students with scores above the reference PYS, we find no statistically significant change. Furthermore, we can clearly reject the hypothesis that effects for higher-scoring students are similar in magnitude to those for students below the PYS. Consequently, preferences towards higher ability peers appear to be heterogeneous based on the student's own relative ability. These relative peer preferences align closely with the findings presented in Section V.C.3.[16]

Our main findings are robust to various alternative modeling choices and sample definitions. While our preferred estimates compare programs whose PYS scores differ by no more than one point over the preceding three years, Figure A.4 shows that alternative similarity criteria—such as stricter comparisons (within one point over four years) or more relaxed comparisons (within two points over two years)—yield qualitatively and quantitatively similar impacts on student demand and characteristics. Additionally, Figure A.5 shows our findings remain unchanged when redefining the number of applicants, shifting from counting any student who ranks a program anywhere on their ROL to counting only students who place the program among their top $x \in \{1,...,7\}$ choices. All these alternative definitions consistently yield similar results.

Overall, this panel-based event-study approach provides additional, robust evidence supporting the existence of relative peer preferences. Students, particularly those below the typically observed ability level, are less likely to apply to a program when observable peer ability increases. Our findings are inconsistent with either structural program quality differences or strategic application behavior driving these results. These findings align closely with those from our main analysis in Section V.C.3, confirming that relative peer ability significantly influences applicant decisions.

F Details for Estimating Peer Preferences

This appendix provides detailed intuition and formal derivations underlying the decomposition of deterministic tastes into group-by-rank fixed effects and residuals that underpin the identifying assumption discussed in Section V.C.3.

We begin by restating the two-stage utility model. Let students θ belong to mutually exclusive

[15] A small number of observations are dropped as a result of logging the outcome variable. The results are quantitatively and qualitatively the same if we use the log of the outcome plus one.

[16] One limitation of this analysis is that while we identify relative peer preferences, we do not observe their exact functional form. Section V.C.3 explicitly estimates these preferences.

groups g defined by $g(\theta)$. Stage-0 utility, prior to students learning their ATAR scores, is given by:

$$U^0_{\theta c} = \phi^0_{g(\theta)c} + \sum_{j=1}^{J} \alpha_j (PYS_c - B^0_{g(\theta)})^j + \varepsilon^0_{\theta c}.$$

Here, $\phi^0_{g(\theta)c}$ represents baseline deterministic preferences for non-peer amenities at program c, identical within groups g, and $B^0_{g(\theta)}$ is the common prior belief about students' scores within each group. Between stage 0 and stage 1, students update their beliefs perfectly upon observing their actual score S_θ, and deterministic tastes may experience systematic drift $d_{g(\theta)r_\theta(c)}$ depending on group g and the program's stage-0 rank position r. Thus, stage-1 utility becomes:

$$U^1_{\theta c} = \phi^0_{g(\theta)c} + d_{g(\theta)r_\theta(c)} + \sum_{j=1}^{J} \alpha_j (PYS_c - S_\theta)^j + \varepsilon^1_{\theta c}.$$

To explicitly isolate deterministic taste heterogeneity, we rewrite the deterministic component of stage-1 utility into two intuitive parts:

$$\psi_{g(\theta)r_\theta(c)} = \mathbb{E}\left[\tilde{\phi}^0_{g(\theta)c} \,|\, g,r\right] + d_{g(\theta)r_\theta(c)}$$
$$\zeta_{\theta c} = \tilde{\phi}^0_{g(\theta)c} - \mathbb{E}\left[\tilde{\phi}^0_{g(\theta)c} \,|\, g,r\right]$$

where $\tilde{\phi}^0_{g(\theta)c} = \phi^0_{g(\theta)c} + \sum_{j=1}^{J} \alpha_j \left(PYS_c - B^0_g\right)^j$. Thus, stage-1 utility simplifies to:

$$U^1_{\theta c} = \psi_{g(\theta)r_\theta(c)} + \zeta_{\theta c} + \sum_{j=1}^{J} \alpha_j (PYS_c - S_\theta)^j + \varepsilon^1_{\theta c}.$$

The term $\psi_{g(\theta)r_\theta(c)}$ captures all deterministic taste heterogeneity that is common within each group-by-rank cell. It bundles three elements: (i) baseline non-peer amenities, $\phi^0_{g(\theta)c}$; (ii) the baseline peer-gap term $\sum_j \alpha_j (PYS_c - B^0_{g(\theta)})^j$ evaluated at the group's prior belief; and (iii) the common within-cell drift d_{gr} that captures systematic re-evaluation between stages. The residual, $\zeta_{\theta c}$, therefore, reflects any remaining within-cell taste differences across programs.

To understand the content of the residual $\zeta_{\theta c}$, consider its explicit decomposition:

$$\zeta_{\theta c} = (\tilde{\phi}^0_{g(\theta)c} - \mathbb{E}[\tilde{\phi}^0_{g(\theta)c} \,|\, g,r]) + \sum_{j=1}^{J} \alpha_j \left[(PYS_c - B^0_{g(\theta)})^j - \mathbb{E}\left[(PYS - B^0_{g(\theta)})^j \,|\, g,r \right] \right].$$

The first part reflects program-specific amenity differences that remain within a given group-by-rank cell. Such differences are expected to be small, since conditioning on group and pre-ROL rank already compares programs perceived as similar by applicants. The second part captures deviations in the peer-gap polynomial from the cell mean, also expected to be minor, as similarly-ranked programs typically have a similar PYS. The common drift is fully accounted for by the group-by-rank mean preference terms.

To illustrate why residual deviations are plausibly negligible in practice, consider the limiting

scenario where groups are defined based on each student's complete stage-0 ROL. Under this complete conditioning, each resulting group-by-rank cell includes only programs ranked identically by all students in the group. Thus, by construction, all deterministic taste heterogeneity is fully absorbed by cell fixed effects, causing $\zeta_{\theta c}$ to collapse exactly to zero. In practice, we use broader group definitions to maintain empirical feasibility. Residual deviations $\zeta_{\theta c}$ therefore arise due solely to the coarser conditioning. Empirically, we assess the practical relevance of these residuals in Section V.C.3 by progressively refining group definitions. Across specifications which vary the number of groups from 1 to over 180 (which are allowed to vary with eight stage-0 ranks), we verify that our peer-preference estimates are remarkably robust.

G Additional Tables and Figures

Table A.1: Student and ROL Summary Statistics

	Mean	SD	P25	P50	P75
Pre- and Post-ROL Sample (2010-2016, N = 173,694 students)					
Students					
Student ATAR Score	72.4	18.5	60.0	75.0	88.0
# of Programs Ranked	5.9	2.1	4.0	6.0	8.0
All Programs in a Student's ROL					
Avg. PYS	78.9	9.2	72.1	78.4	85.9
Avg. Pre-ATAR PYS	79.3	9.4	72.6	79.1	86.5
Avg. PYS/Score Gap	6.5	14.3	-3.2	2.7	14.0
Avg. Pre-ATAR PYS/Score Gap	6.9	14.8	-3.4	3.6	15.1
Only Top-Ranked Program in a Student's ROL					
PYS	80.4	11.4	71.8	80.1	90.2
Pre-ATAR PYS	81.2	12.0	72.6	81.0	91.0
Score Gap	7.7	13.9	-1.0	4.6	14.1
Pre-ATAR Score Gap	9.0	15.3	-0.8	6.5	17.3
Post-ROL Sample (2003-2016, N = 285,598 students)					
Students					
Student ATAR Score	73.2	18.0	61.0	76.0	88.0
# of Programs Ranked	5.6	1.9	4.0	6.0	7.0
All Programs in a Student's ROL					
Avg. PYS	78.8	8.9	72.2	78.5	85.4
Avg. Score Gap	5.6	13.9	-3.9	1.8	13.0
Only Top-Ranked Program in a Student's ROL					
PYS	80.3	11.3	70.6	80.0	90.0
Score Gap	7.4	13.9	-1.0	4.1	14.0

This table displays summary statistics on students (ATAR score, number of programs ranked), all programs in student ROLs (PYS and score gaps), and the top-ranked program in student ROLs (PYS and score gaps). The pre-and-post ROL sample is restricted to students who rank at most 8 programs on the pre-ROL, as discussed in Appendix E. The Post-ROL Sample includes all students.

Table A.2: Rank-order logit estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Score Diff	0.331***	-0.0916***	-0.0997***	-0.106***	-0.165***	-0.106***	-0.0803***	-0.0920***
	(0.00536)	(0.00853)	(0.00873)	(0.00902)	(0.0115)	(0.00883)	(0.00960)	(0.0101)
Score Diff^2	-0.0967***	-0.0953***	-0.0953***	-0.0996***	-0.0981***	-0.0988***	-0.103***	-0.0988***
	(0.00250)	(0.00304)	(0.00304)	(0.00333)	(0.00332)	(0.00312)	(0.00338)	(0.00414)
Score Diff^3	0.00789***	0.0161***	0.0161***	0.0172***	0.0168***	0.0170***	0.0174***	0.0168***
	(0.000501)	(0.000584)	(0.000584)	(0.000641)	(0.000624)	(0.000602)	(0.000656)	(0.000785)
Score Diff $\times$ Old								0.00247
								(0.00799)
Score Diff^2 $\times$ Old								0.00536
								(0.00423)
Score Diff^3 $\times$ Old								-0.00101
								(0.000810)
Pre-rank	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Share pre-rank	No	No	Yes	No	No	No	No	No
Pre-rank $\times$ ROL length	No	No	No	Yes	No	No	No	No
Pre-rank $\times$ University	No	No	No	No	Yes	No	No	No
Pre-rank $\times$ Field	No	No	No	No	No	Yes	No	No
Pre-rank $\times$ Top Univ-Field	No	No	No	No	No	No	Yes	No
Pre-rank $\times$ Old program	No	No	No	No	No	No	No	Yes
N	275535	275535	275535	275535	259051	266307	260913	275535

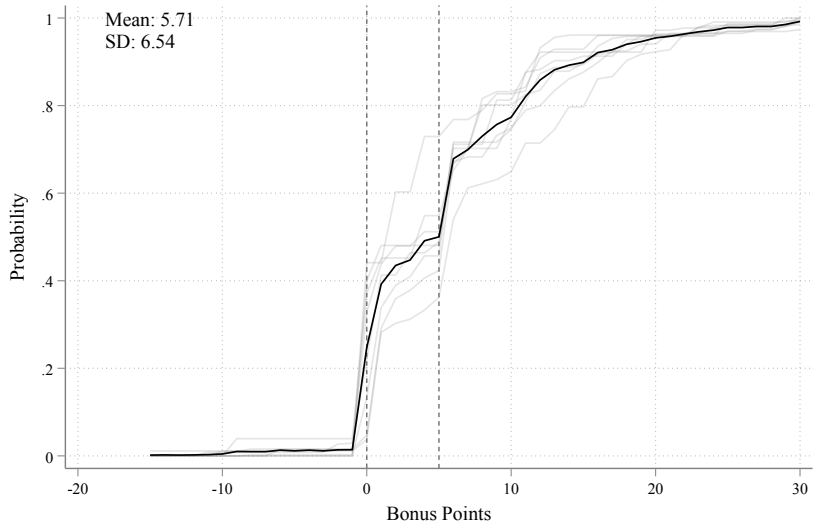
This table displays coefficients of the score difference between the student and program estimated from a rank-order logit model on final student choices. The score difference is measured in units of 10 for readability. Column (1) includes no further controls. Column (2) includes dummies for this program's pre-ROL rank. Column (3) further includes controls for the share of students who rank this program in their pre-ROL for each rank. Column (4) interacts pre-ROL rank dummies with ROL length. Column (5) interacts pre-ROL rank dummies with university. Column (6) interacts pre-ROL rank dummies with field of study. Column (7) interacts pre-ROL rank dummies with the university-field of the individuals pre-rank ROL top choice. Column (8) augments the specification from column (2) to include heterogeneous score difference and heterogeneous pre-rank coefficients by a measure of program age (i.e., "old" programs are those that are open in all previous 5 years up to the focal year). The sample is restricted to individuals who do not add or drop programs after the revelation of student ATAR scores. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table A.3: Effects of observable PYS

	(1) Ln(Average ATAR)	(2) Ln(Number of Applicants)	(3) Ln(Number of Applicants below $PYS_{t=-2}$)	(4) Ln(Number of Applicants above $PYS_{t=-2}$)	(5) CYS
PYS	0.00380***	-0.0372***	-0.0694***	-0.00454	0.762***
	(0.000971)	(0.00948)	(0.0129)	(0.00943)	(0.0868)
N	21026	21026	21000	20800	20949

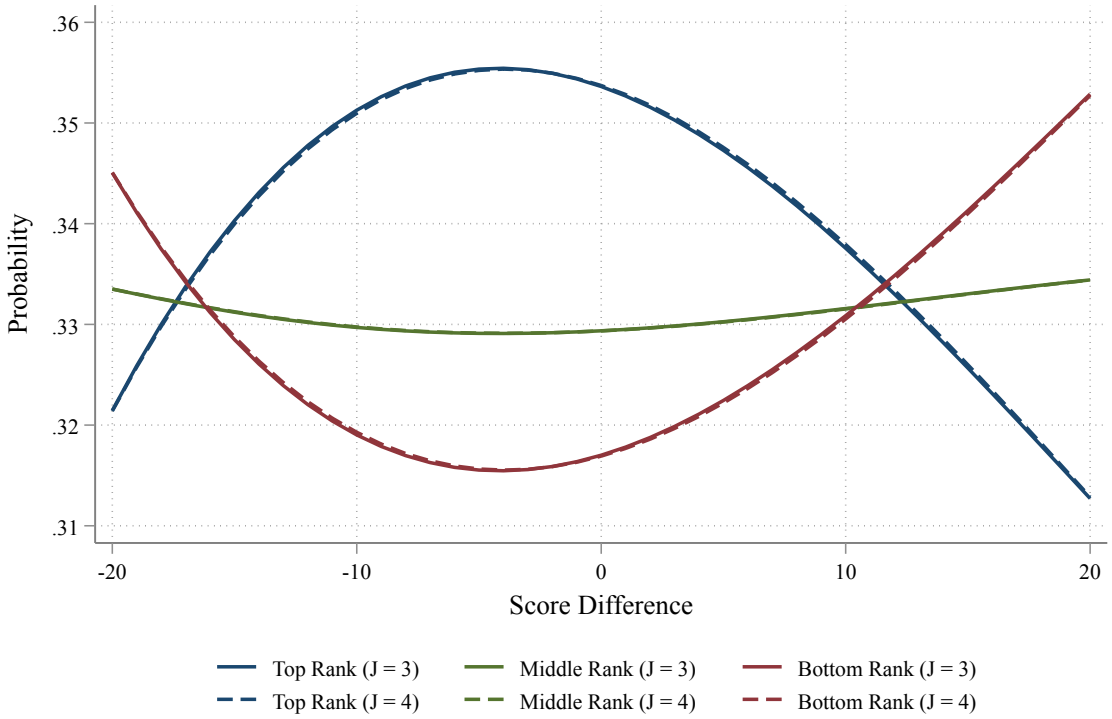
This table displays averages of the post-period event-study coefficients of the effect of the Previous Year Statistic (PYS) on outcomes for five years before and after the event period. Figure A.3 displays regression estimates for each period. Regressions are estimated according to Equation A.16 and include group-by-time and program-by-treatment year fixed effects. Data is at the program-year-relative year level. Coefficients are estimated relative to $t=-2$. Standard errors in parentheses calculated using the Delta method and clustered at the program level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Figure A.1: Empirical CDF of bonus points



This figure shows the empirical CDF of bonus points. The black line plots the CDF estimated across all eight ranks, while the gray lines show the CDF estimated using the first to eighth ranks, respectively. Sample includes all applicants eligible for the main round in 2015 and 2016. Individual by program observations are eligible if a student was not admitted to a program higher on the rank-order list and the student restrictions detailed in Appendix C.

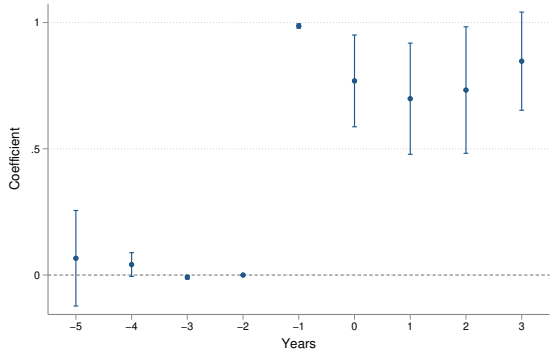
Figure A.2: Updating program choices on relative scores, by polynomial degree



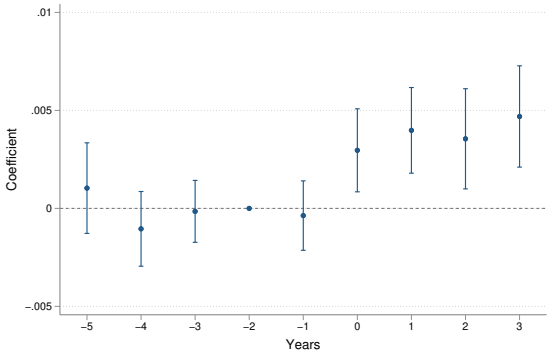
This figure plots predicted probabilities from a rank-order logit model for two different specifications of polynomial degrees of score difference, $J=3$ and $J=4$. Each line represents the probabilities for a program to be rank 1, rank 2, or rank 3 by the score difference between the program PYS and the student ATAR. We compute predicted probabilities for being ranked first, second, or third by evaluating the estimated rank-order logit on a grid of covariate values $x \in [-20,20]$. The rank-order logit estimates come from column (2) in Table A.2.

Figure A.3: Event-study estimates of effect of observable PYS

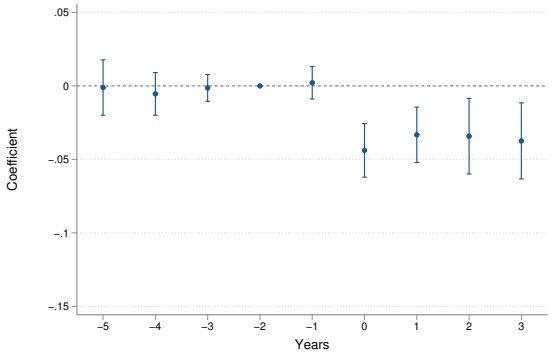
(a) Current Year Score (CYS)



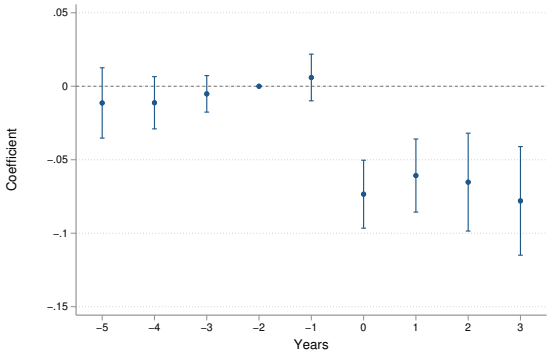
(b) Ln(Average ATAR)



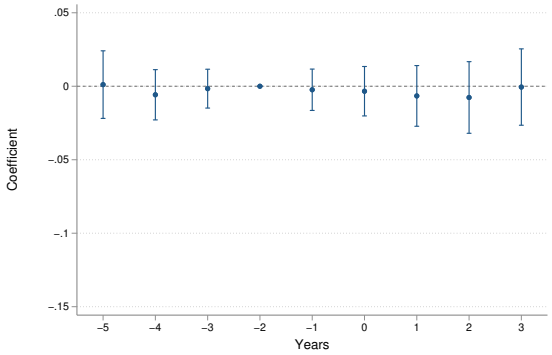
(c) Ln(Number of applicants)



(d) Ln(Number of applicants below $PYS_{t=-2}$)

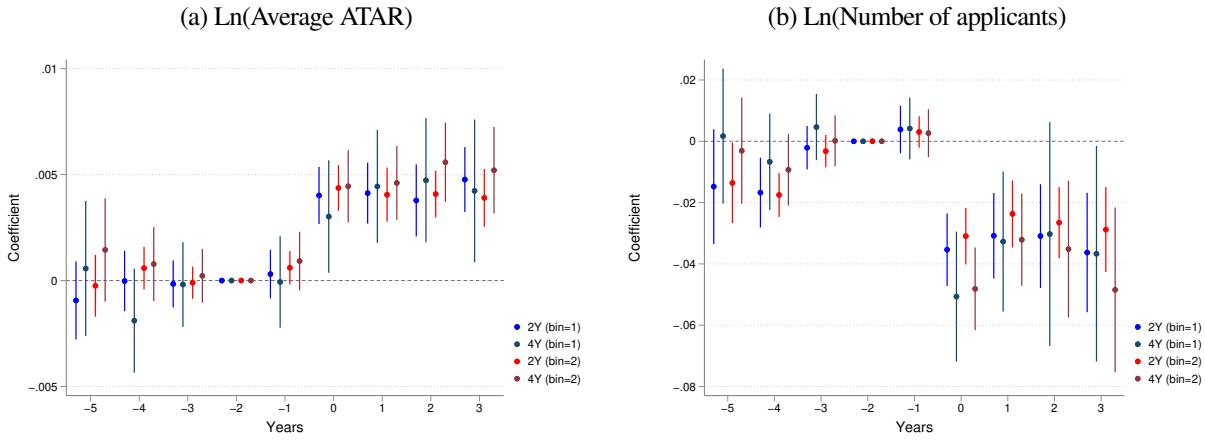


(e) Ln(Number of applicants above $PYS_{t=-2}$)



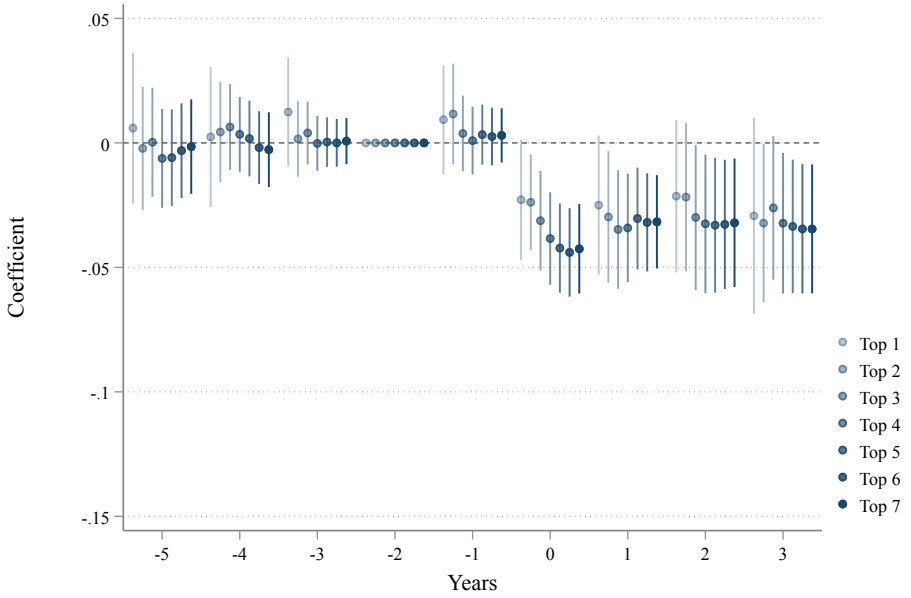
This figure displays regression estimates of the effect of a change in a program's Previous Year Statistic (PYS) on outcomes for five years before and three years after the event period. Regressions are estimated according to Equation A.16 and include group-by-time and program-by-year fixed effects. Coefficients are estimates relative to $t=-2$. Figures show point estimates and 95% confidence intervals clustered at the program level.

Figure A.4: Robustness by different grouping criteria



This figure displays regression estimates of the effect of a change in a program's Previous Year Statistic (PYS) on outcomes for five years before and three years after the event period. Regressions are estimated according to Equation A.16 and include group-by-time and program-by-year fixed effects. Program groups vary by number of previous years and size of score bin, indicated in the legend. Coefficients are estimates relative to $t = -2$. Figures show point estimates and 95% confidence intervals clustered at the program level.

Figure A.5: Robustness in number of applicants by rank



This figure displays regression estimates of the effect of a change in a program's Previous Year Statistic (PYS) on the number of applicants (by rank) for five years before and three years after the event period. The outcome of number of applicants is measured variously based on the top x ranks, where $x \in [1,7]$. Regressions are estimated according to Equation A.16 and include group-by-time and program-by-year fixed effects. Coefficients are estimates relative to $t = -2$. Figures show point estimates and 95% confidence intervals clustered at the program level.