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

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

November 28, 2023

Abstract

School-choice clearinghouses often advise students to "rank their true preferences" despite not allowing students to express preferences over peers. We evaluate the consequences of doing so. Empirically, we find students have preferences over relative peer ability in the college admissions market in New South Wales, Australia. Theoretically, we show stable matchings exist even with peer preferences under mild conditions, but finding one via one-shot mechanisms is unlikely. The status quo procedure frequently employed by clearinghouses is to inform applicants about the assignment of students in the previous cohort, inducing a tâtonnement process which potentially provides useful information about likely peers in the current cohort. We theoretically argue this process likely leads to an unstable outcome, and we find instability in our empirical setting. We propose a mechanism that yields stability and incentivizes truthful reporting in the presence of peer preferences.

*For helpful comments we thank Atila Abdulkadiroglu, Mohammad Akbarpour, Eduardo Azevedo, Yan Chen, Aram Grigoryan, 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 Boston College, Brown, CMU/Pitt, Penn, MSR New England, ITAM, Stanford, UCSB, University of Tokyo Market Design Center, Wharton, Matching in Practice Workshop '21, EAMMO '21, and 2021 NBER Summer Institute (Education). We are grateful to Camilla Adams, Clemens Lehner, Sergio Nascimento, and Joanna Tasmin for excellent research assistance.

This paper replaces and subsumes an earlier paper "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

‡Brown University, 57 Waterman Street, Providence, RI 02906. Email: ricardo_fonseca@brown.edu

§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 now commonly used to allocate seats at schools and colleges in at least 46 countries (Neilson, 2019). Creating a stable matching–one in which no agent wants to "block" by deviating with a willing partner (or remain unmatched)–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 pre-determined 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 assuming that student preferences do not depend on the characteristics of their peers.

In this paper, we seek to answer four questions, which we believe must be investigated simultaneously to understand the role of peer preferences in present-day school choice markets: (i) Do students have peer preferences? (ii) Do stable matchings exist when students have peer preferences? (iii) Do "status quo" matching markets deliver stable matchings and what are the consequences if not? (iv) Do better mechanisms exist in the face of peer preferences?

We study these questions theoretically and empirically. First, using data from the centralized matching market for college admissions in New South Wales (NSW), Australia's largest state, we show that students' ordinal preference rankings over programs are affected by information about potential peers' ability. Specifically, students on average prefer not to match with a program where they are near the bottom of the ability distribution. Second, we develop a theoretical large market matching model where students have arbitrary peer preferences, and we show that a stable matching exists under mild conditions. Third, we use our model to show that the status-quo matching procedure is likely to lead to instability. This theoretical analysis generates testable implications of stability. Applying this test to the data, we find that the NSW market does not find a stable matching. Finally, we propose a new matching mechanism which we theoretically show find or approximates a stable matching.

Our analysis is informed by the centralized market for admissions to college-subject "programs" in NSW. Each student submits a Rank Order List (ROL) over programs. An important consideration for students when forming ROLs is the information about programs available (see, e.g., Agte et al., 2023; Larroucau and Rios, 2020b; Allende et al., 2019; Cohodes et al., 2022; Hastings and Weinstein, 2008). As the set of students attending each program changes from year to year, information about potential peers may be particularly informative if students directly care about peers. Reflecting a common practice around the world, the NSW clearinghouse, by law,

prominently displays information about the student body from the previous cohort and advises students to use this information "as a guide when deciding on ... preferences."[1] Specifically, this information takes the form of a single summary statistic—corresponding generally to the median—of the distribution of standardized test scores on the Australian Tertiary Admissions Rank (ATAR) exam, which we call the Previous Year Statistic (PYS). Using student ROLs, program rankings over students, and program capacities, NSW creates each cohort's matching through the (student-proposing) deferred acceptance algorithm of Gale and Shapley (1962). Deferred acceptance is well known to generate a stable matching in the absence of peer preferences.

We begin our analysis by asking: *Do students have preferences over peer ability?* We use two causal identification strategies to show this that they do. Our first research design exploits panel variation in the PYS to test how changes to available peer information affect student ROLs. Comparing programs with a similar evolution of past peer ability in an event-study framework, we find an increase to a single year's PYS leads to a decrease in program popularity overall, driven entirely by students with ATAR scores below the PYS. Importantly, the timing of these effects is neither consistent with perceived changes in program quality, nor based on strategic application behavior.

Our second research design investigates how relative peer ability affects preferences for the same student. To do so, we leverage a novel feature of the NSW market—students submit ROLs prior to learning their own ATAR score (but after observing program PYSs) and can make adjustments after learning it. Using multiple ROLs for the same student allows us to relax typical truthtelling assumptions made in the literature, by focusing on whether each student "switches" the order of two programs from their initial to final ROL. We find an increasing relationship between the likelihood a student demotes a program on her ROL and the amount by which the PYS exceeds her ATAR score and no relationship for programs below her ATAR; this asymmetric relative peer preference is entirely consistent with the previous research design. Importantly, the empirical findings from these two designs are not consistent with other candidate explanations studied in the market design literature.[2]

We next ask: *Does a stable matching exist in the presence of peer preferences?* To study this

[1] Providing information on the previous cohort's matching as a guide for current applicants is common in both decentralized and centralized higher education markets. For example, *U.S. News and World Report* annually publishes standardized test scores of the entering class from the previous year at U.S. universities.

[2] Our definition of peer preference is broad, and we do not seek to distinguish between "direct" preferences over peers, or "indirect" factors that are frequently unmodeled in matching papers, such as career concerns or uncertainty about future financial opportunities. For example, a student may plausibly prefer not to attend a program where she is overmatched by her peers if she is worried that she will not receive enough attention from her professors. The results of our paper apply to any setting in which the inclusion of the peer ability distribution into students' utility functions captures students' ordinal preferences at the time of application.

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 the highest-ability peers, the lowest-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 informs how to account for 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).[3] 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 our previous evidence that both peer preferences and stable matchings exist, we ask: *Do "status quo" matching markets deliver stable matchings?* We first show theoretically that canonical static mechanisms–including deferred acceptance–are unlikely to result in a stable matching without sufficiently 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 a ROL to a centralized matchmaker who delivers 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."[4] Importantly and unfortunately, we show that the status quo procedure may produce a

[3] 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.

[4] Best responding to the previous period's distribution is analogous to the Cournot updating procedure in exchange

matching that is far from stable in all time periods for nearly any functional form of peer preferences. This additionally suggests that even careful measurement of the functional form of peer preferences is insufficient to predict ex-ante whether any particular market will yield a stable matching.

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.

Since quantifying the degree of instability in the market is crucial to understanding how well the market functions, we calculate a lower bound on the fraction of students involved in blocking pairs i.e. the share of students assigned to the "wrong" program. To do so, we combine a research design using variation in observable peer quality characteristics across programs over time with a theoretically-motivated formula of blocking pairs. Across the years in our data sample, we estimate this lower bound is 3% of students, with disadvantaged students experiencing a significantly higher share.

We believe our estimates of the lower bound on the level of instability in NSW have important policy implications. To see why, consider the National Residency Matching Program (NRMP), which assigns new medical school graduates to U.S. hospitals for residency. The presence of couples, who prefer to be matched in the same geographic region as one another, potentially lead to instability under the previous matching procedure. The NRMP was redesigned in the late 1990s primarily due to the presence of couples, and now finds a stable matching whenever one exists (Roth and Peranson, 1999). In the decade leading up to the redesign, an average of 4% of new doctors were members of couples (see Table 1 of Roth and Peranson (1999)). The peer preferences differ in NSW and the NRMP ("anonymous" versus "couple" preferences), which calls for a different approach to ensure stability. Nevertheless, the NSW is larger than the American medical market, and our finding that a lower bound on the share of NSW students in blocking pairs is similar to the *total* share of coupled doctors in the NRMP suggests that the impact of peer preferences is a large concern in NSW.

We further link our estimates of blocking pairs to an observable impact of instability: attrition. Using a panel fixed effects approach, we find that an increase in our estimated share of students at a program involved in blocking pairs causes a reduction in the completion rate among commencing students. Specifically, a one standard deviation increasing in the share of students estimated to be

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).

in blocking pairs translates into a 0.06 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. These results show that 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 account for peer preferences?* We propose a mechanism that induces a tâtonnement process *within* each cohort of students, and does not require detailed information about the "functional form" of peer preferences.[5] Moreover, it has desirable incentive properties, suggesting that it may not disadvantage unsophisticated students (Pathak and Sönmez, 2008; Song et al., 2020). Crucially, we show that unlike the status quo process, the tâtonnement process induced in our mechanism never cycles as in Scarf (1960),[6] meaning that our mechanism always generates a (approximately) stable matching.

Related Literature

Recent papers find empirical evidence that would-be peers affect student preferences over programs (Rothstein, 2006; Beuermann et al., 2019; Allende, 2020; Che et al., 2022), and matter above and beyond value-added measures (Abdulkadiroğlu et al., 2020; Beuermann and Jackson, 2019). These papers find that (parents of) students prefer, on average, programs where peers have higher ability. Our analysis corroborates this finding, but importantly differs in that we study how a student's *relative* ability affects the desirability of a program. Previous papers in this literature generally do not consider relative peer preferences as their models assume a constant effect of peer scores on the preferences of all students.[7] Our analysis reveals a more nuanced "functional form" of peer preferences in the NSW market than has been presented in the existing literature.[8]

[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. We therefore refrain from considering alternative mechanisms that require students to only submit ROLs over programs.

[6]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.

[7]Beuermann et al. (2019) 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 big-fish preferences.

[8]The "functional form" and root causes of peer preferences potentially depend on a number of factors that vary across markets (for a similar thesis, see Sacerdote, 2014). Cultural norms, including the so-called "tall poppy" syndrome

We identify that students value programs enrolling higher ability peers on average, but that this is tempered by concerns over their *relative* ability in the class (as in Frank, 1985; Azmat and Iriberri, 2010; Tincani, 2018). This peer utility effect is asymmetric, and similar to the analogous function in Card et al. (2012); students face a utility loss only if their ATAR score is below the PYS, and the utility loss is increasing in the difference.

Our finding of preferences over relative peer ability accords with the well-documented "big-fish-little-pond effect," wherein a student being lower in the ability distribution leads to psychic costs (Marsh et al., 2008; Pop-Eleches and Urquiola, 2013) and declines in achievement (Carrell et al., 2013).[9] These preferences are also reflected in the common practice of "redshirting" kindergartners; parents delay entry for an additional year so their child can be among the oldest and most cognitively developed in the class (Dhuey et al., 2019).

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 of couples in matching markets, as previously discussed (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. Recent research also studies stability with peer preferences in large, finite, matching markets (Greinecker and Kah, 2021), and our model is most similar to that in a contemporaneous paper (Leshno, 2022), which also studies 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. In Appendix C

in Australia wherein students react negatively to those who overachieve relative to peers (see, e.g. Feather, 1989), possibly lead Australians to avoid programs they perceive their peers to be "overachieving." Student autaonomy may also play a role. Pop-Eleches and Urquiola (2013) and Ainsworth et al. (2020) study the same high school admissions market and find that while students at the bottom of the ability distribution at a program suffer from psychic costs, their parents nevertheless prefer to send their children to programs with higher-achieving students. It stands to reason that the direction of peer preferences could differ in college admissions markets, due to increased student autonomy.

[9]At the primary and secondary school levels, a series of recent papers show that a student's ordinal "ability" ranking within her school and class has a negative effect on educational achievement; that is, students perform worse when they have higher achieving peers (see Dobbie and Fryer Jr. (2014); Elsner and Isphording (2017); Elsner et al. (2018); Murphy and Weinhardt (2020); Yu (2020); Zárate (2019); Carrasco-Novoa et al. (2021)). Abdulkadiroğlu et al. (2014) do not find a large effect of peer ability on student performance.

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.

The remainder of the paper is structured as follows: Section II discusses the NSW Tertiary Education System and provides evidence of peer preferences; Section III defines peer preferences in a matching environment and shows the existence of a stable matching; Section IV theoretically and empirically discusses instability in status-quo markets that do not account for peer preferences; Section V presents a new mechanism that finds a stable matching in the presence of peer preferences; Section VI concludes. Omitted proofs and additional examples are relegated to the Appendix.

II Empirical Evidence of Peer Preferences

In this section, we describe details of the New South Wales college admissions system and our administrative data from this market. We then discuss how we leverage this context and data to implement two causal research designs to estimate the presence of preferences over the relative abilities of peer classmates and show results. Throughout, we let θ represent a generic student, and c a generic program.

II.A The New South Wales Tertiary Education Admissions System

We study college admissions in NSW (and the Australian Capital Territory) from 2003 to 2016.[10] Students apply for admission at the university-field level (for example, Economics at University of Sydney) through a centralized clearinghouse. We refer to these 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 are not aware of

[10] 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).

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 a ROL of up to nine programs to the United Admissions Centre (UAC), the centralized clearinghouse which processes applications to all major universities in NSW.[11] 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 learning their own ATAR scores, but are able to costlessly change their ROLs after learning their ATAR score. Students are incentivized to submit initial ROLs early in the application process, as fees for stating initial ROLs increase over time.

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).[12] Program rankings over students are determined by the sum of a student's ATAR score and program-student specific "bonus" points, which are awarded at the discretion of the program. Students can receive up to 10 bonus points at each program. Importantly, the bonus points awarded to each student typically differ across programs, are not known in advance, and the criteria for bonus points are typically not known to students.[13] Therefore, bonus points serve as a significant source of admissions uncertainty.

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*

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

[12] 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.

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

order of preference up to the maximum of nine course preferences."[14]

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. Going forward, we will refer to the clearly-in statistic for the cohort admitted in the previous year as the *Previous Year's Statistic (PYS)* for a particular program, and the clearly-in statistic for the current year as the *Current Year's Statistic (CYS)*. While the CYS would equal the minimum score required for entry to a program without the presence of bonus points, in practice the CYS roughly corresponds to the median ATAR score of matriculating students (Bagshaw and Ting, 2016). Students are therefore typically aware of the possibility of admission to a program even if their ATAR score is below the CYS.[15]

When creating their ROLs, students do not know the CYS at any program. However, they can consult programs' PYSs as a guide–this information is made prominently available, by law, on the clearinghouse website. Between 2003 and 2017–which contains our window of analysis–the only information about peer ability in the previous cohort revealed to applicants is the PYS. 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."[16]

II.B Data

We use data from the UAC clearinghouse for our analysis. Our data contain the universe of applications from graduating high schoolers processed by UAC for 2003-2016. We identify each student via a unique student identification number. Over this time 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, and observe the program field of study.[17] For a subset of years

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

[15] 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.

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

[17] 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

(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 do not have information about socioeconomic background or bonus points at the application level.

We do not observe the final assigned program of each student. When required for our analysis, we simulate the matching using student ROLs, program CYSs, and researcher-assigned bonus points at the student-program level. We assign bonus points by drawing from various, plausible distributions. These distributions are, for each student-program pair: (1) bonus = 3, (2) bonus = 3 plus a randomly assigned integer 0-7 from a uniform distribution, and (3) bonus = 7. These regimes generate admitted shares of students to have $ATAR \geq CYS$ of 67.5%, 50.9%, and 49.6%, respectively. Since across programs, roughly half of all enrolling students have ATAR scores below the CYS of their program (Bagshaw and Ting, 2016), the latter two regimes appear closest to the true empirical distribution. We take the final one as our preferred distribution given it is closest, and present results from all bonus regimes for robustness.

Finally, we link estimated instability measures to attrition rates for students commencing in a specific year attending a university assigned to one of 12 fields of study. The data are calculated by the Australian Government Department of Education.[18] To link these records, we aggregate our instability measures to the year-university-broad field of study. The broad field of study is known for all programs in our application records.

II.C Empirical analysis of peer preferences

II.C.1 Parameter of interest

In this section, we seek to identify applicant preferences over peer quality, as measured by the program PYS. These peer preferences are broadly defined; they could reflect direct preferences over the peer population itself, such as for social connections or study partners, or more indirectly through the programs, such as in zero-sum grading or competition for recommendations. This def-

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.

[18]The data can be viewed here: https://app.powerbi.com/view?r=eyJrIjoiNTA4MTZjZmMtZjRjNS00NzcxLWEzZTktODZmNDZkNGEwM2Y4IiwidCI6ImRkMGNmZDE1LTQ1NTgtNGIxMi04YmFkLWVhMjY5ODRmYzQxNyJ9.

inition of peer preferences tracks closely to the assumptions employed in Section III. Not included in our parameter of interest would be a preference for peers as a signal of university characteristics themselves, such as fixed teacher quality or prestige, or student-program "fit" (Rothstein and Yoon, 2008), which we address later.

II.C.2 Motivating evidence and empirical challenges

Table 1 displays summary statistics of the data, including student ATAR scores, program PYSs for those listed on student ROLs, and "score gaps," defined as the difference between the PYS of a ranked program and the student's ATAR, for the for the sample containing post-ROLs only (2003-2016) and for the sample containing pre- and post-ROLs (2010-2016). Across the samples, the data are quite similar and contain a few notable features. First, students include programs in their ROLs with PYS scores on average higher than their own; for example, we observe an average score gap of 5.8 in the post-ROL sample. Second, students tend to show even more desire for "reach"-type programs—the top choices feature a higher average score gap of 7.5, and the vast majority of students list a program with a positive score gap. It therefore does not appear that students are afraid to include reach programs high in their ROLs, providing one indication students understand the incentive properties of the deferred acceptance mechanism.

The simple statistics above, while consistent with students preferring programs with peers who are higher performing than themselves, are unlikely to reflect true preference information for a number of reasons. First, students may choose programs for other reasons correlated to peer quality, such as prestige, institutional quality or other omitted variables, rather than having preferences for relative peer quality itself. Second, stated preferences on ROLs may not reflect true preferences over programs and peers. Third, analyzing student choices in relation to peer quality generates concerns whether these relationships may reflect mechanical effects as a result of econometric issues, such as the "reflection problem"(Manski, 1993).

In the rest of this section, we describe two distinct research designs to estimate preferences over relative peer quality that address the issues above. The designs utilize different assumptions over student truth-telling to recover preferences from ROLs and employ different identifying variation in relative student-program comparisons.

II.C.3 Panel-based analysis

Students are required by law to be shown a program's PYS when making application decisions. Our first research design employs variation in this information in a panel-based empirical design that compares total and relative demand using counterfactual comparisons between programs matched

Table 1: Student and ROL Summary Statistics

	Mean	SD	P25	P50	P75
Pre- and Post-ROL Sample (2010-2016, N = 116,341 students)					
Students					
Student ATAR Score	72.8	18.5	60.0	76.0	88.0
# of Programs Ranked	6.1	2.0	5.0	6.0	8.0
All Programs in a Student's ROL					
Avg. PYS	79.1	9.3	72.3	78.9	86.3
Avg. Pre-ATAR PYS	79.7	9.1	72.8	79.6	86.8
Avg. PYS/Score Gap	6.3	14.0	-3.2	2.6	13.4
Avg. Pre-ATAR PYS/Score Gap	6.8	14.4	-3.3	3.5	14.8
Only Top-Ranked Program in a Student's ROL					
PYS	80.8	11.7	72.4	80.7	91.0
Pre-ATAR PYS	81.7	11.3	74.0	82.0	91.3
Score Gap	7.5	13.7	-1.0	4.4	14.0
Pre-ATAR Score Gap	8.9	14.6	-0.8	6.2	17.0
Post-ROL Sample (2003-2016, N = 491,512 students)					
Students					
Student ATAR Score	73.4	18.1	61.0	76.0	88.0
# of Programs Ranked	7.0	2.2	5.0	8.0	9.0
All Programs in a Student's ROL					
Avg. PYS	79.2	8.8	72.7	78.9	85.8
Avg. Score Gap	5.8	13.8	-3.4	2.0	12.5
Only Top-Ranked Program in a Student's ROL					
PYS	81.0	11.3	72.1	81.0	91.0
Score Gap	8.0	14.0	-0.7	4.9	14.2

This table displays summary statistics on students (ATAR score, 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 Section II.C.3. The Post-ROL Sample includes all students, as discussed in Section II.C.4.

on similar levels and trends in a transparent, event-study design. A simple example illustrates the empirical design. Consider two programs, c and c', which have the same observable-to-the-student information on peer quality, the PYS, for multiple prior years. 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'. How do students respond to this newfound increase in the observable PYS of program c?

Interpreting student responses as preferences requires an important consideration. In keeping with much of the related literature (e.g. Hastings et al., 2009; Abdulkadiroğlu et al., 2017; Luflade, 2019), we restrict our sample of applicants to students who submitted fewer choices than the maximum allowed (i.e., nine) to ensure that observed ROLs reflect ordinal student preferences.[19] A student who prefers weakly fewer than nine programs to her outside option has a weakly dominant strategy to include all such programs and in order of their 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:

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

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 traces out the relationship between PYS_{cy}, the available peer information for program c in a specific year y, and outcomes for relative periods t before and after the change in PYS. As such, the panel is constructed to treat each program-focal year as a new potential event of changing student information and generates relative period structure around these events, in the same way a traditional event-study would around a potential event period.[20]

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 $t=-3$.[21] We do so by defining groups $g=g(c,y)$, mapping

[19] 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.

[20] A two-way fixed effects approach could also be used to estimate the effect of the observable PYS on program application characteristics. We detail this design and results in Section E. Relying instead on more general parallel trends assumptions, we find quantitatively and qualitatively similar results. Our preferred design outlined above is not subject to negative weighting concerns across group cells, and naturally generalizes to heterogeneity in responses across the distribution of students relative to a well-defined benchmark. This latter analysis is precluded when using the fixed effects approach as there is no natural reference period to base the above-or-below-PYS cutoff, and hence is more limited.

[21] We do not match on periods -4 or earlier nor on other outcomes, as such periods and outcomes can be potentially

different programs in the same focal year to a common grouping.[22] The fixed effects, $\delta_{g(c,y),t}$, are relative time by group fixed effects that impose comparisons between programs within group in 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 $\beta_{-2}=0$.[23] 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 us to 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 also the cutoff score 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}. This relates to a natural concern that the changes to peer quality we study may be driven by differences in systematic program-specific quality improvements rather than as-good-as-random perturbations to cutoff scores in the prior year. This type of violation to our exclusion restriction has likely empirical implications: if we thought PYSs were driven by an unobservable-to-the-econometrician program difference, such as an advertised improvement in teachers, we would expect similar patterns in student application behavior in year $t=-1$ as our year of interest $t=0$ for our outcomes of interest. This is because students would not be responding to the PYS in year $t=0$, instead they would be responding in both $t=0$ and $t=-1$ to the program quality difference.

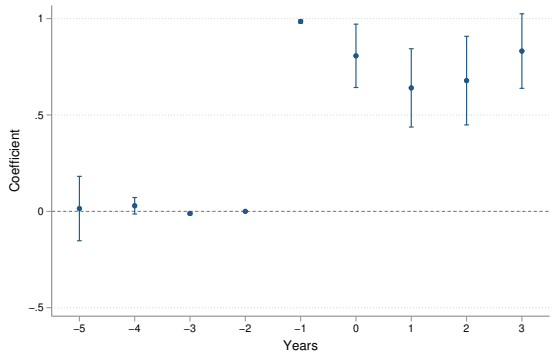
used to test identifying assumptions. We additionally require programs to have at least 15 applicants in the first matching period, -3, to study well-defined programs who consistently receive applicants.

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

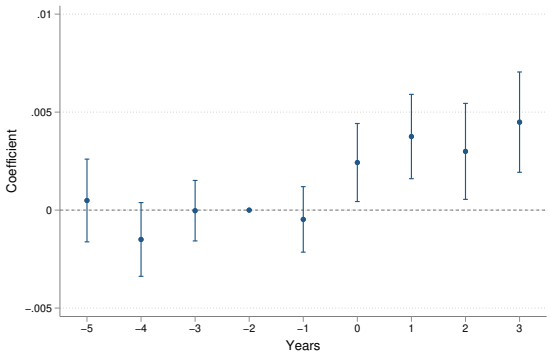
[23] If we do not include this unit fixed effect and do not normalize the coefficients, across all outcomes, none of

Figure 1: Event-study estimates of effect of observable PYS

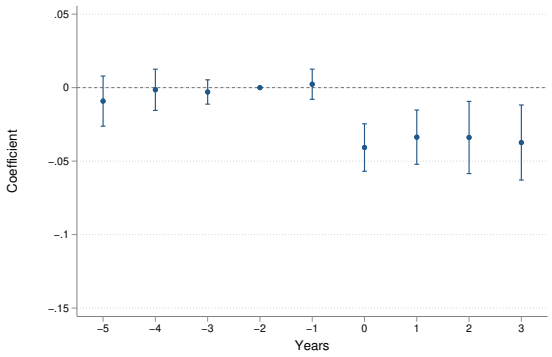
(a) Current Year Score (CYS)



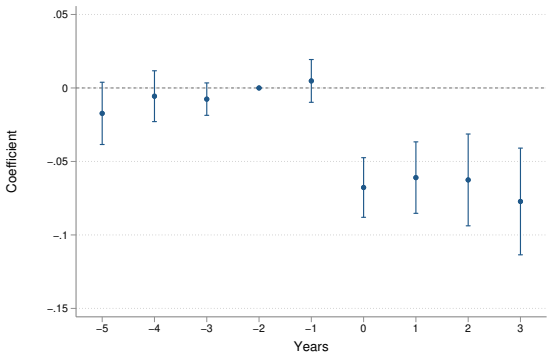
(b) Ln(Average ATAR)



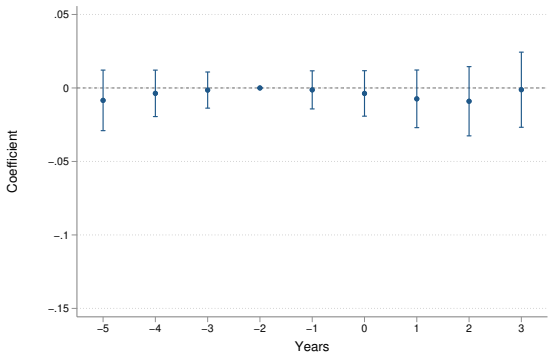
(c) Ln(Number of applicants)



(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 1 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.

the pre-period coefficients are statistically significant, as expected.

Figure 1 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). Prior to the reveal of this test statistic in period 0, we find no relationship between an increase in PYS and average applicant ATAR score or number of applicants. This is consistent with parallel trends prior to the event time, and provides some evidence in favor of the identifying assumption holding. Notably, the small and insignificant coefficient in period -1 implies we find no evidence consistent with a structural change in the program associated with a change in the CYS. An increase in the CYS is uncorrelated with average applicant ATAR score and log number of applicants.

We find that an increase of one point in the observable PYS leads to economically significant changes to the applicant distribution. The average student test score increases by more than 0.2% and the number of applicants decreases by more than 4% in period 0. Both effects are highly statistically significant and are relatively stable or increase in magnitude for the next 3 periods in the event window. The decline in students indicates that fewer students are attracted to an increase in PYS peer quality than are discouraged by it. That there is also an increase in average student score suggests compositional differences in student responses.

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 visualize this potential heterogeneous response in this setting by studying the number of applicants with ATAR scores above and below the PYS in period $t = -2$, the reference period, as outcome variables. 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 2 reports the average of the post-period coefficients, showing reductions of about 7% of applicants in this group.[24] For students with scores above this value, we find no differential effect and can clearly

[24] A small number of observations are dropped as a result of logging the outcome variable. The results are

Table 2: 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	0.00342***	-0.0364***	-0.0671***	-0.00534	0.739***
	(0.000934)	(0.00900)	(0.0121)	(0.00898)	(0.0793)
N	21848	21848	21818	21622	21757

This table displays averages of the post-period event-study coefficients of the effect of the Previous Year Statistic (PYS) on outcomes. Figure 1 displays regression estimates for each period. Regressions are estimated according to Equation 1 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. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

reject the effects are similarly sized to students below the PYS. Consequently, preferences towards higher quality peers appear to be heterogeneous based on the student's own relative ability.

With the results in hand, it is worthwhile considering econometric challenges associated with identifying peer effects (Manski, 1993; Angrist, 2014). Given we model a student's application choice as a function of the PYS, which itself is a function of other (past) students' application choices, a natural question is whether mechanically positive correlations could arise through common student group-based preference shocks. This is unlikely for two reasons. First, of theoretical interest, our dependent (e.g., student average ATAR) and independent (i.e., PYS) variables correspond to different years of application behavior, breaking any mechanical correlation, and the independent variable is a non-linear function dependent on many factors rather than a simple average of application outcomes that we actually study, further mitigating any relationship. Second, of empirical interest, period -1 is the period when any mechanical correlation between outcome and treatment variables would occur as these variables are contemporaneous. Our data is inconsistent with this. We do not see an increase in total applicants or applicant test scores in period -1, nor do we see heterogeneous differences across students above or below the past PYS. In subsequent periods, we observe the number of students who apply decreasing rather than a mechanical increase, as we may worry if there were a mechanical positive correlation. Therefore, for both theoretically- and empirically-founded reasons, mechanical forces appear to have a negligible impact on observed peer effects.

Together, these results are consistent with relative peer preferences. Students, specifically those below the typically observed ability, are less likely to apply to a program when observable peer

quantitatively and qualitatively the same if we use the log of the outcome plus one.

ability increases. Our results are not consistent with structural program quality differences nor strategic application behavior driving these results. The next section strengthens these conclusions by using an alternative research design and reassuringly finds similar results.

II.C.4 Choice updating

To further investigate relative peer preferences, we exploit 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 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 ROLs truthfully capture student ordinal preferences, which a recent literature has challenged.[25] Fack et al. (2019) argue students face a cost to reporting longer ROLs. Therefore, if a student has little admission likelihood for some programs (e.g., a "reach" program) then she may optimally omit some desired programs from her ROL. Even 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 will reflect her true ordinal preferences over c and c' in any *weakly undominated* strategy.

Under this relaxed assumption of students playing weakly undominated strategies, we look within person at how relative rankings over programs respond to new information about a student's ability relative to that of her peers. We focus primarily on *switches*, defined as an instance where program c is ranked higher than program c' on the pre-ROL, both c and c' are on the post-ROL, and c' is ranked above c on the post-ROL. In this case, c' is *promoted* and c is *demoted*. The assumption of undominated strategies allows us to say that the student prefers program c' to c after observing information about her own ability relative to students at these programs.

Our research design closely follows these theoretical ideas. The estimating equation is:

$$Y_{\theta,c} = \sum_{j=-10, j\neq 0}^{10} \beta_j \mathbb{1}[PYS_c - ATAR_\theta = j] + \phi_\theta + \varepsilon_{\theta,c} \tag{2}$$

where $Y_{\theta,c}$ are binary variables measuring whether student θ either promotes or demotes a program c. The sample includes all programs listed on the pre-ROL, as this is the relevant choice set for

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

promotions or demotions, and does not restrict to students with fewer than nine programs on their ROL.[26] The binary variables $\mathbb{1}[PYS_c - ATAR_\theta = j]$ take value 1 if the difference between program c's PYS and student θ's ATAR score is j, and are implemented as such to allow for unrestricted functional form in how student choices are affected by relative peer program differences. We additionally include individual fixed effects, ϕ_θ. Together, this estimating equation compares whether the same student is more or less likely to demote or promote a program based on the program's relative PYS as compared to a program that has the same PYS as the student's ATAR (i.e., the omitted group). If students do not have preferences for relative peer ability, we expect the estimates of β_j to be zero. We two-way cluster the standard errors at the level of the individual and program.

Several assumptions are required to interpret this regression as evidence for relative peer preferences. First, we require our estimated relative comparisons not to be confounded by omitted variables. The central concern is that students' promotion or demotion decisions are correlated with other features of programs, besides their relative peer ability. Many important features of the programs, such as location, subject matter, and prestige, are unchanged before and after students learn their ATAR score and hence are controlled for by the individual fixed effects. Therefore, most central for this assumption are program characteristics that students evaluate with respect to their own score besides peer quality, such as difficulty of instruction. As discussed, only if students update on their own "fit" within the program regardless of peers would the interpretation be confounded.[27] In support of the assumption, Appendix D.1 re-estimates this specification for programs that are newer as opposed to older, based on the presumption students will update on their fit more for lesser-known programs, and we find non-differential results. Furthermore, the research design in Section II.C.3 does not require this assumption and finds substantively similar results.[28]

Second, we assume pre-ROLs reflect relative preferences over any two ranked programs (given initial beliefs), 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.[29] Specifically, Table 1 shows that the average program PYS and average score gap are comparable across the pre- and post-ROLs, indicating that students are likely constructing pre-ROLs following a similar process they use to construct post-ROLs.

[26] In contrast, the analysis in Section II.C.3 limited the sample to applicants with fewer than nine programs to obviate strategic application concerns.

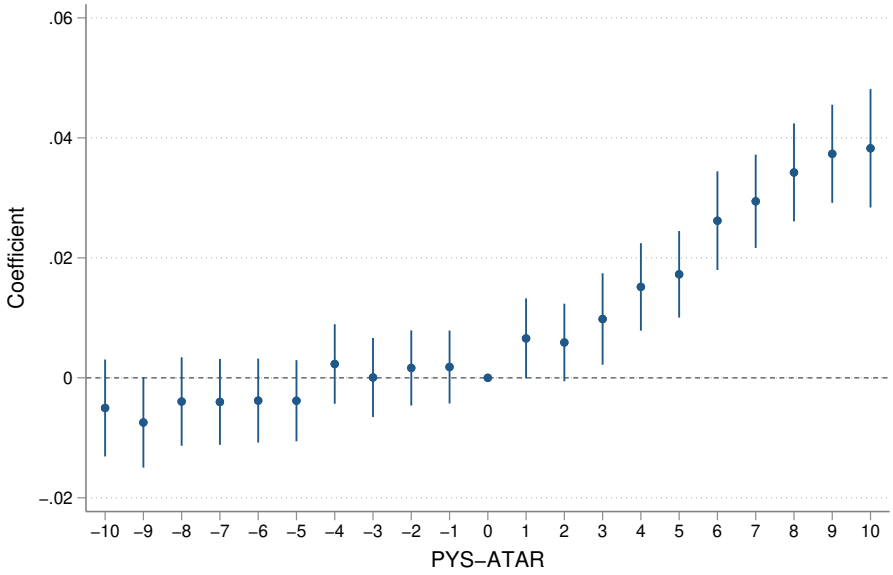
[27] Rothstein and Yoon (2008) discusses how students may be wary of their fit or mismatch with a program.

[28] Appendix D.1 also re-estimates the specification from this section in the same way and finds non-differential results, as expected.

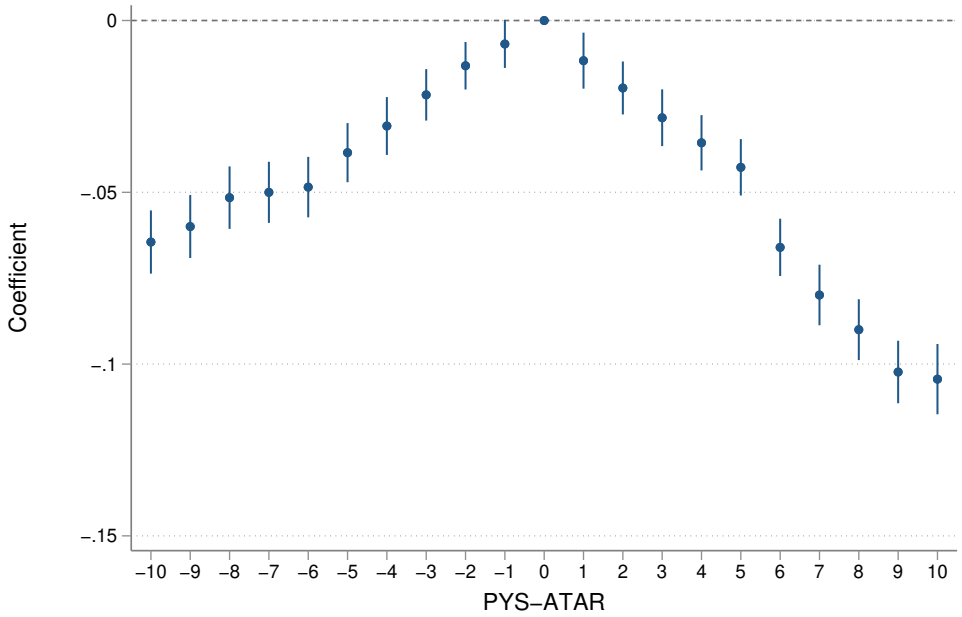
[29] 50% of students submit different pre- and post-ROLs, with 33% rankings changing between the two ROLs on average.

Figure 2: Updating program choices on relative scores

(a) Demote



(b) Promote



This figure displays regression estimates of the impact of a difference in program-student score differences on the decision for a student to demote or promote a program. Regressions estimated according to Equation 2 and include student fixed effects. Coefficients are relative to programs with no PYS to ATAR difference. The sample includes all programs on the "pre-ATAR" ROLs with score differences less than 10 points. Figures show point estimates and 95% confidence intervals clustered at the student and program level.

Figure 2 shows the results for demotions (Panel A) and promotions (Panel B). In Panel A, we find an asymmetric pattern—students demote programs with PYSs progressively higher than their ATAR scores but are no more likely to demote programs with PYSs below their ATAR scores. The asymmetry exactly at the student ATAR score shows students tend to progressively disprefer programs with peer scores above their own but do not do so for programs below. In Panel B, we also find evidence of dispreference for programs that begins at the student ATAR. Somewhat differently than for demotion, students are less likely to promote programs progressively lower than their ATAR although this relationship is weaker than for programs above their ATAR. This pattern can be explained by a student's preference to attend more selective programs, but not ones where the student is overmatched academically. Together, the results are consistent with students having an aversion to being relatively worse than their peers, as evidenced by changes to application behavior exactly at their score, while still valuing program quality. The weak assumptions on student truth-telling further strengthen our findings of relative peer preferences.

Furthermore, we find these choices are consequential, suggesting that relative peer ability is an important component of students' preferences over programs. As discussed in Section II.B, we simulate the true matching using all students' post-ROLs, and also in separate counterfactual worlds where only a single student's pre-ROL is used to determine her assigned program. Across the distributions governing bonus points, we find that changes to ROLs are outcome relevant; between 31% and 48% of students are allocated to different programs if their pre-ROL is used instead of their post-ROL.

II.C.5 Alternative explanations

In Appendix D.2 we evaluate multiple alternative explanations for why the observation of relative peer ability could affect student ROLs. We show each of these explanations predicts a discontinuous drop in the likelihood of ranking programs with PYSs just above (as opposed to just below) the student's ATAR score, which is inconsistent with our empirical findings. Therefore, these alternatives are unlikely to be the drivers for the effects we document. The alternative explanations are:

- Students cannot reason through optimal behavior in the matching mechanism (Li, 2017),
- Students have other non-classical preferences, including loss aversion, which makes them disprefer rejection (Dreyfuss et al., 2021; Meisner and von Wangenheim, 2019; Meisner, 2021),
- Students are uncertain about their preferences over programs, and will optimally acquire more (costly) information over programs to which they have a higher chance of admission (Grenet et al., 2022; Immorlica et al., 2020; Hakimov et al., 2021). Risk-averse students are

more likely to highly rank programs for which they have gathered information.

III Model Setup and Existence of Stable Matchings

We present a large-market matching model with student preferences over the distribution of peer abilities, following our empirical environment. We present sufficient conditions for the existence of stable matchings.

A continuum of students is to be matched to a finite set of 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 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 $\theta=(u^\theta,r^\theta)$. $u^\theta(c|\alpha)\in\mathbb{R}$ 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 regardless of the assignments of other students.

$r^{\theta,c}\in[0,1]$ 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 $c\in C$, $\eta(\{\hat{\theta}\in\Theta|r^{\hat{\theta},c}=r^{\theta,c}\})=0$. Because program c_0 represents the outside option without a binding capacity constraint, r^{θ,c_0} can be viewed as student θ's "ability."

With these preliminaries, 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 $E=[\eta,q,N,\Theta]$. It will often be useful to denote the ordinal preferences of student θ. 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 induced 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 $u^\theta(c_i|\alpha)\geq u^\theta(c_j|\alpha)$ ($u^\theta(c_i|\alpha)>u^\theta(c_j|\alpha)$).

To capture that peer preferences depend on the ability profile of students at each program, we consider the distribution of scores at each program given an assignment. 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. 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 empirical setting.

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, and not on their identities. 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. Our model, and the remainder of our analysis, can easily be extended in a straightforward manner to include more dimensions than just student ability (for example, we could include race,

gender, or height dimensions to each θ), and therefore, anonymity is the substantive restriction in **A2**. **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.

III.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 does 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.[30] 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 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 score 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 lemma, a direct corollary of the supply and demand lemma of Azevedo and Leshno (2016) and Leshno (2022) holds:

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 a large class of markets. We prove this result by constructing an operator whose fixed points corresponds to stable matchings, and

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

show, using a fixed-point theorem, that at least one fixed point exists.[31]

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).

Remark 2. *There need not be a unique stable matching. For example, if $N \geq 2$, all programs offer students similar intrinsic utility, and sufficiently many 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.*

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

Theorem 1 tells us that a stable matching exists in many markets. 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 the beliefs of students 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.

IV.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 a 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 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

[31] Our proof generalizes the argument of Leshno (2022) by application of a high-dimensional fixed-point theorem that is necessary to consider peer preferences that depend flexibly on λ. Grigoryan (2021) uses the same fixed-point theorem we use in a matching market with complementarities. More broadly, fixed-point arguments are often used to show existence results in the literature (see, for example, Pycia and Yenmez, 2023).

$\varphi^\theta(\tilde{\succ})\tilde{\succeq}^\theta\varphi^{\theta'}(\tilde{\succ})$. That is, a student is not matched to a program she ranks below 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.[32] 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.[33]

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.[34] 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 $\tilde{\succ}^\theta = \succeq^{\theta|\mu_*}$.[35] 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 and let μ_* be a stable matching. If for all $r\in[0,1)^{N+1}$ and all $\tilde{\succ}$ it is*

[32] 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.

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

[34] 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 IV.

[35] As we discuss in the proof of Proposition 1, for any stable mechanism φ, if almost 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.

the case that $\eta(L_{\succsim,\varphi,r}) > 0$ 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 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.

IV.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. Does this process always lead to a stable matching in the long run? If so, a patient market designer may be content to rely on this status quo. Unfortunately, we show that for almost any collection of peer preferences, this process is not guaranteed to lead to a stable matching even in the long run.

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)*—generates the matching in each period. The market is initialized with an arbitrary assignment $\mu_0 \in \mathcal{A}$.[36] 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 (as discussed in our empirical setting) students use information from the previous period in a Cournot-updating fashion; that is, each period t student has a Dirac belief that μ_t will equal μ_{t-1}.[37]

The main theoretical result of this section finds that the TIM process is not guaranteed to generate a stable matching in any time period for almost *every* specification of peer preferences.

[36] 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.

[37] 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, $\lambda(\mu_{t-1}),...,\lambda(\mu_{t-k})$.

We present it here in an informal manner, before describing economic intuitions and 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* $\{f^{\theta,c}(\cdot)\}_{\theta\in\Theta,c\in C\setminus\{c_0\}}$*, there exists a market* E *with the same collection of peer preference functions and* μ_0 *for which the TIM process does not yield or approximate a stable matching at any time* t*.*

Several observations are in order, and these serve as preliminary considerations before proving Theorem 2. First, there is a unique stable matching in a market in which peer preferences are defined by μ_{t-1}, and this matching coincides with μ_t. This result follows directly from Assumption **A3** and Grigoryan (2022).

Remark 3. *Fix a market* $E=[\eta,q,N,\Theta]$ *and let* μ_t *be the matching constructed at time* $t\geq 1$ *in the TIM process. 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* $\succeq$*:* $\zeta^{\eta,\mu_{t-1}}(\{\theta\in\Theta^{\mu_{t-1}}|r^{\theta}\in R \text{ and } \succeq^{\theta|\alpha}=\succeq\})=\eta(\{\theta\in\Theta|r^{\theta}\in R \text{ and } \succeq^{\theta|\mu_{t-1}}=\succeq\})$*. 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* $x\in[0,1]$*.*

Given our assumption on beliefs, the previous remark implies each student θ has a weakly dominant strategy 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 students report their true preferences 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.

Second, we observe that Theorem 2 does not imply that the TIM process *cannot* ever yield a stable matching, only that it *need not* do so. An important consideration in proving Theorem 2 is identifying when a stable matching is generated. The following result connects stability in the TIM process to classical intuitions surrounding price convergence and equilibrium in exchange economies: 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 the result, 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 $\mu_1, \mu_2, \ldots$ 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.*

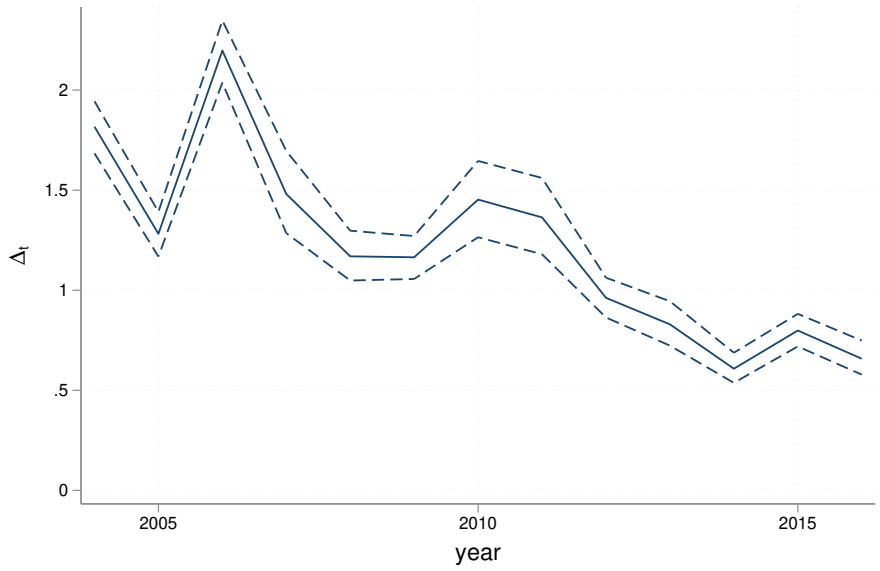
Remark 4. *Proposition 2 is particularly amenable to empirical testing for two reasons. First, suppose students have preferences over* summary statistics *of the ability distribution, as in our empirical setting. A summary statistic of abilities at program c is defined as a function $s^c : \Lambda \to [0,1]$. 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.

Third, there are markets for which the TIM process does not converge to a stable matching for almost any initial condition μ_0. Example 1 in the appendix constructs a market such that the TIM process diverges for any μ_0 which is not the (unique) stable matching. Non-convergence of the TIM process is therefore not a pathological outcome.

Returning to Theorem 2, we show that any collection of peer preferences that admits a *negative externality group*–informally, a set of students who reduce one anothers' utilities–admit markets in which the TIM process does not converge.

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.

Focusing on peer preferences that admit a negative externality group is important for two reasons. First, we show that the existence of a negative externality group is likely from a topological perspective; peer preferences generically admit a negative externality group. Second, negative externality groups are economically meaningful, as they capture the notion that there can exist certain students who are undesirable to others. Whenever a negative externality group exists, the TIM process can cycle as in Scarf (1960), which by Proposition 2 implies that the TIM process never finds a stable matching.

IV.C Assessing market stability

By Proposition 2, a natural empirical test of stability assesses whether the difference between every program's PYS and CYS is close to, or converges to, zero. Figure 3 shows substantial difference between these statistics over time, and clearly rejects a null of no difference across all years. While this test suggests that the market has not converged to a stable matching, the theory does not provide a way to translate differences in observable student-program distributions over time to a direct measure of market instability. Estimating whether the level of instability is economically relevant is critical for understanding market functioning and the potential benefits of modifying the matching process. We develop an alternative approach to estimate the degree of instability below.

We consider a standard notion of instability. A matching is said to be more unstable the higher the proportion of students involved in blocking pairs—recall that a student and program form a blocking pair when the student prefers the program to her current assignment, and where the program either has an empty seat or has admitted a student it assigns a lower priority to.

Consider an ideal experiment to calculate the share of students involved in blocking pairs in the matching that arises from the status quo process using outdated peer information. First, we inform applicants of the characteristics of current students across programs, elicit student ROLs, and allocate the spots to create a supposed "stable" matching. Second, we inform the same students of peer characteristics across programs from the newly-created matching and then re-elicit ROLs to observe changes to student preferences. A student is a member of a blocking pair if and only if, using her updated ROL and the original ROLs for all other students, she is assigned to a different program.

An example motivates our approach to approximating this ideal experiment. Consider all students with an ATAR score ℓ (e.g., 90) in a year t and define the share of students who matriculate in year t to program c to be $\hat{f}_{t,\ell,c}$. This probability distribution function incorporates information on student ROLs and the program where students are ultimately matched. These students would have preferred current information on their prospective peers (i.e., the CYS) when applying but only were able to observe past peer information (i.e., the PYS). To provide updated counterfactual information on peer characteristics of program c, we use ROLs for similar students (same ATAR) in year $t+1$ who do observe the CYS in year t (i.e., their PYS). We can then define the counterfactual distribution $\hat{g}_{t,\ell,c}$ to represent where students of score ℓ with these updated preferences from year $t+1$ would have been assigned in year t by using the admissions standards in year t, derived from the original ROLs. In contrast to the ideal case though, there are many other possible factors beyond peer information that may lead to differences between the $\hat{f}$ and $\hat{g}$ distributions. Disentangling these factors is the main empirical challenge we address below.

We define an observed measure of misallocation in year t at program c as

$$\hat{M}_{t,c} = \frac{1}{2} \sum_{\ell \in \{30,\ldots,100\}} |\hat{g}_{t,\ell,c} - \hat{f}_{t,\ell,c}|$$

which sums matriculation differences over all ATAR scores.[38] If the terms $\hat{g}_{t,\ell,c} - \hat{f}_{t,\ell,c}$ only include differences arising from updated peer information, then $\hat{M}_{t,c}$ underestimates the true share of students in blocking pairs in year t at program c. This is because the ideal experiment counts blocking

[38] Dividing by two avoids double counting students. An analogous expression, summing over programs to calculate aggregate misallocation, appears in Equation A.9 in the Proof of Theorem 1.

pairs at the student level, while our approach aggregates students of the same ATAR score and therefore misses any two students with the same ATAR score who "trade" seats at different programs.

Our goal is therefore to isolate how updated information on peer composition affects the matriculation difference $\hat{g}_{t,\ell,c} - \hat{f}_{t,\ell,c}$. To do so, we translate this to a simple regression-based framework. We first define the policy of updating information on programs as the difference between the unobserved CYS and the observed PYS. This information on peer composition can affect choices for, and matriculation to, program c in time t for students of score ℓ by directly revealing information on peer composition for this program, and indirectly by revealing information on programs likely to be an outside option. To incorporate relevant outside options for a program c, we define the top five most common other programs besides c ranked by number of applicants in year t for rank r as $o(c,t,r)$. For example, the most common outside option will have rank $r=1$. In estimation, we flexibly allow peer information of these outside options to affect the matriculation differences. As other less common outside options may also affect matriculation, we view this choice as underestimating the role of indirect information on the ultimate allocation. Further differences in these empirical distribution functions can arise not just from updated information but through heterogeneity in multiple sources, including across students, programs, and time, which we control for directly.

We estimate the effect of updated peer information using the following regression equation

$$\hat{g}_{t,\ell,c} - \hat{f}_{t,\ell,c} = \phi(CYS_{t,c} - PYS_{t,c}) + \sum_{j=1}^{5} \psi_j (CYS_{t,o(c,t,j)} - PYS_{t,o(c,t,j)}) + \gamma_\ell + \delta_t + \sigma_c + \varepsilon_{t,\ell,c} \quad (3)$$

which analyzes the difference in matriculation distributions as a linear function of own program updating, indirect program updating, and other determinants.[39] The parameter ϕ represents the effect of updating own program information on matriculation for a specific ATAR score level ℓ, while ψ_j is the effect for the top outside options. The difference $\hat{g}_{t,\ell,c} - \hat{f}_{t,\ell,c}$ reflects variation in ROLs across years for comparable students subject to the same matriculation rules, so empirically we are most concerned with omitted variables that may lead to differential application behavior for reasons outside of direct peer information. In particular, we include time fixed effects δ_t to control for aggregate variation in preferences and the marketplace, student score fixed effects γ_ℓ to control for variation in applicant behavior across ATAR groups, and program fixed effects σ_c to control for time-invariant

[39] Note that while the form of this regression—a change in the dependent variable regressed on a change in an independent variable—appears to be a first-differences estimator, it is not. Instead, the outcome difference $\hat{g}_{t,\ell,c} - \hat{f}_{t,\ell,c}$ depends on the match in year t and hence $\hat{g}_{t,\ell,c}$ is not equal to $\hat{f}_{t+1,\ell,c}$. Consequently, this analysis is not a transformation of the panel-based analysis in Section II.C.3.

differences across programs such as whether some programs are on average more or less popular.

One additional empirical concern is that an increase in the CYS may be mechanically correlated with increased student demand in the same year, reflected in f. In this case, a spurious shock could drive both variables, leading to omitted variables bias. Notably, the analysis in Section II.C.3 investigated whether student application behavior responded to changes in observable information through the PYS, or through changes in the unobservable CYS. It finds strong evidence against student application behavior being correlated to CYS and strong evidence in favor of the role of observable information through the PYS, reducing the plausibility of bias from this potential source.

Using the estimated parameters from Equation 3, we define our preferred measure of instability in program c at year t as

$$M^*_{t,c} = \frac{1}{2} \sum_{\ell \in \{30,\ldots,100\}} |\hat{\phi}(CYS_{t,c} - PYS_{t,c}) + \sum_{j=1}^{5} \hat{\psi}_j (CYS_{t,o(c,t,j)} - PYS_{t,o(c,t,j)})|, \tag{4}$$

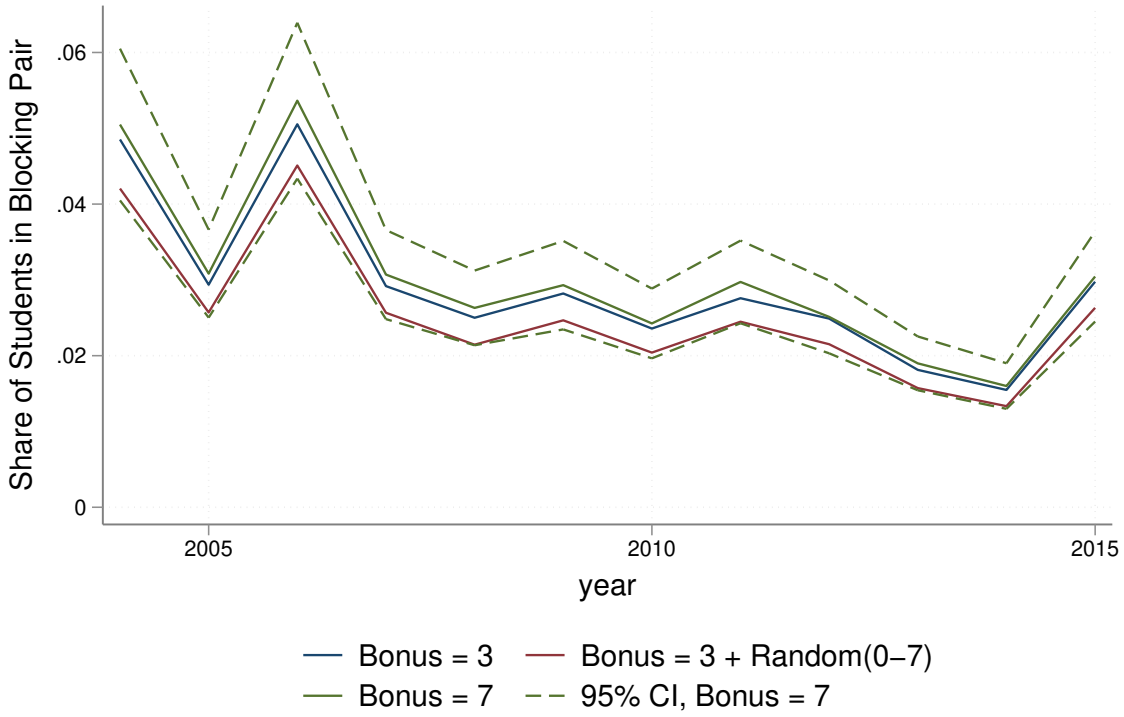
which only incorporates the effects of peer information ($\hat{\phi}$ and $\hat{\psi}_j$) multiplied by the change in available peer information. We can then define total instability within a year as $M^*_t = \sum_{c \in C \setminus \{c_0\}} M^*_{t,c}$ and average instability $M^* = \frac{1}{T} \sum_{t \in T} M^*_t$. Three reasons imply that this quantity is a lower bound of the true share of blocking pairs. First, as discussed, the difference $\hat{M}_{t,c}$ undercounts students with the same ATAR score who form offsetting blocking pairs. Second, the attenuated difference between $\hat{f}$ and $\hat{g}$ represents non-classical measurement error from the true difference and similarly attenuates the coefficients.[40] Third, we incorporate the effects of outside options by only focusing on the top five outside option programs, which can undercount the impact of the reveal of information from other programs. Together, these results imply that our estimated share of blocking-pairs will be conservative for the true share of blocking pairs in the ideal experiment.

Table A.2 presents our regression estimates. Columns (1-3) differ as a result of the bonus points used to simulate the matches. We find that the direct impact of peer information has a large effect on the matriculation distribution, with large and highly statistically significant coefficients. Variation in indirect peer information from the outside option schools has a much more limited role. Alternative choices towards these indirect effects are therefore likely to be inconsequential. The results are also highly consistent across the bonus point regimes, indicating that these distributional

[40] In a bivariate regression, a scalar $\kappa \in [0,1)$ multiplied by the outcome will result in an estimated coefficient of κ multiplied by the true coefficient. In practice, it is possible for the attenuation factor to be stochastic, although the same attenuation result holds as long as κ does not strongly covary with the regressors.

choices have limited importance in determining our ultimate parameters of interest.

Figure 4: Share of Students in Blocking Pairs, by Year (M_t^*)



This figure plots the estimated share of blocking pairs by year across three different distributions of bonus points. These shares are calculated from Equation 4 across all years. Bonuses are assigned using three regimes: (1) bonus = 3 for all applications, (2) bonus = 3 + randomly assigned from a uniform distribution over integers 3-7 at the program level, (3) bonus = 7 for all applications. Bootstrapped 95% CIs taken from 1000 draws over the parameter estimates from Equation 3 are shown for bonus regime (3).

Figure 4 shows our estimated lower bound on the share of students involved in blocking pairs, M_t^*, across the study period for the three regimes of bonus points used to simulate the matches. The results are highly consistent across the bonus point distributions; we estimate a lower bound of the (across year) average share of students in blocking pairs, M^*, between 2.55% and 3.05% due to inaccurate peer information. Our preferred specification of 7 bonus points has 95% confidence interval of [2.48%, 3.62%].[41] We view this is an economically significant share of students involved blocking pairs. As a benchmark, the National Residency Matching Program had an upper bound of 4% of agents in blocking pairs due to spousal preferences (i.e. 4% of participants were coupled with one another), causing its redesign. Our lower bound M_t^* also appears to be relatively consistent across time, and there is little evidence to suggest a long-term convergence to stability.

[41] For the regime with bonus = 3, the 95% CI is [2.21%, 3.63%], and for regime with bonus = 3 + uniform[0, 7], the 95% CI is [2.00%, 3.11%].

We also use our framework to consider how instability affects traditionally disadvantaged groups, such as indigenous and low SES individuals. We observe enrollee shares of students across demographic groups at the level of the university-year, so to calculate group-specific instability rates we re-weight our overall instability measures by the share of the student population from this group at each university. For our preferred bonus point scheme, we calculate low SES and indigenous students are approximately 25% more likely to be in blocking pairs compared to the overall student average. These higher rates of blocking pairs for disadvantaged groups may represent a significant educational barrier.

Finally, we provide evidence on an important and observable impact of instability: program attrition. We define and measure attrition as occurring when a student, who commences in year t, does not complete a program before or during year $t+4$, and otherwise, we say the student completes the program.[42] Whenever a blocking pair is consummated (either with a different program or the student's outside option), attrition occurs, and therefore, we expect attrition to be higher at programs with more students who are members of blocking pairs, which we previously estimated as $M^*_{t,c}$.

Table 3: Relationship Between Completion Rate and $M^*_{t,b}$

	Completion rate (%)		
	(1)	(2)	(3)
$M^*_{t,b}$	-5.156*	-5.281*	-5.006*
	(3.061)	(2.751)	(2.897)
Linear field-of-study trends	No	Yes	No
Year by field-of-study fixed effects	No	No	Yes
N	1675	1675	1675

This table presents the relationship between the 4-year completion rate of commencing students and estimated $M^*_{t,b}$. The university-program-year-specific $M^*_{t,c}$ estimated according to Equation 4 is aggregated up to the university-field-year level to match the level of the completion rate records. All columns control for year and university-program fixed effects in a two-way fixed effects specification, with columns (2) and (3) additionally controlling for linear field of study year trends and field of study-year fixed effects, respectively. Standard errors are bootstrapped accounting for first- and second-step estimation with 2000 draws, with the bootstrap clustered at the university-field level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

[42] 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.

We investigate the relationship between the completion rate and the share of students involved in blocking pairs across program years in Table 3. For privacy reasons, we do not observe completion at the program level and instead merge in the completion rate at the university-year-broad field of study level. There are 12 broad fields of study. To match this data aggregation, we aggregate the program-year specific $M^*_{t,c}$ to calculate a university-year-broad field of study level sum, $M^*_{t,b}$. Using a two-way fixed effects design that controls for year and university-broad field of study fixed effects, column (1) shows marginally statistically significant evidence that a one unit increase to the estimated $M^*_{t,b}$ decreases completion by 5.16 percent ($p=0.09$).[43] More easily interpreted, a one standard deviation increase in the estimated $M^*_{t,b}$ accounts for a sizable 0.059 decrease in the standard deviation of the completion rate, after accounting for university-field and year heterogeneity in both measures. Columns (2-3) show our estimate is robust to accounting for differential time trends either through linear field of study time trends ($p=0.05$) or through field-year fixed effects ($p=0.08$), respectively. Importantly, the latter specification isolates variation within fields of study; therefore, differential trends across fields, potentially as a result of relative labor demand shocks, cannot explain our findings. Overall, given the limited statistical power arising from merging these outcomes at the university-broad field level, we view these findings as highlighting important impacts of instability due to peer preferences on student attrition.

V An Approximately Stable Mechanism

Given theoretical and empirical evidence in previous sections 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.[44] 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 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.

Students in each cohort are assigned to one of many smaller submarkets, and students in each

[43] We account for the generated regressor and the clustering by field-year through a two-step clustered bootstrap routine with 2000 draws. There are 139 clusters.

[44] 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.

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 $\sum_{\ell=1}^{T} \eta_t(\Theta) = 1$, 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.[45] 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:*

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

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

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 are unwilling or unable to misreport. 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.

Proposition 3. *Let $\epsilon > 0$.*

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 $\tau = |\Lambda_0^\gamma|$) if each student θ reports $\succeq^{\theta|\lambda_\tau}$ at each step τ.*

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.[46] *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).*

VI Discussion and Conclusion

Our analysis provides new evidence on the presence of peer preferences, and on the barriers they pose to constructing a stable matching, in status quo school-choice markets. We show that the status quo process 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 exhibits 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 leads to the observable consequence of greater attrition. Finally, we estimate a *lower bound* on the share of students involved in blocking pairs due to peer preferences which is large compared to an upper bound on the share of agents involved in blocking pairs due to peer preferences in the market for American medical doctors.

As instability in the medical market led to a redesign of the matching mechanism in use (Roth and Peranson, 1999), we 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 Bo and Hakimov (2019); Luflade (2019)) and finds a stable matching in the presence of peer preferences.

[46] To see this point, consider Example 1 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 yield 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 $\frac{1}{2}$.

In ongoing work, Fonseca et al. (2023) present sufficient conditions for the status quo process to converge to a stable matching. Intuitively, these conditions ensure the market exhibits a form of aggregate gross substitutes, and hence the tâtonnement-like process we study is guaranteed to converge.[47] Indeed, results in Fonseca et al. (2023) indicate that the degree of instability in the NSW market is likely smaller than that we would expect to find in many other markets; implementing a mechanism that finds a stable matching when students have peer preferences may be even more consequential in other markets.

Finally, our work 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 to 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

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

Atila Abdulkadiroğlu, Joshua Angrist, and Parag A. Pathak. The elite illusion: Achievement effects at boston and new york exam schools. *Econometrica*, 82(1):137–196, 2014.

Atila Abdulkadiroğlu, Yeon-Koo Che, and Yosuke Yasuda. Expanding "choice" in school choice. *American Economic Journal: Microeconomics*, 7(1):1–42, 2015.

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

Atila Abdulkadiroğlu, Parag A. Pathak, Jonathan Schellenberg, and Christopher R. Walters. Do parents value school effectiveness? *American Economic Review*, 110(5):1502–39, 2020.

Patrick Agte, Claudia Allende, Adam Kapor, Christopher Neilson, and Fernando Ochoa. Search and biased beliefs in education markets. mimeo, 2023.

Robert Ainsworth, Rajeev Dehejia, Cristian Pop-Eleches, and Miguel Urquiola. Information, preferences, and household demand for school value added. NBER WP 28267, 2020.

[47] For example, consider a class of markets with positive network effects, which represent peer preferences present in a Japan school choice market (see https://web.archive.org/web/20230324050348/https://news.yahoo.co.jp/articles/124f47d03ca41a70512b5b39e2f04df8718f2c1a). Fonseca et al. (2023) show that in markets with sufficiently few programs, these preferences guarantee convergence to stability.

Claudia Allende. Competition under social interactions and the design of education policies. mimeo, 2020.

Claudia Allende, Francisco Gallego, and Christopher Neilson. Approximating the equilibrium effects of informed school choice. mimeo, 2019.

Joshua D Angrist. The perils of peer effects. *Labour Economics*, 30:98–108, 2014.

Georgy Artemov, Yeon-Koo Che, and YingHua He. Strategic 'mistakes': Implications for market design research. mimeo, 2020.

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

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

Eryk Bagshaw and Inga Ting. Nsw universities taking students with atars as low as 30. *The Sydney Morning Herald*, 2016.

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

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

Diether W. Beuermann and C. Kirabo Jackson. The short and long-run effects of attending the schools that parents prefer. NBER WP 24920, 2019.

Diether W. Beuermann, C. Kirabo Jackson, Laia Navarro-Sola, and Francisco Pardo. What is a good school, and can parents tell? evidence on the multidimensionality of school output. mimeo, 2019.

Inacio Bo and Rustamdjan Hakimov. The iterative deferred acceptance mechanism. mimeo, 2019.

George W. Brown. Iterative solutions of games by fictitious play. In Tjalling C. Koopmans, editor, *Activity Analysis of Production and Allocation*, pages 374–376. Wiley, 1951.

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

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

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

Diego Carrasco-Novoa, Sandro Diez-Amigo, and Shino Takayama. The impact of peers on academic performance: Theory and evidence from a natural experiment. mimeo, 2021.

Scott E. Carrell, Bruce I. Sacerdote, and James E. West. From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3):855–882, 2013.

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

Yeon-Koo Che, Dong Woo Hahm, Jinwoo Kim, Se-Jik Kim, and Olivier Tercieux. Prestige seeking in college application and major choice. mimeo, 2022.

Li Chen and Juan Sebastián Pereyra. Self-selection in school choice. *Games and Economic Behavior*, 117:59–81, 2019.

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

Sarah Cohodes, Sean Corcoran, Jennifer Jennings, and Carolyn Sattin-Bajaj. When do informational interventions work? experimental evidence from new york city high school choice. NBER WP 29690, 2022.

Elizabeth Dhuey, David Figlio, Krzysztof Karbownik, and Jeffrey Roth. School starting age and cognitive development. *Journal of Policy Analysis and Management*, 38(3):538–578, 2019.

Will Dobbie and Roland G. Fryer Jr. The impact of attending a school with high-achieving peers: Evidence from the new york city exam schools. *American Economic Journal: Applied Economics*, 6(3):58–75, 2014.

Bnaya Dreyfuss, Ori Heffetz, and Matthew Rabin. Expectations-based loss aversion may help explain seemingly dominated choices in strategy-proof mechanisms. *American Economic Journal: Microeconomics, Forthcoming*, 2021.

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

Bryan Ellickson, Birgit Grodal, Suzanne Scotchmer, and William R Zame. Clubs and the market. *Econometrica*, 67(5):1185–1217, 1999.

Benjamin Elsner and Ingo E. Isphording. A big fish in a small pond: Ability rank and human capital investment. *Journal of Labor Economics*, 35(3):787–828, 2017.

Benjamin Elsner, Ingo E. Isphording, and Ulf Zölitz. Achievement rank affects performance and major choices in college. mimeo, 2018.

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

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

Norman T. Feather. Attitudes towards the high achiever: The fall of the tall poppy. *Australian Journal of Psychology*, 41(3):239–267, 1989.

Ricardo Fonseca, Bobak Pakzad-Hurson, and Matthew Pecenco. Entry and exit in school choice markets. mimeo, 2023.

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

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

Diego Gentile Passaro, Fuhito Kojima, and Bobak Pakzad-Hurson. Equal pay for *Similar* work. mimeo, 2023.

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

Julien Grenet, YingHua He, and Dorothea Kübler. Preference discovery in university admissions: The case for dynamic multi-offer mechanisms. *Journal of Political Economy, Forthcoming*, 2022.

Aram Grigoryan. School choice and the housing market. mimeo, 2021.

Aram Grigoryan. On the convergence of deferred acceptance in large matching markets. mimeo, 2022.

Pablo Guillen, Onur Kesten, Alexander Kiefer, and Mark Melatos. A field evaluation of a matching mechanism: University applicant behaviour in australia. *The University of Sidney Economics Working paper Series*, 2020.

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

Rustamdjan Hakimov, Dorothea Kübler, and Siqi Pan. Costly information acquisition in centralized matching markets. mimeo, 2021.

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

Justine S. Hastings and Jeffrey M. Weinstein. Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly Journal of Economics*, 123(4): 1373–1414, 2008.

Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. Heterogeneous preferences and the efficacy of public school choice. mimeo, 2009.

Nicole S. Immorlica, Jacob D. Leshno, Irene Y. Lo, and Brendan J. Lucier. Information acquisition in matching markets: The role of price discovery. mimeo, 2020.

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

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

Tomás Larroucau and Ignacio Rios. Dynamic college admissions. mimeo, 2020b.

Jacob D Leshno. Stable matching with peer effects in large markets - existence and cutoff characterization. mimeo, 2022.

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

Margaux Luflade. The value of information in centralized school choice systems. mimeo, 2019.

Anthony Manny, Helen Yam, and Robert Lipka. 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*, 2019.

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

Herbert Marsh, Marjorie Seaton, Ulrich Trautwein, Oliver Lüdtke, K.T. Hau, Alison O’Mara, and Rhonda G. Craven. The big-fish–little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, 20:319–350, 2008.

Vincent Meisner. Report-dependent utility and strategy-proofness. mimeo, 2021.

Vincent Meisner and Jonas von Wangenheim. School choice and loss aversion. mimeo, 2019.

Yiannis Moschovakis. *Notes on Set Theory, Second Edition*. Springer, 2006.

Richard Murphy and Felix Weinhardt. Top of the class: The importance of ordinal rank. *Review of Economic Studies*, 87(6):2777–2826, 2020.

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

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

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

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

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

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

Junping Qiu and Rongying Zhao. *College admissions cutoffs and application guide: 2007-2008, Second Edition*. Science Press: Beijing, 2007.

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

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

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

Jesse Rothstein and Albert Yoon. Mismatch in law school. NBER WP 14275, 2008.

Jesse M Rothstein. 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, 2006.

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

Bruce Sacerdote. Experimental and quasi-experimental analysis of peer effects: Two steps forward? *Annual Review of Economics*, 6:253–272, 2014.

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

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

Yan Song, Kentaro Tomoeda, and Xiaoyu Xia. Sophistication and cautiousness in college applications. mimeo, 2020.

Sándor Sóvágó and Ran I. Shorrer. Obvious mistakes in a strategically simple college-admissions environment. mimeo, 2018.

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

Han Yu. Am i the big fish? the effect of ordinal rank on student academic performance in middle school. *Journal of Economic Behavior & Organization*, 176:18–41, 2020.

Román Andrés Zárate. Social and cognitive peer effects: Experimental evidence from selective high schools in peru. mimeo, 2019.

ONLINE APPENDIX

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

A Example Discussed in Section IV

Example 1. *Let $N=1$. To economize on notation, we suppress program indices; all notation refers only to the non-outside-option program c. c has $q<1$ measure of seats, and let $r^\theta := r^{\theta,c} = r^{\theta,c_0}$ for all $\theta \in \Theta$. Let $s(\lambda(\alpha))$ be the mean score r^θ of students assigned to c in α, that is*

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

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

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

for some $k>0$ and each v^θ is distributed independently and uniformly over $(0,1)$. Any θ is better off enrolling at the program if and only if $v^\theta - f(s(\lambda(\alpha)), r^\theta) \geq 0$, where we break ties in favor of the student attending the program. 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" type.

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

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

We claim that the TIM process does not converge for any $s_0 = s(\lambda(\mu_0)) \neq \frac{1}{2}$ *when* $k \geq \frac{8}{1-\gamma}$. *Recall that as* $s(\cdot)$ *is a function of* λ, *if the sequence* $s_1, s_2, \ldots$ *does not converge, then neither does the sequence* $\lambda_1, \lambda_2, \ldots$.

To show this claim, let $s_0 = \frac{1}{2} - \delta$ *for some* $\delta > 0$ *(by the symmetry of the market, similar logic holds if* $\delta < 0$*). First suppose that* $k\delta \geq 1$. *Then in* μ_1, *none of the students with* $r^\theta < \frac{1}{2}$ *who have strong peer preferences will enroll in* c, *and all other students will. Therefore,*

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

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

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

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

For $k \geq \frac{8}{1-\gamma}$ *we claim that* $s(\lambda(\mu_1)) \geq \frac{1}{2} + \delta$. *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* $s_t, s_{t+2}, s_{t+4}, \ldots$ *is non-decreasing where each element is strictly larger than* $\frac{1}{2}$ *and* $s_{t+1}, s_{t+3}, s_{t+5}, \ldots$ *is non-increasing where each element is strictly smaller than* $\frac{1}{2}$. *Therefore, the TIM process does not converge.*

B 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 Lemma 13 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,

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

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

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

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

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

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 follows

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, $\lambda^x_{\ell'} > \lambda^x_* - \epsilon$ and $\lambda^{x'}_{\ell'} < \lambda^{x'}_* + \epsilon$. Therefore, $\lambda^x_{\ell'} - \lambda^{x'}_{\ell'} > \lambda^x_* - \lambda^{x'}_* - 2\epsilon = \delta - 2\epsilon = 0$, which implies that $\lambda_{\ell'}$ is not non-decreasing. Contradiction with $\lambda_{\ell'} \in \Lambda$. The non-decreasing property of λ_* establishes that λ'^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, $\lambda^1_\ell \leq 1$ because $\lambda_\ell \in \Lambda$ for all ℓ. Therefore, for $0 < \epsilon \leq \delta$ and any ℓ, $||\lambda_\ell - \lambda_*||_\infty \geq |\lambda^1_\ell - \lambda^1_*| \geq 1 + \delta - \lambda^1_\ell \geq \delta \geq \epsilon$. 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 $|\lambda^x_\ell - \lambda^x_*| < \epsilon$ and $|\lambda^{x'}_\ell - \lambda^{x'}_*| < \epsilon$ for all $\ell > L$. Consider any $\ell' > L$. Then by the preceding argument, $\lambda^x_{\ell'} < \lambda^x_* + \epsilon$ and $\lambda^{x'}_{\ell'} > \lambda^{x'}_* - \epsilon$. Therefore, $\lambda^{x'}_{\ell'} - \lambda^x_{\ell'} > \lambda^{x'}_* - \lambda^x_* - 2\epsilon = x' - x + \delta - 2\epsilon = x' - x$, 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

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

The first equality follows because each student $\theta \in \Theta(\alpha,\alpha')$ is double counted in the RHS of the top line.[2] 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

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

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. 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 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 $\tilde{\succ}$ is a Bayes Nash equilibrium of φ but that $\mu = \varphi(\tilde{\succ})$ 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 $\tilde{\succ}$. Let p be the associated cutoff vector. Now consider reported preferences $\hat{\succ}$ where $\hat{\succ}^{\theta'} = \tilde{\succ}^{\theta'}$ for all $\theta' \neq \theta$ and $\hat{\succ}^\theta$ 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 $\hat{\succ}$ and $\tilde{\succ}$. Since (θ,c) block μ it must be that $r^{\theta,c} \geq p^c$. But then $\varphi^\theta(\hat{\succ}) = c$ since c is a stable mechanism. Contradiction with $\tilde{\succ}$ being a Bayes Nash equilibrium. □

Proof of Part 2: Let $\tilde{\succ}$ be a Bayes Nash equilibrium, and suppose for contradiction that $\varphi(\tilde{\succ}) = \mu_*$. By Remark 3 and the ongoing assumption that μ_* is stable, it must be that μ_* is associated with some cutoff vector p satisfying $p^c \leq \max\{1-q^c,0\} < 1$ for all $c \in C$, where the strict inequality follows from Assumption **A3**. Let $\tilde{p}$ be an $N+1$ dimensional vector such that $p < \tilde{p} < 1$.

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 $\theta^c \in \Theta^c$: $p < r^{\theta^c} < r^\theta$ 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 is admitted to her top-ranked program according to her submitted preferences $\tilde{\succ}^\theta$. Therefore, stability implies that $\mu_*(\theta)$ is θ's top-ranked program according to $\tilde{\succ}^\theta$. By the equilibrium hypothesis, it must be that the $\tilde{\succ}^\theta$-maximal program is the same as the $\succeq^{\theta|\mu(\sigma^\theta,\tilde{\succ})}$-maximal program. That is, it must be that, for equilibrium profile $\tilde{\succ}$, 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 $\succeq^{\theta|\mu(\sigma^\theta,\tilde{\succ})}$ for almost all θ with $r^\theta\geq\tilde{p}$. But this contradicts the ongoing assumption that $\eta(L_{\tilde{\succ},\varphi,\tilde{p}})>0$. □

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 $B := \{\theta | (\theta, c) \text{ block } \mu_t \text{ for some } c \in C\}$. 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 be the set of students involved in at least one blocking pair at μ_t. We wish to show that there exists $\epsilon > 0$ such that if $\eta(B) < \epsilon$ then $\|\lambda_t - \lambda_{t+1}\|_\infty < \delta$.

Consider three alternative markets $E_t = [\zeta^{\eta,\mu_{t-1}}, q, N, \Theta^{\mu_{t-1}}]$, $E_{t+1} = [\zeta^{\eta,\mu_t}, q, N, \Theta^{\mu_t}]$, and $E_\gamma = [\zeta^\gamma, q, N, \Theta^\gamma]$. Let $\mathcal{P}^c$ be the set of strict linear orders over C that list only program $c \in C$ above c_0. We define ζ^γ and Θ^γ implicitly as follows for $\gamma \in (0,1)$:

- For any open set $R \subset [0,1]^{N+1}$, any assignment α, and any $\succeq \in \mathcal{P}$: $\zeta^\gamma(\{\theta \in \Theta^\gamma | r^\theta \in R \text{ and } \succeq^{\theta|\alpha} = \succeq\}) = (1-\gamma)\eta(\{\theta \in \Theta \cap B | r^\theta \in R \text{ and } \succeq^{\theta|\mu_{t-1}} = \succeq\})$,
- for any open set $R \subset [0,1]^{N+1}$, any assignment α, and any $\succeq \in \mathcal{P}$: $\zeta^\gamma(\{\theta \in \Theta^\gamma | r^\theta \in R \text{ and } \succeq^{\theta|\alpha} = \succeq\}) = (1-\gamma)\eta(\{\theta \in \Theta \setminus B | r^\theta \in R \text{ and } \succeq^{\theta|\mu_t} = \succeq\})$, and
- for any open set $R \subset [0,1]^{N+1}$, any assignment α, any $c \in C \setminus \{c_0\}$, and any $\succeq \in \mathcal{P}^c$: $\zeta^\gamma(\{\theta \in \Theta^\gamma | r^\theta \in R \text{ and } \succeq^{\theta|\alpha} = \succeq\}) \geq \frac{\gamma}{N}\eta(\{\theta \in \Theta | r^\theta \in R\})$.

That is, ζ^γ specifies student types such that $1-\gamma$ fraction of students selected "uniformly at random" among those involved in blocking pairs at μ_t in market E have the same ordinal preferences as in market E_t and $1-\gamma$ fraction of students selected "uniformly at random" among those not involved in blocking pairs have the same ordinal preferences as in market E_{t+1}. In the limiting market, E_0, those involved in blocking pairs at μ_t in E have the same ordinal preferences as in market E_t and all others have the same ordinal preferences as in E_{t+1}. Let μ_t and μ_{t+1}, be the (unique) stable matchings in E_t and E_{t+1}. Recall that by Remark 3, μ_t and μ_{t+1} are the outcomes of the TIM process at times t and $t+1$, respectively.

We proceed with the proof first by showing that there is a unique stable matching μ_γ in market E_γ which for sufficiently small γ coincides with μ_t for at least $1-\frac{\delta}{2}$ measure of students. Then we show that for sufficiently small γ and ϵ, μ_γ coincides with μ_{t+1} for at least $1-\frac{\delta}{2}$ measure of students. This implies that μ_t coincides with μ_{t+1} for at least $1-\delta$ measure of students, completing the proof.

Lemma A.7. *For any $\gamma \in (0,1)$ there is a unique stable matching μ_γ in market E_γ, and there exists $\gamma^* > 0$ such that for all $\gamma < \gamma^*$, $|\lambda^{c,x}(\mu_t) - \zeta^\gamma(\{\theta \in \mu_\gamma(c) | r^{\theta,c_0} \leq x\})| < \frac{\delta}{2}$ for all $c \in C$ and all $x \in [0,1]$.*

Proof. To see that there is a unique stable matching μ_γ in market E_γ, note that E_γ satisfies Assumption **A3**: $\frac{\gamma}{N}$ satisfies **A3** in E_γ by construction. Therefore, there is a unique stable matching μ_γ in E_γ by Remark 3.

We now show that $|\lambda^{c,x}(\mu_t) - \zeta^\gamma(\{\theta \in \mu_\gamma(c) | r^{\theta,c_0} \leq x\})| \to 0$ in γ for all $c \in C$ and all $x \in [0,1]$. By Lemma A.6, θ blocks matching μ_t in market E (i.e. $\theta \in B$) only if $D_E^\theta(p_t, \lambda_t) \neq D_E^\theta(p_t, \lambda_{t-1})$ (excepting a measure zero set of students for which $D_E^\theta(p_t, \lambda_t)$ or $D_E^\theta(p_t, \lambda_{t-1})$ is not a singleton, by Assumption **A1**), where we index demand by market. Fix $\gamma \in (0,1)$. By construction, $|D_E^c(p_t, \lambda_{t-1}) - D_{E_\gamma}^c(p_t, \lambda_{t-1})| \leq \gamma$. Therefore, by Lemma A.5 $|\lambda^{c,x}(\mu_t) - \zeta^\gamma(\{\theta \in \mu_\gamma(c) | r^{\theta,c_0} \leq x\})| \to 0$ as $\gamma \to 0$ for all $c \in C$ and all $x \in [0,1]$. □

Lemma A.8. *There exists $\gamma^* > 0$ and $\epsilon^* > 0$ such that if $\gamma < \gamma^*$ and $\epsilon < \epsilon^*$ then $|\lambda^{c,x}(\mu_{t+1}) - \zeta^\gamma(\{\theta \in \mu_\gamma(c) | r^{\theta,c_0} \leq x\})| < \frac{\delta}{2}$ for all $c \in C$ and all $x \in [0,1]$.*

Proof. Follows directly from Lemmas A.5 and A.7. □

The previous two Lemmas establish that for sufficiently small ϵ, if $\eta(B) < \epsilon$ then $\|\lambda_t - \lambda_{t+1}\|_\infty < \delta$ as desired. □

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
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.

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 $\tilde{E}$.*

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 $\mathbf{f}^{\alpha(c')}_{\ell,L}$ 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 $\tilde{\alpha}(c')_{\ell,L}\subset\{\tilde{\theta}\in\tilde{\Theta}|f^{\tilde{\theta},c'}-g^{c'}\in\mathbf{f}^{\alpha(c')}_{\ell,L}$ 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 $\tilde{\theta} \in \tilde{\Theta}_L^i$,

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

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 $\tilde{\theta} \in \tilde{\Theta}_L^i$,

$$g^{c'}(\tilde{\lambda}_L^{c'}(\tilde{\Theta}_L^O \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}_L^i)) - g^{c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L) > -\frac{\delta}{2}. \tag{A.13}$$

Combining Equations A.12 and A.13 implies that for sufficiently large L and for all $\tilde{\theta} \in \tilde{\Theta}_L^i$, $f^{\tilde{\theta},c'}(\tilde{\lambda}_L^{c'}(\tilde{\Theta}_L^O \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}_L^i)) + g^{c'}(\tilde{\lambda}_L^{c'}(\tilde{\Theta}_L^O \cup \tilde{\alpha}(c')_L \setminus \tilde{\Theta}_L^i)) > f^{\tilde{\theta},c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L) + g^{c'}(\tilde{\lambda}^{c'}(\tilde{\alpha}(c')_L)$. 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 $\tilde{\Theta}_L^I$ as in the openness proof, and let L be sufficiently large such that $\gamma(\delta) + \frac{1}{\delta} > \sup_{\tilde{\theta} \in \tilde{\Theta}_L^I} f^{\tilde{\theta},c'}(\tilde{\lambda}(\tilde{\alpha}_L(c'))) - f^{\tilde{\theta},c'}(\tilde{\lambda}(\tilde{\alpha}_L(c') \setminus \tilde{\Theta}_L^I)) + \frac{1}{\delta}$.[3] For all $c \neq c'$ let $g^c(\cdot) = 0$, and for any measurable set $\beta \subset \tilde{\mathcal{A}}$ let

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

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

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

$\tilde{\eta}(\tilde{\Theta}^I_L)>0=\tilde{\eta}(\tilde{\Theta}^O)$ 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 $\beta,\beta'\in\tilde{\mathcal{A}}$,

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

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. □

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 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 $T>T^*$. □

C 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

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

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 M of summary statistics $\{s^m(\lambda)\}_{m=1,...,M}$. We claim the peer preferences above cannot be represented via a utility function over $\{s^m(\lambda)\}_{m=1,...,M}$. 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 $\mathcal{A}$.[4] Therefore, it suffices to show that there does

[4]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 WLOG $f^{\theta,c}(\lambda)=f^{\theta,c}(\lambda')+\delta$ 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 $f^{\theta,c}(\lambda)\neq f^{\theta,c}(\lambda'')$).

not exist a function $h^{\theta,c}:[0,1]^M \to \Lambda$ such that $h^{\theta,c}(s^1(\lambda(\alpha)),...,s^M(\lambda(\alpha))) = \lambda^c(\alpha)$ 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 M arguments, it cannot be surjective, implying that there is some α such that $h^{\theta,c}(s^1(\lambda(\alpha)),...,s^M(\lambda(\alpha))) \neq \lambda^c(\alpha)$.

D Alternative explanations for empirical findings

In this section, we discuss why alternative models listed in Section II.C.5—which are not based on peer preferences–are unlikely to explain the observed patterns in our data.

D.1 Mismatch/"fit" preferences

One alternative is that students do not use the PYS to learn about the skill distribution of peers, but rather as an indication of their mismatch or fit with a particular program. It is worthwhile noting, as mentioned in Section II.C.1, that forms of mismatch or "fit" that directly arise from peers through, for example, zero-sum/curved grading, would not challenge our interpretation. That said, it could still be the case that students may be uninformed about a program and interpret the PYS as a signal, for example, of how difficult the material is for them or the prestige of the program.

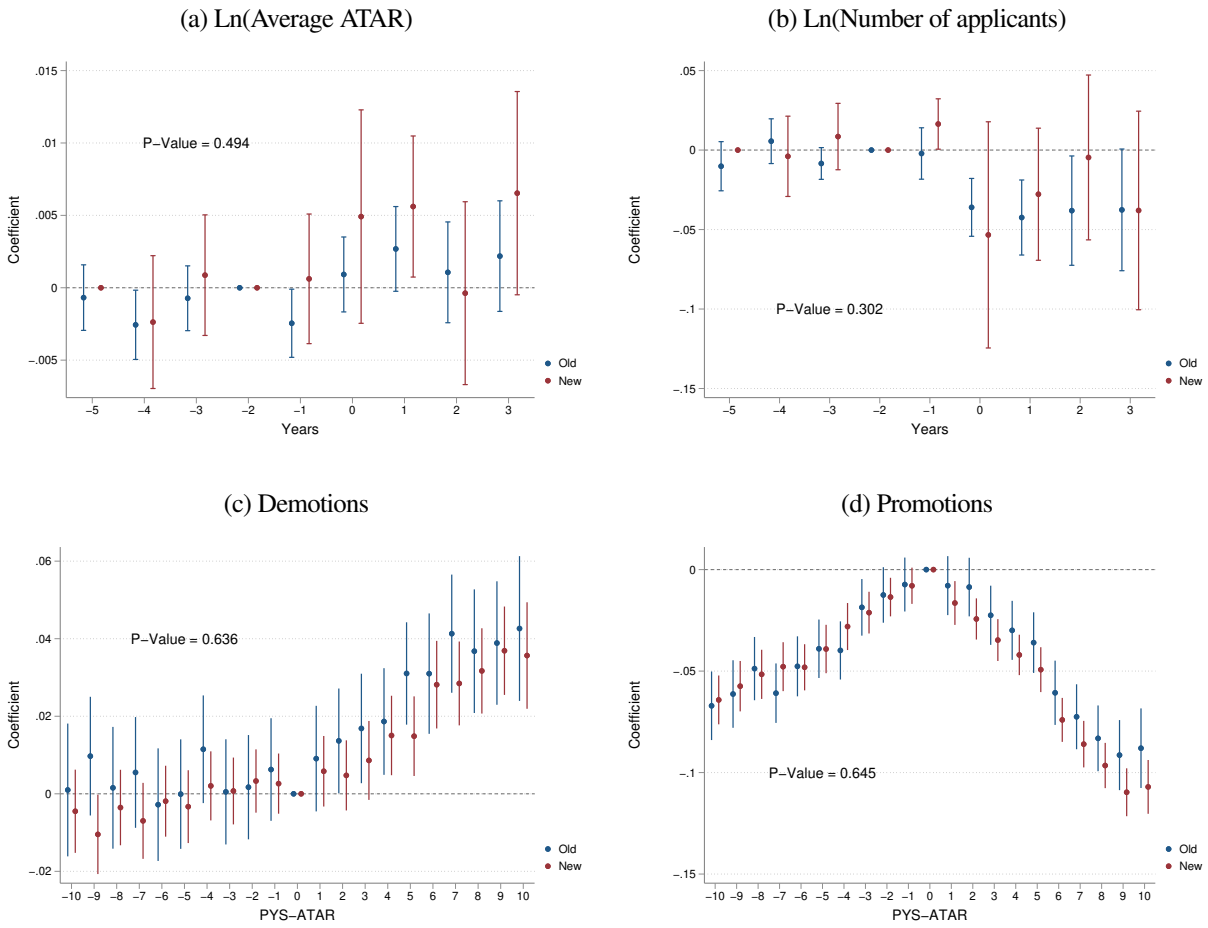
A straightforward prediction is that programs in which students have a stronger prior–perhaps coming from more, or more informative signals excluding the PYS–will be less affected by the signal provided by the PYS. Empirically, we implement a test of the "strength" of the signal provided by analyzing responses to the PYS heterogeneously by program age. This assumes that students have more information about long-standing programs, which we believe is plausible.

We test the potential for our estimated effects to interact with program age for both of our identification strategies. We re-estimate our models after splitting the sample based on whether the sample has been open for all previous 5 years or not (this is the length of the pre-period in the across-person analysis). Students are likely to be less certain for programs that have not been continuously open for the previous few years.

Figure A.1 presents the results for both research designs. In Panels A and B, we show whether the number of students and their average test scores changes are different across program age. We find similar reductions in the number of students who apply and their test scores. We cannot reject an F-test of no difference between these groups across all post-period coefficients for each outcome, separately ($p=0.49$, $p=0.30$). In Panels C and D, we report the within-person analysis separately by program age. As evidenced by the figure, we find neither quantitative nor qualitative differences of effects across these samples. As before, we cannot reject an F-test of differences across the samples, indicating again that students do not appear to be reacting differently across programs with more or less information ($p=0.64$, $p=0.65$).

Both analyses find consistent evidence of non-differential student responses across program

Figure A.1: Heterogeneous effects of PYS by program age



This figure displays regression estimates from the across-person event-study in Panels A and B, and the within-person analysis in Panels C and D. For Panels A and B, regressions estimated according to Equation 1 separately for programs being open in all previous 5 years up to the focal year or not. For Panels C and D, regressions estimated according to Equation 2 separately for programs open in all previous 5 years up to the focal year or not. Figures show point estimates and 95% confidence intervals clustered at the program level (Panels A and B) and additionally individual level (Panels C and D). Joint *p*-values come from joint tests of equality of all within-period (Panels A and B) or within PYS-ATAR (Panels C and D) differences between the all vs not groups.

age/past openness. Given that students likely have more uncertainty about newly-opened programs, our evidence is inconsistent with changes in PYS being used as a relative signal for university difficulty or mismatch.

D.2 Alternatives with "missing mass" prediction

In this section, we describe three alternative models of student behavior proposed in the literature. Common to all of these alternative models is a de facto cost of rejection from a program for students. We first show these models predict a discontinuous jump in the likelihood of applying to programs with PYSs equal to the student's ATAR score. We then discuss evidence from our

empirical analysis, which does not match these predictions.

For these alternative models, 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 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.

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. Let $\Delta = p(0) - \lim_{x \to 0^+} p(x)$.

D.2.1 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 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.2.2 Non-classical preferences

Non-classical utility functions can also explain some non-standard behavior in matching markets. Dreyfuss et al. (2021); Meisner and von Wangenheim (2019) 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 (2021) 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.2.1, 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.2.3 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 (Hakimov et al., 2021; Grenet et al., 2022; Immorlica et al., 2020), where students have incentives not to

"waste" information acquisition costs on programs they will be rejected from.

Prima facie evidence from our within-person analysis does not support the "exogenous factors" hypothesis. From a timing standpoint, only one month separates our observation of the pre- and post-ROLs. Moreover, the fact that adjustments to students' ROLs are predicted by their realized test scores does not support this alternative hypothesis. Specifically, for exogenous preference changes to rationalize our findings, it would have to be that programs with PYSs closer to a student's eventual ATAR score are systematically receiving a positive "shock" relative to other programs.

We further investigate the potential strategic choice of agents 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 $U(\hat{v}^{\theta,c_2})$).[5] 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

[5] 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.

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.

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.

Stylized facts from empirical analysis

All three alternative models find that students are discontinuously more likely to submit ROLs with a slight safety program than with a slight reach program. Our empirical setting with a relatively large number of programs with PYSs distributed richly over the support of student ATAR scores offers a clean test of the these models.

Our first stylized fact is based on prima facie evidence that students frequently list "reach" schools first on their ROLs; as we discuss in Section II.C.2, 75% of students top-rank a program on their ROL where the program's PYS exceeds the student's ATAR and at which they will be rejected without receiving bonus points. Figure A.2 plots the proportion of top-ranked programs on a student's post-ROL by the difference between the program's PYS and the student's ATAR score. We also plot the counterfactual proportion of top-ranked programs if students top ranked each program with a PYS within 10 points of her ATAR score with equal probability. In this simple across-person correlation, we see no evidence of a decrease in students ranking programs just above their ATAR score.

Our second stylized fact comes from Figure 2 where we similarly see no discontinuity in the promotion and demotion probabilities for programs with score gaps just around zero. Again, the modeling in the previous sections implies that, averaging over all students, we should expect discontinuous changes around a zero score gap under these alternative models.

Finally, we view these alternative models as ex-ante less likely in this institutional context. The clearinghouse we study prominently displays information and advice to dispel the types of incorrect beliefs discussed in Section D.2.1.[6] Also, explanations around optimal information acquisition discussed in Section D.2.3 are unlikely to be as important for research designs that exploit switches

[6] See Section II.A for more information.

between existing program preferences, which we employ in Section II.C.4.

E Alternative panel-based research design

Complementing the event-study design in Section II.C.3, we test the impact of observable PYS on ROLs using a two-way fixed effect specification. We estimate regressions of the form:

$$Y_{c,y} = \beta PYS_{c,y} + \alpha_c + \delta_y + \epsilon_{c,y} \tag{A.15}$$

where $Y_{c,y}$ denotes outcomes for program c in year y, such as log average applicant scores. We include year and program fixed effects (α_c and δ_y, respectively) to control for time-invariant characteristics of programs, such as prestige and difficulty, and aggregate time trends. We are interested in β: when a program has a higher PYS, does it attract fewer low scoring students?

The results are presented in Columns (1-3) of Table A.1. When a program's PYS increases by one point, the applicants tend to be higher scoring on average (column 1) and fewer students rank the program on their ROLs (column 2-3). These results of a decrease in applicants and increase in the average ATAR score of applications are entirely consistent with those presented in Section II.C.3, which uses an event-study approach based on commonly evolving group comparisons.

This test assumes students react to the observable PYS due to peers and not to changing program quality or peers from the distant past. To provide information on this, columns (4-6) augment the previous specifications with lagged values of the PYS (from two and three years before the current year). These years have little predictive power and do not change the coefficient from the current year. Consequently, responses to the PYS are not based on a trend of changes in the PYS over time.

Additional Tables and Figures

Table A.1: Across Time Student Response to Program PYS

	(1) Average ATAR	(2) # of Stud.	(3) % of Stud.	(4) Average ATAR	(5) # of Stud.	(6) % of Stud.
Previous Year's Statistic (PYS)	.39*** (.015)	-4*** (.4)	-.0089*** (.00099)	.33*** (.018)	-4.7*** (.51)	-.013*** (.0014)
2 Years Ago Statistic				.062*** (.014)	.14 (.34)	.00061 (.00086)
3 Years Ago Statistics				.041*** (.012)	-.32 (.38)	.002** (.00098)
Dep. Var. Mean	70.26	146.36	.38	70.26	146.36	.38
N	15230	15230	15230	8892	8892	8892

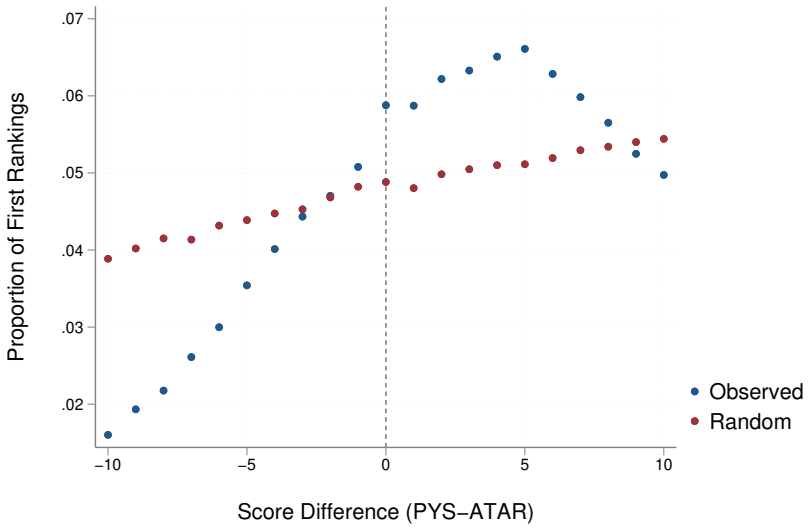
This table presents the relationship between student responses and program PYS. Outcomes include (1) average student score, (2) the number of students who apply, and (3) the percent of students who apply. Columns (4)-(6) present the relationship between student responses and program PYS in the three years prior. Program and year fixed effects are included. Standard errors in parentheses, clustered at the program level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table A.2: Measuring Instability using Counterfactual Enrollment

	$(\hat{g}-\hat{f})\times 100{,}000$		
	(1) Bonus = 3	(2) Bonus = 3 + Random(0-7)	(3) Bonus = 7
CYS - PYS	-.27***	-.25***	-.32***
	(.053)	(.041)	(.043)
CYS - PYS, OO 1	.028	-.035	.0025
	(.052)	(.04)	(.043)
CYS - PYS, OO 2	.0074	.023	.0079
	(.05)	(.038)	(.04)
CYS - PYS, OO 3	-.084*	-.071*	-.055
	(.049)	(.037)	(.039)
CYS - PYS, OO 4	-.048	-.063*	-.11***
	(.05)	(.038)	(.04)
CYS - PYS, OO 5	-.13***	-.044	-.037
	(.051)	(.039)	(.042)
Dep. Var. Mean	-2.28	-1.92	-1.99
N	26683	31006	30431

This table presents the relationship between instability and score differences. CYS - PYS is the difference in the score of the current and previous year statistics. OO 1-5 are the CYS - PYS for the top 5 outside option programs. Bonuses are assigned using three regimes: (1) bonus = 3 for all applications, (2) bonus = 3 + a randomly assigned from a uniform distribution over integers 0-7 at the student × program level for all applications, and (3) bonus = 7 . Program, year, and ATAR score fixed effects are included. Standard errors clustered at the program level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Figure A.2: Proportion of First-Ranked Programs, by Score Gap



This figure plots the relationship between the program-student score difference and student application choices. In blue, we the points correspond to the proportion of students who rank a program first on their list with this difference. For reference, the red points plot the relationship between the program-student score difference and the proportion of students who would rank a program first if they randomly chose programs. The program-student score difference is the gap between the top-ranked program's PYS and the student's ATAR score. The sample is restricted to students who first-rank a program within this range. Students can receive up to 10 bonus points, so there is a positive probability of admission to all programs with score gaps in the presented range.