# Essentially Stable Matchings*

Peter Troyan[†]
Department of Economics
University of Virginia

David Delacrétaz[‡]
Department of Economics
University of Melbourne

Andrew Kloosterman[§]
Department of Economics
University of Virginia

This draft: July 2, 2018
First draft: July 18, 2016

**Abstract**

We propose a solution to the conflict between fairness and efficiency in matching markets. A matching is *essentially stable* if any priority-based claim initiates a chain of reassignments that results in the initial claimant losing the object (i.e., the claim is *vacuous*). We study the structure of the essentially stable set, and classify popular Pareto efficient mechanisms using our criterion: those based on Shapley and Scarf's TTC mechanism are not essentially stable, while Kesten's EADA mechanism is. Besides reconciling the conflict with efficiency, vacuous claims are simple and straightforward to explain, making essential stability well-suited for practical applications.

**JEL Classification:** C78, D47, D61, D63, I20
**Keywords:** matching, stability, fairness, efficiency, school choice

---

*Authors' names appear in random order. We thank (also in random order) Umut Dur, Thayer Morrill, Alex Teytelboym, Mustafa Oğuz Afacan, Martin van der Linden, Georgy Artemov, Bettina Klaus, and Samson Alva for helpful discussions regarding this project.

[†]P.O. Box 400182, Charlottesville, VA, 22904. troyan@virginia.edu.
[‡]111 Barry Street, The University of Melbourne, VIC 3010. ddelacretaz@unimelb.edu.au
[§]P.O. Box 400182, Charlottesville, VA, 22904. ask5b@virginia.edu

# 1 Introduction

There exists a trade-off between efficiency and fairness in matching problems. The celebrated deferred acceptance (DA) mechanism of Gale and Shapley (1962) always produces a fair matching, and in fact produces the most efficient matching among all fair ones. However, it does not always produce a Pareto efficient matching: there may be unfair matchings that Pareto dominate it. Are all unfair matchings equally unfair? We argue that the answer is no, and in fact the fairness criterion typically used in the literature excludes many matchings unnecessarily. We propose a new fairness criterion that takes these matchings into account and is not at odds with efficiency.

There are many real-world examples of problems that fit into our framework, but perhaps the largest and most important is public school choice as instituted in many cities across the United States and around the world. Fairness is a crucial concern for many school districts because they must be able to justify why one student is admitted to a school and another is rejected, or else be vulnerable to legal action. This is typically done by assigning priorities to each student at each school according to some set criteria (which may vary across school districts) and then running a well-defined matching mechanism that takes these priorities and the student preferences as inputs.

In this framework, the standard approach in the literature is to use the mathematical definition of *stability* as a formal fairness criterion.[1] Given a matching, a student is said to have a (justified) *claim* to school $A$ if she prefers $A$ to her assignment and she has higher priority than another student who is assigned to $A$. A matching is *stable* if there are no claims. In other words, stable matchings are "fair" because they *eliminate justified envy* (Abdulkadiroğlu and Sönmez, 2003), and a school district can easily explain why some student $j$ was not admitted to a school $A$ (even though she prefers it) and another student $i$ was: $i$ has higher priority than $j$ at $A$.

At first glance, the classic use of stability as the standard for fairness seems very reasonable, because it ensures that there are no claims. However, this simple definition actually misses a subtle (and important) point: if a student were to have a claim to a seat at a school, granting her claim displaces a student currently assigned to that school. This student will

[1]In fact, some authors simply define the word "fair" to mean "stable" (see, e.g., Balinski and Sönmez (1999)). In this paper, we use the term in its normative sense, and will argue that the definition of fairness as being equivalent to stability is too restrictive: there are matchings that are not stable according to the formal definition, but are fair according to a reasonable understanding of the term. See also footnote 11 in Section 2.

then have to be reassigned, and, using the same justification as the initial student, she can claim her favorite school at which she has high enough priority. This will displace yet another student, and so on. Eventually, this chain of reassignments will end when some student is reassigned to a school with an empty seat (or determines that everything that is available is unacceptable to her and takes her outside option). It is possible that the student who made the initial claim will be displaced by some student further down the chain. Since the initial student will ultimately not receive the school to which she laid claim, her claim actually is not justified, but rather is *vacuous*.

The widespread success of DA as an assignment mechanism suggests that fairness is indeed important in many real-world markets.[2] However, the efficiency losses from eliminating all claims (e.g., by running a stable mechanism such as DA) can be significant in practice. For example, using data from eighth-grade assignment in New York City, Abdulkadiroğlu et al. (2009) show that, by moving from the DA matching to a Pareto efficient matching, over 4,000 students could be made better off each year (on average) without making a single student worse off. We thus propose a new definition that expands on the set of stable matchings by allowing some claims to remain: namely, those that are vacuous. While this is a weaker fairness criterion, we argue that it captures the essential feature of using stability as a fairness standard, in the sense that if a student were to try to assert her claim, we can clearly explain to her why doing so will be for naught. For example, if some student $i$ were to claim a seat at school $A$, one could convince her to relinquish this claim by walking her through its ultimate effects in the following way:

> "Yes, I agree that your claim at $A$ is valid because you have higher priority than student $j$. But, if I assign you to $A$, then $j$ will need to be reassigned, and she will claim $B$, and her claim at $B$ is equally as valid as yours was at $A$. This will displace student $k$, who will then claim $A$, and, because she has higher priority than you at $A$, I will need to give it to her. So, while you may claim $A$, ultimately, you will not receive it anyway."

For this reason, we call a matching in which all claims are vacuous an *essentially stable* matching.

Imposing the less stringent constraint of essential stability allows us to implement a larger set of fair matchings. In contrast to the stable set, we show (Corollary 2) that the essentially

[2]See Abdulkadiroğlu et al. (2005a) and Abdulkadiroğlu et al. (2005b) for discussions of implementing DA in New York City and Boston, respectively.

stable set always contains at least one Pareto efficient matching, i.e., a Pareto efficient and essentially stable (PEES) matching always exists. In other words, essential stability provides a solution to the trade-off between fairness and efficiency, as there is always a matching that possesses both properties.

We analyze the structure of the set of essentially stable matchings and show that it retains some, though not all, of the properties that the stable set possesses. One of the most well-known results in matching theory in that the set of stable matchings forms a lattice, which in particular implies that there exists a student-optimal stable matching and a student-pessimal stable matching. We show that the latter is also the student-pessimal *essentially* stable matching (Theorem 1). However, the student-optimal stable matching is in general not Pareto efficient, and Corollary 2 (see the previous paragraph) implies that the symmetric property will not hold: the student-optimal stable matching is not necessarily the student-optimal essentially stable matching. Further, we show that there may be multiple PEES matchings, i.e., there may not be a unique student-optimal essentially stable matching. Last, we show that one of the other most famous results of stable matching theory, the Rural Hospital Theorem (Roth, 1986), does extend to the set of essentially stable matchings: all schools that are not filled to capacity at some essentially stable matching are assigned the same set of students at all of them (Theorem 2).

After defining essential stability and analyzing the structure of the essentially stable set, we last turn to mechanisms for implementation. The ultimate reason for relaxing stability in the first place is to achieve the Pareto frontier, and so we consider the three main Pareto efficient mechanisms that have been proposed in the literature: the top trading cycles (TTC) mechanism of Shapley and Scarf (1974), extended to school choice settings by Abdulkadiroğlu and Sönmez (2003); the DA+TTC mechanism that first runs DA and then searches for "improvement cycles" over the DA matching; and the efficiency-adjusted deferred acceptance (EADA) mechanism of Kesten (2010). We show that neither TTC nor DA+TTC are essentially stable, i.e., they are *strongly unstable*. EADA, on the other hand, is essentially stable; therefore it produces a PEES matching.[3]

While we have studied the three most commonly-proposed Pareto efficient mechanisms in the literature, there are others. In fact, we show that there may exist multiple PEES matchings for any given problem (and hence, multiple PEES mechanisms as well). Thus, while the main contribution of this paper is to solve the conflict between fairness and Pareto

[3]Readers familiar with the EADA mechanism will know that it is Pareto efficient so long as all students "consent". This will be discussed in more detail below.

efficiency in a simple and intuitive way, it also raises the question of whether some PEES matchings can be argued as more desirable than others in a meaningful way. One criterion that could potentially discriminate amongst PEES mechanisms would be strategyproofness, the property that truthfully revealing one's preferences is a dominant strategy, but this criterion is well-known to be quite difficult to satisfy and the results of Alva and Manjunath (2017) apply here to show that the only essentially stable and strategyproof mechanism is DA (Theorem 4). Determining how to select the "best" PEES matching for any particular problem is an interesting direction for future work.

**Related Literature**

Most closely related to our paper is a growing literature that investigates weaker definitions of stability that are compatible with efficiency. Most of the papers in this literature are loosely based on the idea that a student with a claim must propose an alternative matching that is free of any counter-claims (and possibly some other conditions too) or else her initial claim can be disregarded. Work in this vein includes Alcalde and Romero-Medina (2015), who introduce the concept of $\tau-$*fairness*, and Cantala and Pápai (2014), who introduce the concepts of *reasonable stability* and *secure stability.* Morrill (2016) introduces the concept of a *legal assignment,* where, in legal terminology, a student $i$'s claim at a school $c$ is not redressable (and thus can be disregarded) unless $i$ can propose an alternative assignment (i.e., matching) that is "legal" and at which she is assigned to $c$ (see also Ehlers and Morrill (2017)). He introduces an iterative procedure for finding the set of legal assignments, which he shows is equivalent to the *von Neumann-Morgenstern stable set* (Von Neumann and Morgenstern, 1944).[4] Tang and Zhang (2016) introduce their own new definition of weak stability for school choice problems that is also closely related to vNM stable sets. While sharing a similar motivation, all are independent properties, and in Appendix B, we show formally that these other concepts are distinct from ours.

A different strand of related literature focuses on mechanisms (rather than matchings), and in particular on a class of mechanisms that, besides simply asking students to report their preferences, also asks them if they "consent" to having their priority violated. The goal is then to use a mechanism that ensure students cannot gain from not consenting, or in other words the mechanism should be *no-consent-proof.* This is the original approach taken in Kesten's paper introducing the EADA mechanism, and was further expanded by

[4]Ehlers (2007) studies vNM stable sets in the context of marriage markets (see also Wako (2010)).

Dur et al. (2015), who show that EADA is in fact the unique constrained efficient mechanism that Pareto dominates DA and is no-consent-proof.[5] The reason that EADA is no-consent-proof is that a student's own assignment is unaffected by her consent decision, and so all students are actually indifferent between consenting and not consenting. While related, there is an important conceptual distinction between the approaches. Essential stability is a novel fairness criterion that relaxes stability; no-consent-proofness is a procedural justification for why, given a particular mechanism, students should affirmatively consent to violations of the classical definition of stability. In other words, essential stability is a property of matchings, while no-consent-proofness is a property of mechanisms, which are more complex objects.[6]

We believe that a main advantage of essential stability is the ease with which it can be explained to non-experts, which makes it particularly well-suited for practical applications. Essential stability puts a simple condition on any given matching: all claims (if any) must be vacuous. The distinction between vacuous and non-vacuous claims can, in turn, be made clear to policymakers and market participants by fixing a particular matching and walking them through a reassignment chain. As such, our approach to justifying why some claims are not valid could, for example, be easily (visually) communicated in a simple information brochure for policymakers and/or parents. In contrast, convincing policymakers to use a no-consent-proof mechanism requires explaining the inner workings of the mechanism in detail and ensuring that all market participants understand precisely why their consent decision is irrelevant to their own eventual outcome, no matter what everyone else reports (and even then, students may still choose not to consent, since consenting does not benefit them). Other relaxations of stability are also more difficult to explain to non-experts as the conditions they specify are not directly on a matching. For example, to determine whether a certain priority violation is *legal* in the sense of Morrill (2016) or Ehlers and Morrill (2017), one must first construct the entire set of legal matchings, and then check that any alternative matching where the student is assigned to the school with which she blocks lies outside of this "legal set". At the same time, while our approach is arguably more practical, the legal set possesses elegant mathematical structure and provides a convincing argument for relaxing stability in

[5] Dur et al. (2015) also have a relaxation of stability called *partial stability* that exogenously takes a subset of priority violations as allowable. A partially stable matching is then constrained efficient if it is not Pareto-dominated by any other partially stable matching.

[6] Of course, the definition of essential stability can be easily extended to mechanisms in the natural way by defining a mechanism as essentially stable if it always produces an essentially stable matching. While it so happens that EADA is both no-consent-proof and essentially stable, there is no a priori logical relationship between these properties: there are no-consent-proof mechanisms that are *not* essentially stable, as well as essentially stable mechanisms that are *not* no-consent-proof.

a way that is compatible with Pareto efficiency. In fact, many legal matchings are also essentially stable (e.g., EADA) and we view the two concepts, as well as the others in this literature, as complementary.

From a broader perspective, our paper also contributes to a growing literature on how to define stability when agents may anticipate more than one step of blocking, a question that has received considerable attention in other game-theoretic contexts. The central concept in this literature is called farsightedness: an outcome is stable if there does not exist a series of blocks that culminate in better outcomes for every agent who participates in it. Farsightedness was first introduced by Harsanyi (1974) as a criticism of von Neumann-Morgenstern stable sets. In more recent work, Ray and Vohra (2015) and Dutta and Vohra (2017) carefully address farsightedness in coalition formation games. Page Jr et al. (2005) and Herings et al. (2009) consider related issues in network formation games. While similar in the sense that both look more than one step ahead, the ultimate effect of essential stability is actually opposite that of farsightedness: farsightedness excludes myopically stable outcomes by providing a series of blocks that makes the initial outcome ultimately unstable, while essential stability includes myopically unstable outcomes by showing that a series of reassignments nullifies the original block. This allows expending the set of admissible matchings in order to reach the Pareto frontier.

The remainder of the paper is organized as follows. We formally introduce essential stability in Section 2. Section 3 is devoted to the set of essentially stable matchings and its structure. In Section 4, we study whether well-known mechanisms are essentially stable. Section 5 concludes. All proofs not in the main text can be found in the appendix.

# 2 Preliminaries

## 2.1 Model

There is a set of students $S$ who are to be assigned to a set of schools (or "colleges") $C$. Each $i \in S$ has a strict preference relation $P_i$ over $C$ and each $c \in C$ has a strict priority relation $\succ_c$ over $S$. Let $P = (P_i)_{i \in S}$ denote a profile of preference relations, one for each student, and $\succ = (\succ)_{c \in C}$ denote a profile of priority relations.[7] Each $c \in C$ has a capacity $q_c$, which is the number of students that can be assigned to it. Let $q = (q_c)_{c \in C}$ denote a profile

[7]We also compare the priority of sets of students based on *responsiveness* (see Roth (1985)). For any $I \subset S$, any $i, j \in S \setminus I$, and any $c \in C$, $I \cup \{i\} \succ_c I \cup \{j\}$ whenever $i \succ_c j$.

of capacities. We assume $S$, $C$, $\succ$, and $q$ are fixed throughout the paper, and associate a **market** with its preference profile $P$. For concreteness, we use the school choice terminology throughout as it is the best-known application; however, the model can be applied to many other real-world assignment problems. Examples include the military assigning cadets to branches, business schools assigning students to projects, universities assigning students to dormitories, or cities assigning public housing units to tenants.[8]

A **matching** is a correspondence $\mu : S \cup C \to S \cup C$ such that, for all $(i, c) \in S \times C$, $\mu(i) \in C$, $\mu(c) \subseteq S$, $|\mu(c)| \leq q_c$, and $\mu(i) = c$ if and only if $i \in \mu(c)$.[9] A matching $\nu$ **Pareto dominates** a matching $\mu$ if $\nu(i) R_i \mu(i)$ for all $i \in S$, and $\nu(i) P_i \mu(i)$ for at least one $i \in S$.[10] A matching $\mu$ is **Pareto efficient** if it is not Pareto dominated by any other matching $\nu$. Note that Pareto efficiency is evaluated only from the perspective of the students $S$, and not the schools $C$. This is a standard view in the mechanism design approach to school choice, beginning with the seminal papers of Balinski and Sönmez (1999) and Abdulkadiroğlu and Sönmez (2003).

In addition to Pareto efficiency, in many applications (particularly in school choice), market designers also care about fairness. The most common fairness criterion is *stability*.[11] Given a matching $\mu$, we say student $i$ **claims a seat at school** $c$ if (i) $cP_i\mu(i)$ and (ii) either $|\mu(c)| < q_c$ or $i \succ_c j$ for some $j \in \mu(c)$. We will sometimes use $(i, c)$ to denote $i$'s claim to $c$. If no student claims a seat at any school, then we say $\mu$ is **stable**. Stability is a fairness criterion in the sense that it ensures priorities are respected: a student only misses out on a school she wants if that school is filled to capacity with higher-priority students.

[8]For more detail on these markets, see Sönmez (2013), Sönmez and Switzer (2013), Fragiadakis and Troyan (2016), Chen and Sönmez (2002), Chen and Sönmez (2004), Sönmez and Ünver (2005), Sönmez and Ünver (2010), and Thakral (2015).

[9]We assume all students are assigned to a school and vice-versa. While in practice some students may prefer taking an "outside option" to some schools, our model is without loss of generality, as we could simply model the outside option as a particular school $o \in C$ with capacity $q_o \geq |S|$.

[10]$R_i$ denotes the weak part of $i$'s preference relation $P_i$. Given the assumption that preferences are strict, $aR_ib$ but not $aP_ib$ if and only if $a = b$. We analogously denote by $\succeq_c$ the weak part of $c$'s priority relation $\succ_c$.

[11]The term stability was first introduced by Gale and Shapley (1962) and is standard in the two-sided literature. In one-sided matching models such as school choice, some papers use other terminology such as "fairness" (Balinski and Sönmez, 1999) or "elimination of justified envy" (Abdulkadiroğlu and Sönmez, 2003) to refer to concepts that are mathematically equivalent to stability. To avoid confusion, we stick with "stability" for our formal definition because we think it is the most familiar, and reserve the word "fairness" for the normative concept. See also footnote 1.

## 2.2 Motivating Example

Deferred acceptance (DA) is one of the benchmark mechanisms that form the foundation for both the theory and practice of a myriad of matching markets.[12] DA is an enormously successful mechanism in the field because it produces the *student-optimal stable matching:* that is, the DA outcome is stable, and, for any other stable matching $\nu$, $\nu$ is Pareto dominated by $\mu^{DA}$. The prevalence of DA in the field suggests that stability plays an important role in many settings as a fairness standard. However, as discussed in the introduction, eliminating all claims comes at the price of efficiency: DA is not a Pareto efficient mechanism. The example below illustrates this point, and serves to motivate our new definition. Throughout the paper, given preferences $P$, the matching produced by the DA mechanism is denoted as either $DA(P)$, or, if the preferences are understood, as $\mu^{DA}$. Student $i$'s assigned school under the DA matching is denoted as either $DA_i(P)$, or, if the preferences are understood, as $\mu^{DA}(i)$.

**Example 1.** Let there be 5 students, $S = \{i_1, i_2, i_3, i_4, i_5\}$, and 5 schools with capacity 1, $C = \{A, B, C, D, E\}$. The priorities and preferences are given in the following tables.

| $\succ_A$ | $\succ_B$ | $\succ_C$ | $\succ_D$ | $\succ_E$ |
|---|---|---|---|---|
| $i_1$ | $i_2$ | $i_3$ | $i_4$ | $\vdots$ |
| $i_2$ | $i_3$ | $i_4$ | $i_5$ | |
| $i_4$ | $i_1$ | $i_2$ | $i_3$ | |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | |

| $P_{i_1}$ | $P_{i_2}$ | $P_{i_3}$ | $P_{i_4}$ | $P_{i_5}$ |
|---|---|---|---|---|
| $\dagger B$ | $\dagger C*$ | $B*$ | $\dagger A$ | $D$ |
| $\boxed{A*}$ | $A$ | $\dagger D$ | $C$ | $\boxed{\dagger E*}$ |
| $\vdots$ | $\boxed{B}$ | $\boxed{C}$ | $\boxed{D*}$ | $\vdots$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | |

The table on the right indicates three different potential matchings, a matching $\mu^{\square}$ (denoted by boxes $\square$), and two Pareto efficient matchings $\mu^*$ (denoted by stars $*$) and $\mu^\dagger$ (denoted by daggers $\dagger$). The DA matching $\mu^{DA}$ in this example is $\mu^{\square}$, which therefore can readily be shown to be stable. It is, however, not Pareto efficient: it is easy to see that it is Pareto dominated by both $\mu^*$ and $\mu^\dagger$. This of course implies that $\mu^*$ and $\mu^\dagger$ are both unstable. For instance, at $\mu^*$, student $i_4$ claims the seat at $C$ (because $i_4 \succ_C i_2 = \mu^*(C)$) and at $\mu^\dagger$, student $i_3$ claims the seat at $B$ (because $i_3 \succ_B i_1 = \mu^\dagger(B)$) and student $i_5$ claims the seat at $D$ (because $i_5 \succ_B i_3 = \mu^\dagger(D)$).

[12]In particular, we consider the student-proposing version of the DA mechanism. DA is very well-known, and so we do not provide a formal definition here. Such a definition can be found in (for example) Gale and Shapley (1962) or Abdulkadiroğlu and Sönmez (2003).

There are simpler examples to show that the DA matching may not be Pareto efficient. We present this one to illustrate the main point of our paper, which is that not all instability is the same. We argue that $\mu^*$ is truly unstable while $\mu^\dagger$ is not.

To understand our argument, consider $\mu^\dagger$ first. Suppose student $i_3$ claims the seat at school $B$. If we grant $i_3$'s claim and assign her to $B$, then student $i_1$ becomes unmatched. Student $i_1$ must be assigned somewhere, and (using the same logic as $i_3$), she can ask to be assigned to $A$, her next most-preferred school where she has higher priority than the student who is matched to it (student $i_4$). Granting $i_1$'s claim just as we did $i_3$'s, she is assigned to $A$ and now student $i_4$ is unmatched. Student $i_4$ then asks for $C$, which is her most preferred school where she has high enough priority to be assigned. Student $i_2$ is now unmatched, and asks for $B$,[13] which means student $i_3$ is removed from $B$. In summary, student $i_3$ starts by claiming $B$. If her request is granted based on the fact that $i_3 \succ_B i_1$, then we must also grant the next request of $i_1$, since she has the same justification for claiming $A$ as $i_3$ did for claiming $B$. Continuing, we see that ultimately another student with higher priority than $i_3$ at $B$ (in this case $i_2$) ends up claiming it, and so $i_3$'s initial claim is unfounded, or as we will call it, vacuous. Student $i_5$'s claim to $D$ also begins a chain of reassignments where eventually $i_4$ takes $D$ away, so $i_5$'s claim is also vacuous.

Now let us contrast this with the instability found in matching $\mu^*$. Assume that student $i_4$ claims the seat at $C$, and this request is granted. Following similar logic to the above, $i_2$ then asks for $B$, and $i_3$ asks for $D$. This is the end of our reassignments because $D$ is the school that $i_4$ gave up to claim $C$. In this case, the original claimant's (student $i_4$) request does *not* result in her ultimately losing the school she claimed to a higher priority student, and therefore this claim is not vacuous in the manner that the claims at $\mu^\dagger$ were.

Thus, both $\mu^*$ and $\mu^\dagger$ are unstable, but in different ways (which will be made more precise in a moment). While student $i_3$ can protest $\mu^\dagger$ and request $B$, if she does so, she will ultimately be rejected from $B$. What is more, while in practice students may not fully understand the workings of the mechanism, if a student were to protest, it would be very easy to walk her through the above chain of reassignments to show her that granting her request would ultimately not be beneficial to her, and therefore convince her to relinquish her claim. Thus, to the extent that the goal of imposing stability is to prevent students from protesting their assignment, matchings like $\mu^\dagger$, while not fully stable in the classical sense, are "essentially" stable.

[13]Note that her next most preferred school is $A$, but school $A$ is now assigned to $i_1$ and $i_1 \succ_A i_2$, so she cannot get $A$ and must go to $B$.

On the other hand, the instability of matching $\mu^*$ is a much stronger type of instability, because $i_4$ will ultimately benefit from claiming $C$, and so we would not be able to convince her to relinquish her claim. Our new definition of stability is designed to capture this idea and, in the process, recover inefficiencies by expanding the set of permissible matchings to include those like $\mu^\dagger$, but still exclude those like $\mu^*$.

## 2.3 Essentially Stable Matchings

We now formalize the intuition from the previous example. Recall that, fixing a matching $\mu$, we use the notation $(i, c)$ to denote $i$'s claim to a seat at $c$.

**Definition 1.** Consider a matching $\mu$ and a claim $(i, c)$. The **reassignment chain initiated by claim** $(i, c)$ is the list

$$i^0 \to c^0 \to i^1 \to c^1 \to \cdots \to i^K \to c^K$$

where,

- $i^0 = i$, $\mu^0 = \mu$, $c^0 = c$ and for each $k \geq 1$:
- $i^k$ is the lowest-priority student in $\mu^{k-1}(c^{k-1})$,
- $\mu^k$ is defined as: $\mu^k(j) = \mu^{k-1}(j)$, for all $j \neq i^{k-1}, i^k$, $\mu^k(i^{k-1}) = c^{k-1}$, and student $i^k$ is unassigned,[14]
- $c^k$ is student $i^k$'s most preferred school where she can claim a seat at $\mu^k$,
- and terminates at the first $K$ such that $|\mu^K(c^K)| < q_{c^K}$.

We should note a few features of reassignment chains. First, a student $i$ may claim a seat at a school $c$ either because $c$ is not filled to capacity, or because she has higher priority than someone currently assigned there. Whenever the former occurs, that is the end of the reassignment chain (if this occurs immediately, then the original matching $\mu$ was wasteful). Second, when a student is rejected, she can claim a seat at a school she prefers to the school

[14]Note that $\mu^k$ does not fully satisfy our definition of a matching since student $i^k$ is unassigned. However, this is not important, as $i^k$ is immediately reassigned in the next step.

that rejected her.[15] Last, note that a school or a student can appear multiple times in a reassignment chain.[16]

For a reassignment chain $\Gamma$ started by a claim $(i, c)$, if there exists $k \neq 0$ such that $i^k = i$, we say that the reassignment chain **returns to** $i$. If the reassignment chain returns to $i$, then $i$ will ultimately be removed from the school $c$ that she claimed initially by some student with higher priority. When this is the case, we say that claim $(i, c)$ is **vacuous**.

**Definition 2.** Matching $\mu$ is **essentially stable** if all claims at $\mu$ are vacuous. If there exists at least one claim at $\mu$ that is not vacuous, $\mu$ is **strongly unstable**.

A mechanism $\psi$ is said to be essentially stable if $\psi(P)$ is an essentially stable matching for all $P$. If $\psi$ is not essentially stable, then we say it is strongly unstable.

We defined a claim $(i, c)$ as vacuous as soon as the induced reassignment chain returns to $i$ and ultimately rejects her from $c$. However, one might argue that maybe $i$ claims a seat at a very good school $c$ and, while she is rejected from $c$ in the reassignment chain, perhaps $i$'s final assignment at the end of the chain is still better than $i$'s original school. A possible alternative would be to define $(i, c)$ as vacuous if $i$ ends up matched to her original school $\mu(i)$ at the end of the reassignment chain. The next result shows that this would not affect Definition 2:[17]

**Proposition 1.** *For any essentially stable matching $\mu$, the reassignment chain initiated by any (vacuous) claim $(i, c)$ ends with $i$ matched to $\mu(i)$.*

The intuition for Proposition 1 is that all of $i$'s claims, which are vacuous at $\mu$ by definition, remain vacuous throughout $\Gamma$ so that any school that $i$ prefers to $\mu(i)$ ends up rejecting her.

---

[15]This and the fact that they start with a student makes reassignment chains slightly different from the notion of rejection chains.

[16]Note also that reassignment chains are well-defined (i.e., they must end in finite steps). This is because, if a student $i^k$ is rejected from a school $c^{k-1}$, then $\mu^k(c^{k-1})$ must be filled to capacity with higher priority students. For all $k' > k$, the lowest priority student in $\mu^{k'}(c^{k-1})$ only increases, and thus, if $i^k$ is ever rejected again later in the chain, the best school at which she can claim a seat is ranked lower than $c^{k-1}$. Since students only apply to worse and worse schools as the chain progresses and preference lists are finite in length, the chain must end.

[17]Another possible concern one might have is that counter-claims along a reassignment chain may themselves be vacuous, which would undermine the argument for disregarding the initial claim. However, as we show in Appendix D, removing these vacuous counter-claims does not affect essential stability, and thus we are justified in working directly with the simpler definition given here. We thank an anonymous referee for raising this point.

Returning to Example 1, we can check that $\mu^\dagger$ is essentially stable, while $\mu^*$ is strongly unstable. As we showed above, at $\mu^\dagger$, the claims $(i_3, B)$ and $(i_5, D)$ are vacuous, because they ultimately result in the initial claimant losing the seat she claimed to a higher priority student. At $\mu^*$, on the other hand, the reassignment chain initiated by $(i_4, C)$ ends with $i_4$ assigned to $C$. Thus, $i_4$'s claim is not vacuous.

# 3 Structure

The set of stable matchings contains a large amount of structure. Two of the canonical results in the theory of stable matching are (1) the stable set is a lattice, which implies the existence of a student-optimal stable matching $\mu^o$ and a student-pessimal stable matching $\mu^p$,[18] and (2) those schools that are not filled to capacity are assigned the same set of students in all stable matchings, a property known as the *Rural Hospital Theorem* (Roth, 1986). In this section, we study the structure of the essentially stable set. We show that it retains some, though not all, of the structure that exists with stable matchings: it does not form a lattice and there may not exist a student-optimal essentially stable matching; however, $\mu^p$ is always the student-pessimal essentially stable matching and the Rural Hospital Theorem holds.

To prove these results, it is useful to first introduce the idea of **rotations**, whose properties are key to our analysis. Given an essentially stable matching $\mu$ and a (vacuous) claim $(i, c)$ at $\mu$, we define the rotation initiated by $(i, c)$ at $\mu$ to be the part of the reassignment chain up to the point where it returns to $i$ (the chain returns to $i$ by definition since the claim is vacuous). That is, the rotation $\hat{\Gamma}$ initiated by $(i, c)$ at $\mu$ is the list

$$i = i^0 \to c = c^0 \to i^1 \to c^1 \to i^2 \to c^2 \to \cdots \to i^{\hat{K}} \to c = c^{\hat{K}} \to i.$$

We denote by $\mu^{\hat{\Gamma}}$ the matching obtained when the rotation is **carried out**: $\mu^{\hat{\Gamma}}(i) = \mu(i)$ and $\mu^{\hat{\Gamma}}(j) = \mu^{\hat{K}+1}(j)$ for all $j \neq i$. In words, carrying out the rotation initiated by $(i, c)$ at $\mu$ means following the reassignment chain up to the point where it returns to $i$ and match $i$ to her original school. We say that a student $j$ **appears** in $\hat{\Gamma}$ if there exists $k = 1, \ldots, \hat{K}$

[18] $\mu^o$ is a matching that weakly Pareto dominates all stable matchings and $\mu^p$ is a matching that is weakly Pareto dominated by all stable matchings; that is, for any stable matching $\mu$ and any student $i$, $\mu^o R_i \mu R_i \mu^p$. In addition, we should also note that the opposite holds for the schools: for any stable matching $\mu$ and any school $c$, $\mu^p \succeq_c \mu \succeq_c \mu^o$ (for this reason, $\mu^p$ is sometimes also called the school-optimal stable matching and $\mu^o$ the school-pessimal stable matching).

such that $j = i^k$ (note that by this definition, $i$ does not appear in $\hat{\Gamma}$). Similarly, school $d$ appears in $\hat{\Gamma}$ if there exists $k = 1, \ldots, \hat{K}$ such that $d = c^k$.[19]

We now present three important properties of rotations, which as we will show, turn out to have important consequences for the structure of the essentially stable set. First of all, the following lemma shows that students and schools that appear in a rotation are affected in a monotonic way.

**Lemma 1.** *For every student $j$ and every school $d$ that appear in $\hat{\Gamma}$:*

$$\mu(j) P_j \mu^{\hat{\Gamma}}(j) \quad \text{and} \quad \mu^{\hat{\Gamma}}(d) \succ_d \mu(d).$$

The statement related to schools is straightforward. At each step, a school replaces a student by another with a higher priority; therefore by the end of the rotation it is assigned a set of students with a higher priority overall.[20] The statement related to students is not as obvious. A student is assigned to her favorite school to which she has a claim and may prefer that school to the one that just rejected her. However, as we show in the proof, that school always rejects her before the end of the rotation. Students and schools that do not appear in the rotation are not affected and are matched identically at $\mu$ and $\mu^{\hat{\Gamma}}$; therefore Lemma 1 immediately implies the following result:[21]

**Corollary 1.**

$$\mu R \mu^{\hat{\Gamma}} \quad \text{and} \quad \mu^{\hat{\Gamma}} \succeq \mu.$$

Throughout a rotation, students get matched to the school they prefer among those to which they have a claim; therefore any student who appears in $\hat{\Gamma}$ does not have any claim at $\mu^{\hat{\Gamma}}$. As students who do not appear in $\hat{\Gamma}$ are unaffected and $\mu^{\hat{\Gamma}} \succeq \mu$, the rotation does not create any new claim, as we formalize below:

**Lemma 2.** *If student $j$ has a claim to school $d$ at $\mu^{\hat{\Gamma}}$, then she has a claim to $d$ at $\mu$.*

Every claim at $\mu^{\hat{\Gamma}}$ is a claim at $\mu$ and, because $\mu$ is essentially stable, any such claim is vacuous at $\mu$. A natural question at this point is whether such a claim can become non-vacuous as a result of a rotation. As we show in the appendix, this is not the case, which implies that rotations preserve essential stability:

[19] We analogously denote by $\mu^{\Gamma} = \mu^{K+1}$ the matching obtained when the whole reassignment chain $\Gamma$ is carried out and say that $j$ and $d$ appear in $\Gamma$ if there exists $k = 1, \ldots, K$ such that $j = i^k$ and $d = c^k$, respectively.

[20] For any school $d$, sets of students are ranked based on responsiveness (see footnote 7).

[21] We write $\mu R \nu$ when all students weakly prefer their assignment at $\mu$ compared to $\nu$: $\mu(i) R_i \nu(i)$ for all $i$. Analogously, $\mu \succeq \nu$ means that $\mu(c) \succeq_c \nu(c)$ for all $c$.

**Lemma 3.** $\mu^{\hat{\Gamma}}$ *is essentially stable.*

We now use these three lemmas to prove our first main result. Consider any essentially stable matching $\mu$. Either $\mu$ is stable, or there exists a (vacuous) claim $(i, c)$ at $\mu$. In the latter case, it is possible to carry out the rotation $\hat{\Gamma}$ induced by that claim in order to obtain $\mu^{\hat{\Gamma}}$. By Lemmas 1 and 3, $\mu^{\hat{\Gamma}}$ is Pareto dominated by $\mu$, assigns to all schools a weakly higher-priority cohort than $\mu$, and is essentially stable. By Lemma 2, any claim at $\mu^{\hat{\Gamma}}$ is a claim at $\mu$. In addition, $(i, c)$ is a claim at $\mu$ by assumption but, as $c$ rejects $i$ at the end of $\hat{\Gamma}$, it is not a claim at $\mu^{\hat{\Gamma}}$. Combining the last two statements implies that $\mu^{\hat{\Gamma}}$ has strictly less claims than $\mu$. Starting from $\mu$, it is possible to carry out rotations, one at a time, until a stable matching $\nu$ (that is, an essentially stable matching with zero claims) is found. By Lemma 1, $\mu R \nu$. Also, $\nu$ is weakly preferred to the student-pessimal stable matching $\mu^p$ and so $\mu R \nu R \mu^p$. The above argument implies the following result:[22]

**Theorem 1.** $\mu^p$ *is the student-pessimal essentially stable matching.*

Next, we turn our attention to the Rural Hospital Theorem and show that it holds for the set of essentially stable matchings. The proof also makes use of rotations and their properties and follows a similar argument to that of Theorem 1. Consider an essentially stable matching $\mu$. Unless $\mu$ is stable, there exists a (vacuous) claim $(i, c)$ that induces a rotation $\hat{\Gamma}$. Let $d$ be a school that is not filled to capacity at $\mu$: $|\mu(d)| < q_d$. That school does not appear in $\hat{\Gamma}$, otherwise the corresponding reassignment chain $\Gamma$ ends before it returns to $i$, which does not happen since $(i, c)$ is a vacuous claim. Therefore, $\mu(d) = \mu^{\hat{\Gamma}}(d)$. By Lemmas 2 and 3, $\mu^{\hat{\Gamma}}(d)$ is essentially stable and has strictly less claims than $\mu$. Starting from $\mu$, it is possible to carry out rotations, one at a time, until a stable matching $\nu$ is found. As none of these rotations affect school $d$, $\mu(d) = \nu(d)$. As shown by Roth (1986), $\nu(d) = \nu'(d)$ for any two stable matchings $\nu$ and $\nu'$. Combining the last two statements yields our next result.

**Theorem 2.** *(Rural Hospital Theorem) For any two essentially stable matchings $\mu$ and $\mu'$ and any school $d$, $|\mu(d)| < q_d$ implies $\mu(d) = \mu'(d)$.*

Theorems 1 and 2 show that the essentially stable set shares some interesting structural similarities with the stable set. It is clear from Example 1 that an analogous result to Theorem 1 does not hold for the student-optimal stable matching. Indeed, the entire motivation

[22] Recall also that when considering the stable set, $\mu^p$ is not only the student-pessimal stable matching, but is also the school-optimal stable matching (see footnote 18). This result also carries over to the essentially stable set: $\mu^p$ is not only the student-pessimal essentially stable matching, but is also the school-optimal essentially stable matching. The proof of this latter claim is analogous to that of Theorem 1 since schools get higher priority students every time a rotation is carried out.

for essential stability is that it allows making students better off, and so the student-optimal stable matching will generally not be the student-optimal essentially stable matching. The natural question is then whether a student-optimal essentially stable matching exists, i.e., is there a unique essentially stable matching that Pareto dominates all others? We answer in the negative with the following counterexample.

**Example 2.** Let there be 4 students $S = \{i_1, i_2, i_3, i_4\}$ and 4 schools $C = \{A, B, C, D\}$, each with capacity 1. The priorities and preferences are given in the following tables.

| $\succ_A$ | $\succ_B$ | $\succ_C$ | $\succ_D$ |
|---|---|---|---|
| $i_3$ | $i_1$ | $i_2$ | $i_4$ |
| $i_1$ | $i_4$ | $i_3$ | $i_3$ |
| $i_2$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\vdots$ | | | |

| $P_{i_1}$ | $P_{i_2}$ | $P_{i_3}$ | $P_{i_4}$ |
|---|---|---|---|
| $\boxed{A}$ | $\dagger A$ | $\boxed{D}$ | $\boxed{B}$ |
| $\dagger B$ | $\boxed{C}$ | $\dagger C$ | $\dagger D$ |
| $\vdots$ | $\vdots$ | $A$ | $\vdots$ |
| | | $\vdots$ | |

$\mu^{\square}$ and $\mu^{\dagger}$ are two Pareto efficient and essentially stable matchings. $\mu^{\square}$ is the DA matching, which in this example happens to be Pareto efficient. $\mu^{\dagger}$ is not stable as $i_1$ has a claim to $A$. However, it is the only claim and the reassignment chain it initiates is

$$i^1 \to A \to i_2 \to C \to i_3 \to A \to i_1 \to \cdots;$$

therefore the claim is vacuous and $\mu^{\dagger}$ is essentially stable. We conclude that a student-optimal essentially stable matching may not exist. Note that Examples 1 and 2 both contain at least one Pareto efficient and essentially stable matching. This is not a coincidence: while the essentially stable set does not form a lattice, Corollary 2 below shows that there will always be at least one PEES matching.

Our analysis reveals an asymmetric structure since an extreme matching exists on one end but not on the other. Clearly, our negative result implies that the set of essentially stable matchings does not form a full lattice; however, given the existence of a student-pessimal essentially stable matching, it seems natural to think it may form a semilattice (more precisely, a meet-semilattice with respect to the partial ordering $R$). We provide a counterexample in Appendix C to show that, perhaps surprisingly, this is in fact not the case.

# 4 Mechanisms

Essential stability allows relaxing the somewhat stringent requirement that is stability in a meaningful way, thus allowing to improve the fate of students without compromising on fairness. It is also desirable to have well-defined mechanisms to find essentially stable matchings. Because the ultimate goal of relaxing stability is to reach efficiency, we focus on three popular Pareto efficient alternatives: top trading cycles (TTC), a two-stage mechanism that combines deferred acceptance with top trading cycles (DA+TTC), and the efficiency-adjusted deferred acceptance (EADA) mechanism of Kesten (2010). Since it is well-known that Pareto efficiency is in general incompatible with stability, the mechanisms we consider are not stable; our goal is to assess whether they are essentially stable or strongly unstable.

**Top Trading Cycles**

Top trading cycles (TTC) is a classic mechanism that first appears in Shapley and Scarf (1974) and was introduced to school choice problems by Abdulkadiroğlu and Sönmez (2003). The intuition behind TTC is that a "priority" at a school is interpreted as an ownership right to a seat at that school, which can be traded away: if $i$ has high priority at $j$'s first choice, and $j$ has high priority at $i$'s first choice, then TTC allows $i$ and $j$ to trade, even though this may violate the priority of some third student $k$ who is not involved in the trade. By continually making all mutually beneficial trades, we eventually end up at a final matching that is Pareto efficient.[23]

Since the seminal paper of Abdulkadiroğlu and Sönmez (2003), there has been an extremely rich literature studying TTC in a school choice context. It is one of the mechanisms most commonly suggested by economists because it is not only Pareto efficient, but is also strategyproof. Despite these appealing features, its use in practice is very rare. One common explanation for this is that TTC is not stable. Here, we strengthen this by showing that TTC is not essentially stable either.

**Proposition 2.** *The top trading cycles mechanism is strongly unstable.*

*Proof.* The proof is by example. Let there be 5 students $S = \{i_1, i_2, i_3, i_4, i_5\}$ and 5 schools $C = \{A, B, C, D, E\}$ with one seat each. The preferences and priorities are given in the table below.

[23]For a formal definition in the school choice context, see Abdulkadiroğlu and Sönmez (2003).

| $\succ_A$ | $\succ_B$ | $\succ_C$ | $\succ_D$ | $\succ_E$ |
|---|---|---|---|---|
| $i_2$ | $i_3$ | $i_1$ | $i_4$ | $\vdots$ |
| $i_1$ | $i_4$ | $i_2$ | $i_3$ | |
| $i_5$ | $i_1$ | $i_3$ | $\vdots$ | |
| $i_4$ | $\vdots$ | $\vdots$ | | |
| $i_3$ | | | | |

| $P_{i_1}$ | $P_{i_2}$ | $P_{i_3}$ | $P_{i_4}$ | $P_{i_5}$ |
|---|---|---|---|---|
| $B$ | $C$ | $C$ | $A$ | $A$ |
| $A$ | $A$ | $D$ | $B$ | $E$ |
| $\vdots$ | $D$ | $B$ | $D$ | $\vdots$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | |

TTC proceeds with each student pointing to her favorite school and each school to its top-priority student. In the initial round, there is one cycle: $(i_1, B, i_3, C, i_1)$. We implement this trade between $i_1$ and $i_3$ and remove them from the market with their assignments. Then, in the next round, $i_2$ forms a self-cycle with $A$ and is therefore assigned to it. In the final round, $i_4$ and $i_5$ are assigned to $D$ and $E$, respectively, and the final TTC outcome is

$$\mu^{TTC} = \begin{pmatrix} A & B & C & D & E \\ i_2 & i_1 & i_3 & i_4 & i_5 \end{pmatrix}.$$

Now, consider student $i_2$, who claims a seat at school $C$. The reassignment chain initiated by this claim is:

$$i_2 \to C \to i_3 \to B \to i_1 \to A.$$

We see that $i_2$'s claim is not vacuous, and so $\mu^{TTC}$ is strongly unstable. $\square$

### DA + TTC

The reason that TTC may produce strongly unstable matchings is that it completely ignores the priorities of students not involved in a trade. An alternative analyzed by Cantala and Pápai (2014) and Alcalde and Romero-Medina (2015) is to first calculate $\mu^{DA}$, and then allow students to trade by running TTC, using the initial DA assignments as the "endowments" in the TTC mechanism.[24] This mechanism will be Pareto efficient and a Pareto improvement over DA (that is, it guarantees that students always receive an assignment that is no worse than their DA school, which is not guaranteed by using TTC alone). As it takes priority into account more than TTC does, it may intuitively appear as "fairer" than TTC; however, the next result shows that this, too, is a strongly unstable mechanism.[25]

[24]Similar "improvement cycles" mechanisms were first proposed by Erdil and Ergin (2008).

[25]Cantala and Pápai (2014) and Alcalde and Romero-Medina (2015) propose alternative stability notions that are satisfied by DA+TTC. See the introduction for further discussion.

**Proposition 3.** *The DA+TTC mechanism is strongly unstable.*

*Proof.* Consider again Example 1. The DA outcome is

$$\mu^{DA} = \begin{pmatrix} A & B & C & D & E \\ i_1 & i_2 & i_3 & i_4 & i_5 \end{pmatrix}.$$

Now, TTC is applied by first giving each student her DA assignment as her initial endowment, and then each student points to her favorite school, and the school points to the student who is endowed with it. Carrying this out, we see that there is only one cycle, $(i_2, C, i_3, B, i_2)$ and so we implement this trade between $i_2$ and $i_3$ and remove them from the market with their assignments. After this, all remaining cycles are self-cycles, and so the final allocation is

$$\mu^{DA+TTC} = \begin{pmatrix} A & B & C & D & E \\ i_1 & i_3 & i_2 & i_4 & i_5 \end{pmatrix}.$$

This is the matching $\mu^*$ introduced in Example 1 which we found had a non-vacuous claim and so we have that $\mu^{DA+TTC}$ is strongly unstable. □

**Efficiency-Adjusted Deferred Acceptance (EADA)**

Combining DA and TTC is one very natural way to improve on deferred acceptance and reach Pareto efficiency. Kesten (2010) takes a different approach: his EADA mechanism starts with the DA matching, but asks students to "consent" to having their priority violated for a school that they cannot obtain in any stable matching. As claims are given up, school seats become less competitive and the mechanism is able to improve the fate of other students. If all students consent, the EADA mechanism produces a matching that is Pareto efficient and improves upon DA. This method of reaching efficiency differs starkly from the trading cycles approaches, as it carefully keeps track of the student priorities. Of course, EADA will not be stable, but as we now show, it is essentially stable.

In the proof of the result below, it actually turns out to be easier to work with the Simplified Efficiency-Adjusted Deferred Acceptance (SEADA) mechanism. SEADA is a simplification of the original EADA mechanism that was introduced by Tang and Yu (2014), who show that the two mechanisms are outcome-equivalent.[26] Following their terminology, we

[26] Further, the definition given below is the version of SEADA in which all students "consent".

say that a school $c$ is **underdemanded** at matching $\mu$ if $\mu(i) R_i c$ for all $i$, that is all students who are not matched to $c$ prefer their assignment.

**SEADA**

**Round 0** Compute the deferred acceptance outcome $DA(P)$. Identify the schools that are underdemanded, and for each student at these schools, make their assignments permanent. Remove these students and their assigned schools from the market.

**Round $r \geq 1$** Compute the DA outcome on the submarket consisting of those students who still remain at the beginning of round $r$. Identify the schools that are underdemanded, and for each student at these schools, make their assignments permanent. Remove these students and their assigned schools from the market.

Let $\mu^0 = DA(P)$, and, for $r \geq 1$, let $\mu^r$ denote the matching at the end of round $r$, defined as follows: if $i$ was removed from the market prior to the beginning of round $r$, then $\mu^r(i) = \mu^{r-1}(i)$; if $i$ remains in the market at the beginning of round $r$, then $\mu^r(i)$ is the school she is assigned at the end of DA on the round $r$ submarket. The final output of the mechanism is $\mu^{SEADA} = \mu^R$, where round $R$ is the final round of the above mechanism. Tang and Yu (2014) show that $\mu^R$ is equivalent to the matching produced by the EADA mechanism.

**Theorem 3.** *The final matching produced by the SEADA mechanism is essentially stable.*

The proof proceeds by defining two alternative preference profiles, one that gives $i$ her SEADA assignment when DA is run and one for which the rejection chain $i$ starts when DA is run is ultimately identical to the reassignment chain. Using the fact that DA is weakly Maskin monotonic (Kojima and Manea, 2010), we show that both profiles must lead to the same assignment for $i$ and so the reassignment chain must end with $i$ back at her SEADA assignment. The full details can be found in the appendix.

Since SEADA is a Pareto efficient mechanism, we have the following immediate corollary, which shows that essential stability reconciles Pareto efficiency and fairness.

**Corollary 2.** *A Pareto efficient and essentially stable matching always exists.*

Note that while the matching produced by EADA is Pareto efficient and essentially stable, there may exist multiple matchings with these properties, in which case the EADA mechanism only finds one of them. For instance, in Example 2 the EADA matching is $\mu^{\square}$, but an alternative PEES matching is given by $\mu^{\dagger}$. This opens the possibility that there are other simple mechanisms besides EADA that produce PEES matchings. Finding such mechanisms and defining criteria to select the "best" one constitutes an interesting question that is mostly beyond the scope of this paper; however, we close by briefly investigating one such criterion that has received a large amount of attention in the literature: strategyproofness.

A (direct) mechanism is **strategyproof** if, for each student, reporting her true ordinal preferences is a weakly dominant strategy. Strategyproofness is a demanding incentive constraint, and, while DA satisfies it, it is also well-known that improving upon DA with respect to efficiency while preserving strategyproofness is effectively impossible, even without considering fairness objectives. More specifically, Kesten (2010) shows that no Pareto efficient mechanism that Pareto dominates DA is strategyproof, while Abdulkadiroğlu et al. (2009) dispense with the Pareto efficiency requirement and show that no mechanism that Pareto dominates DA is strategyproof. Note, however, that these results consider only improvements over DA; we have shown that there are Pareto efficient and essentially stable mechanisms that are Pareto-incomparable with DA, and so they do not preclude the existence of a strategyproof PEES mechanism. However, more recent work by Alva and Manjunath (2017) significantly strengthens these previous negative results by showing that DA is in fact the only strategyproof and stable-dominating mechanism (where a mechanism is **stable-dominating** if it always weakly Pareto dominates *some* stable matching, which need not be the DA matching). Since Theorem 1 shows that any essentially stable mechanism must Pareto dominate the student-pessimal stable matching $\mu^p$, we conclude the following:

**Theorem 4.** *The only essentially stable and strategyproof mechanism is the deferred acceptance mechanism.*

By Theorem 4, all other strategyproof mechanisms besides DA are strongly unstable. More broadly, our results shed a new light on the trade-off between fairness, efficiency, and strategyproofness. Theorem 4 shows that no mechanism can achieve all three properties. Within the context of strategyproofness, the trade-off between efficiency and fairness is very well-understood and effectively comes down to choosing between DA and TTC. Our definition opens up a third option of combining efficiency and fairness by using a PEES mechanism. This latter possibility has received much less attention, and while the final choice is ultimately a policy question, the fact that strategyproofness has been shown to be neither necessary nor sufficient for successful market design in practice is an argument in favor of considering such an option.[27] At the very least, more theoretical, experimental, and empirical investigations are needed to determine the severity of the incentive problem and how it compares with the welfare gains from PEES mechanisms.

# 5 Conclusion

This paper introduces the concept of essential stability, a weakening of classical stability that allows a matching to have some priority-based claims to seats at schools as long as those claims are vacuous. The motivation for this definition is twofold. First, it is compatible with Pareto efficiency, which can significantly improve the welfare of participants. Second, it still adheres to the principle behind imposing stability as a fairness criterion in the first place: students should not have valid claims. Indeed, essential stability makes evidently clear why there is no reason for a student to claim a seat even if she desires it and has high priority. The definition is simple enough that it can easily be explained to non-experts, which we believe constitutes a key advantage for the purpose of practical implementation.

Our paper opens several avenues for future research. First, the existence of multiple PEES matchings raises the question of whether some can be argued to be more desirable than others. If these matchings could be compared in a meaningful way, it may be possible to improve upon the EADA mechanism by selecting the "best" PEES matching in each market. Second, essential stability could constitute a useful concept beyond the model studied in this paper; it could prove particularly valuable, for example, in settings where a stable matching is not guaranteed to exist, such as one-sided matching markets or matching markets with couples. While the right formal definition will likely depend on the particular setting, we hope that the ideas in this paper provide inspiration for thinking about how to appropriately define a fairness criterion that is not only compatible with efficiency, but is also intuitive and convincing to policymakers and market participants.

[27]For example, in two-sided medical residency markets, DA is not strategyproof (Roth, 1984, 1991), but is widely successful, and in school choice, most school districts that use DA actually implement a modified version of the mechanism that is not strategyproof (see, e.g., Calsamiglia et al. (2010)). In light of this, Kesten's (2010) result that truth telling is an ordinal Bayes-Nash equilibrium of the EADA mechanism under imperfect information is encouraging, though more work on the incentives of PEES mechanisms is needed.

# References


Abdulkadiroğlu, A., P. A. Pathak, and A. E. Roth (2005a): "The New York City high school match," *American Economic Review, Papers and Proceedings*, 95, 364–367.

——— (2009): "Strategy-proofness versus Efficiency in Matching with Indifferences: Redesigning the NYC High School Match," *American Economic Review*, 99, 1954–1978.

Abdulkadiroğlu, A., P. A. Pathak, A. E. Roth, and T. Sönmez (2005b): "The Boston Public School Match," *American Economic Review, Papers and Proceedings*, 95, 364–367.

Abdulkadiroğlu, A. and T. Sönmez (2003): "School Choice: A Mechanism Design Approach," *American Economic Review*, 93, 729–747.

Alcalde, J. and A. Romero-Medina (2015): "Strategy-proof fair school placement," *Available at SSRN 1743082*.

Alva, S. and V. Manjunath (2017): "Strategy-proof Pareto-improvement under voluntary participation," Working paper.

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

Calsamiglia, C., G. Haeringer, and F. Klijn (2010): "Constrained school choice: An experimental study," *The American Economic Review*, 1860–1874.

Cantala, D. and S. Pápai (2014): "Reasonably and Securely Stable Matching," Tech. rep., mimeo.

Chen, Y. and T. Sönmez (2002): "Improving efficiency of on-campus housing: An experimental study," *American Economic Review*, 1669–1686.

——— (2004): "An experimental study of house allocation mechanisms," *Economics Letters*, 83, 137–140.

Dubins, L. E. and D. A. Freedman (1981): "Machiavelli and the Gale-Shapley algorithm," *American Mathematical Monthly*, 88, 485–494.

Dur, U., A. Gitmez, and O. Yilmaz (2015): "School Choice Under Partial Fairness," Tech. rep., Working paper, North Carolina State University, 2015.[19].

Dutta, B. and R. Vohra (2017): "Rational expectations and farsighted stability," *Theoretical Economics*, 12, 1191–1227.

Ehlers, L. (2007): "von Neumann–Morgenstern stable sets in matching problems," *Journal of Economic Theory*, 134, 537–547.

Ehlers, L. and T. Morrill (2017): "(Il)legal Assignments in School Choice," .

Erdil, A. and H. Ergin (2008): "What's the Matter with Tie-Breaking? Improving Efficiency in School Choice," *American Economic Review*, 98, 669–689.

Fragiadakis, D. and P. Troyan (2016): "Designing Mechanisms to Focalize Welfare-Improving Strategies," Working paper, Texas A&M University.

Gale, D. and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69, 9–15.

Harsanyi, J. C. (1974): "An equilibrium-point interpretation of stable sets and a proposed alternative definition," *Management science*, 20, 1472–1495.

Herings, P. J.-J., A. Mauleon, and V. Vannetelbosch (2009): "Farsightedly stable networks," *Games and Economic Behavior*, 67, 526–541.

Kesten, O. (2010): "School Choice with Consent," *Quarterly Journal of Economics*, 125, 1297–1348.

Kojima, F. and M. Manea (2010): "Axioms for Deferred Acceptance," *Econometrica*, 78, 633–653.

McVitie, D. G. and L. B. Wilson (1971): "The stable marriage problem," *Communications of the ACM*, 14, 486–490.

Morrill, T. (2016): "Which School Assignments Are Legal?" Working paper, North Carolina State University.

Page Jr, F. H., M. H. Wooders, and S. Kamat (2005): "Networks and farsighted stability," *Journal of Economic Theory*, 120, 257–269.

Ray, D. and R. Vohra (2015): "The farsighted stable set," *Econometrica*, 83, 977–1011.

Roth, A. E. (1984): "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of Political Economy*, 92, 991–1016.

——— (1985): "The College Admissions Problem is Not Equivalent to the Marriage Problem," *Journal of Economic Theory*, 36, 277–288.

——— (1986): "On the Allocation of Residents to Rural Hospitals: A General Property of Two-Sided Matching Markets," *Econometrica*, 54, 425–427.

——— (1991): "A natural experiment in the organization of entry-level labor markets: regional markets for new physicians and surgeons in the United Kingdom," *American Economic Review*, 81, 415–440.

Shapley, L. S. and H. E. Scarf (1974): "On Cores and Indivisibility," *Journal of Mathematical Economics*, 1, 23–28.

Sönmez, T. (2013): "Bidding for Army Career Specialties: Improving the ROTC Branching Mechanism," *Journal of Political Economy*, 121, 186–219.

Sönmez, T. and T. Switzer (2013): "Matching with (Branch-of-Choice) Contracts at United States Military Academy," *Econometrica*, 81, 451–488.

Sönmez, T. and M. U. Ünver (2010): "House allocation with existing tenants: A characterization," *Games and Economic Behavior*, 69, 425–445.

Sönmez, T. and U. M. Ünver (2005): "House allocation with existing tenants: an equivalence," *Games and Economic Behavior*, 52, 153–185.

Tang, Q. and J. Yu (2014): "A new perspective on Kesten's school choice with consent idea," *Journal of Economic Theory*, 154, 543–561.

Tang, Q. and Y. Zhang (2016): "Weak Stability and Pareto Efficiency in School Choice," Working paper, Shanghai University of Finance and Economics.

Thakral, N. (2015): "Matching with Stochastic Arrival," Working paper, Harvard University.

Von Neumann, J. and O. Morgenstern (1944): *Theory of Games and Economic Behavior*, Princeton University Press.

Wako, J. (2010): "A polynomial-time algorithm to find von Neumann-Morgenstern stable matchings in marriage games," *Algorithmica*, 58, 188–220.

# A Omitted proofs

In this appendix, we provide proofs of all results that were not proved in the main text. We first present the proofs of Proposition 1 and Theorem 3. The proofs of all remaining lemmas (including those from the main text and those introduced in the proof of Theorem 3) can be found in the next subsection.

*Proof of Proposition 1.*

Let $\nu_1, \ldots, \nu_N$ be a sequence of matchings, $d_1, \ldots, d_{N-1}$ be a sequence of schools, and $\hat{\Delta}_1, \ldots, \hat{\Delta}_{N-1}$ be a sequence of rotations such that,

- $\nu_1 = \mu^{\hat{\Gamma}}$ and for each $n = 1, \ldots, N-1$:
- $d_n$ is $i$'s most preferred school to which she has a claim at $\nu_n$,
- $\hat{\Delta}_n$ is the rotation initiated by $(i, d_n)$ at $\nu_n$,
- $\nu_{n+1} = \nu_n^{\hat{\Delta}_n}$,

- and $i$ does not have any claim at $\nu_N$.

By Lemmas 2 and 3 and the fact that the claim that starts a rotation disappears when that rotation is carried out, $N$ is finite and $\nu_1, \dots, \nu_N$ are essentially stable. Additionally, by construction, $\mu(i) = \nu_1(i) = \dots = \nu_N(i)$.

Consider the reassignment chain $\Gamma$ and the rotation $\hat{\Gamma}$ initiated by $(i, c)$ at $\mu$. Up to step $\hat{K}$, the two are identical. The rotation ends at step $\hat{K}$, after which $c$ rejects $i$ and the latter is assigned to $\mu(i)$, while the reassignment chain continues until some step $K$, where a student is matched to a school with a free seat. Then, $i$ is unassigned at $\mu^{\hat{K}+1}$ and, for all $j \neq i$, $\mu^{\hat{K}+1}(j) = \mu^{\hat{\Gamma}}(j) = \nu_1(j)$.

Suppose towards an inductive argument that for some $n = 1, \dots, N-1$, there exists $\hat{L}_n$ such that $i$ is unassigned at $\mu^{\hat{L}_n+1}$ and, for all $j \neq i$, $\mu^{\hat{L}_n+1}(j) = \nu_n(j)$. $d_n$ is $i$'s most preferred school among those to which she has a claim at $\nu_n$, by definition, as well as at $\mu^{\hat{L}_n+1}$, by the induction hypothesis. The next steps of $\Gamma$ are then identical to $\hat{\Delta}_n$, the rotation initiated by $(i, d_n)$ at $\nu_n$, until $d_n$ rejects $i$. Let that step be labeled $\hat{L}_{n+1}$; then $i$ is unassigned at $\mu^{\hat{L}_{n+1}+1}$ and, for all $j \neq i$, $\mu^{\hat{L}_{n+1}+1}(j) = \nu_{n+1}(j)$.

By induction, there exists $\hat{L}_N$ such that $i$ is unassigned at $\mu^{\hat{L}_N+1}$ and, for all $j \neq i$, $\mu^{\hat{L}_N+1}(j) = \nu_N(j)$. Since $i$ does not have any claim at $\nu_n(j)$, her most preferred school among those to which she has a claim at $\mu^{\hat{L}_N+1}$ is $\mu(i)$. Since $\mu(i)$ has a free seat, $i$ is matched to it and $\Gamma$ ends. $\square$

*Proof of Theorem 3.*

Consider some arbitrary claim $(i, c)$, and let $\Gamma$ denote the reassignment chain initiated by this claim.[28] We will show that student $i$ must be rejected from $c$ at some point in $\Gamma$, and hence the claim $(i, c)$ is vacuous, and $\mu^R$ is essentially stable.[29]

We start with the following monotonicity lemma, part (i) of which is due to Kojima and Manea (2010). To state it, say that a preference relation $P_i'$ is a **monotonic transformation** of $P_i$ at $c \in C$ if $bR_i'c \implies bR_ic$. Preference profile $P'$ is a monotonic transformation of $P$ at a matching $\mu$ if $P_i'$ is a monotonic transformation of $P_i$ at $\mu(i)$ for all $i$. In words, $P'$ is a monotonic transformation of another preference profile $P$ at a matching $\mu$ if, for all $i$, the ranking of $\mu(i)$ only increases in moving from $P_i$ to $P_i'$.

[28]If there are no claims, then the matching is classically stable, and so is also essentially stable trivially. Also, $\mu^R$ is nonwasteful, and so any claim $(i, c)$ must be because there exists some $j \in \mu^R(c)$ such that $i \succ_c j$.

[29]In an earlier version of this paper, we also prove that every round $r$ matching $\mu^r$ is essentially stable. For simplicity, we focus on the most important one, $\mu^R$, here.

**Lemma 4.** *(i) If $P'$ is a monotonic transformation of $P$ at $DA(P)$, then $DA_i(P')R'_iDA_i(P)$ for all students $i \in S$.*

*(ii) If $P'$ is a monotonic transformation of $P$ at $DA(P)$, then $DA_i(P')R_iDA_i(P)$ for all students $i \in S$.*

Now, consider again the claim $(i, c)$ at $\mu^R$. Because DA on the round $R$ submarket is stable, only students who were removed in a round strictly earlier than $R$ (the final round) can have a claim. That is, $i$ must have been removed in some round $\hat{r} < R$. Define an alternative preference profile $P^{\hat{r}}$ as follows: for any student $j$ removed before round $\hat{r}$, $P_j^{\hat{r}}$ ranks her assignment $\mu^{\hat{r}}(j)$ first, and the remaining schools in the same order as the true $P_j$; for all $j$ not removed before round $\hat{r}$, $P_j^{\hat{r}} = P_j$. Note that this is a simple way to describe preferences so that $DA(P^{\hat{r}}) = \mu^{\hat{r}}$.[30]

Define a second preference profile $\bar{P}_j$ as follows: for each $j \neq i$, $\bar{P}_j$ ranks $\mu^R(j)$ first, and every other school is listed in the same order as the true $P_j$, while for student $i$, $\bar{P}_i$ ranks $c$ first and the remaining schools in the order of the true $P_i$.

**Lemma 5.** *Student i's DA assignment at the end of round $\hat{r}$ is the same as her DA assignment under $\bar{P}$, which is the same as her assignment at the end of round R: $DA_i(\bar{P}) = DA_i(P^{\hat{r}}) = \mu^R(i)$.*

The lemma is formally proved in the "Proofs of lemmas" subsection that follows the proof of this theorem, but the main step is that $\bar{P}$ is a monotonic transformation of $P^{\hat{r}}$ at $DA(P^{\hat{r}})$. Now, it is well-known that the following is an alternative description of the DA mechanism (McVitie and Wilson, 1971; Dubins and Freedman, 1981):

**DA** At each step $t$, arbitrarily choose one student among those who are currently unmatched, and allow her to apply to her most preferred $a$ school that has not yet rejected him. All schools other than $a$ tentatively hold the same students as the last step. School $a$ holds the highest-priority students up to their capacity among those held from last step combined with the new applicant and reject the (at most one) other.

In this new method, the choice of the applicant at each step is arbitrary, in the sense that the order in which they are chosen does not affect the final outcome. So, for any fixed preference profile, one way to find the DA outcome is to have $i$ be the last student chosen to

[30] Raising school $\mu^{\hat{r}}(j)$ for all $j$ removed prior to round $\hat{r}$ to the top of her preferences is a way to effectively "remove" student $j$ from the market, because no student who has not been removed prior to round $\hat{r}$ will ever apply to such a school because it is underdemanded.

enter the market. That is, as long as there is some other student besides $i$ who is tentatively unmatched, we always choose one of these students to make the next application. Once all of these students have been (tentatively) assigned to a school, we allow $i$ to enter by applying to the first school on her preference list. Student $i$'s application then initiates a rejection chain, where $i$ applies to some school $a$, $a$ rejects its lowest-priority student $i^1$, $i^1$ applies to her most preferred school that has not yet rejected her, and so on, until we reach a school $a^K$ with an empty seat, at which point the rejection chain (and the entire DA mechanism) end, and all tentative matchings are made final.

Consider running DA on the preference profile $\bar{P}$ in this manner where $i$ enters the market last. All students $j$ other than $i$ are tentatively matched to $\mu^R(j)$, and then $i$ starts a rejection chain which, as the next lemma proves, turns out be identical to the reassignment chain $\Gamma$.

**Lemma 6.** *The final matching at the end of* $\Gamma$ *is* $DA(\bar{P})$.

By Lemma 6, the outcome of the reassignment chain $\Gamma$ is the same as the outcome of $DA(\bar{P})$. By Lemma 5, $DA_i(\bar{P}) = \mu^R(i)$, and so $i$'s assignment at the end of $\Gamma$ is also $\mu^R(i)$. The only way this is possible is if $i$ is rejected from $c$ at some point in $\Gamma$; that is, the claim $(i, c)$ is vacuous. $\square$

## Proofs of lemmas

*Proof of Lemma 1.*

As the second part of the statement was proved in the main text, we focus on the first part: $\mu(j) P_j \mu^{\hat{\Gamma}}(j)$ for all $j$ appearing in $\hat{\Gamma}$. For any matching $\tilde{\mu}$ and any school $e \in C$, let $\underline{\tilde{\mu}}(e)$ be the lowest-priority student in $\tilde{\mu}(e)$.

Let $j$ be a student who appears in $\hat{\Gamma}$. That student is rejected by her original school $\mu(j)$ at some point in this rotation; therefore she is not matched to it at the end of the rotation: $\underline{\mu}^{\hat{\Gamma}}(\mu(j)) \succ_{\mu(j)} j$ so $\mu^{\hat{\Gamma}}(j) \neq \mu(j)$. It remains to show that $j$ does not prefer $\mu^{\hat{\Gamma}}(j)$ to $\mu(j)$.

Let $d = \mu^{\hat{\Gamma}}(j)$ and, towards a contradiction, suppose that $d P_j \mu(j)$. As $j \in \mu^{\hat{\Gamma}}(d)$ and $d$ appears in $\hat{\Gamma}$, $j \succeq_d \underline{\mu}^{\hat{\Gamma}}(d) \succ_d \underline{\mu}(d)$; therefore $j$ has a claim to $d$ at $\mu$. That claim is vacuous since $\mu$ is essentially stable. Let

$$j = j^0 \to d = d^0 \to j^1 \to d^1 \to j^2 \to d^2 \to \cdots \to j^{\hat{L}} \to d^{\hat{L}} \to j$$

be the rotation initiated by $(j, d)$ at $\mu$, which we denote by $\hat{\Delta}$. Let $\nu^\ell$ $(\ell = 0, 1, \ldots, \hat{L}, \hat{L}+1)$

be the matching (or pseudo matching for $\ell = 1, \ldots, \hat{L}$ since these leave a student unmatched) obtained at step $\ell$ of $\hat{\Delta}$. Then, $\mu^{\hat{\Delta}}(h) = \mu^{\hat{L}+1}(h)$ for all $h \neq j$ and $\mu^{\hat{\Delta}}(j) = \mu(j)$.

For every school $e \in C$, we define $\phi^{\hat{\Gamma}}(e) = \mu^{\hat{\Gamma}}(e) \cup \{h \in S : \underline{\mu}^{\hat{\Gamma}}(e) \succ_e h\}$. That is, $\phi^{\hat{\Gamma}}(e)$ contains the students matched to $e$ at $\mu^{\hat{\Gamma}}$ and those who have a lower priority than all of these students. Our argument proceeds by induction with the following hypothesis:

For some $\ell = 1, \ldots, \hat{L}$ and for every school $e \in C$, $\nu^\ell(e) \subseteq \phi^{\hat{\Gamma}}(e)$.

We begin by showing that our induction hypothesis is satisfied for $\ell = 1$. Consider any school $e \in C$ and recall that $\nu^0 = \mu$. All the students in $\mu(e) \setminus \mu^{\hat{\Gamma}}(e)$ are rejected by $e$ along $\hat{\Gamma}$, which means that they have a lower priority than all students in $\mu^{\hat{\Gamma}}(e)$. We conclude that $\nu^0(e) = \mu(e) \subseteq \phi^{\hat{\Gamma}}(e)$. For all $e \notin \{d, \mu(j)\}$, $\nu^0(e) = \nu^1(e)$ while $\nu^1(\mu(j)) = \nu^0(\mu(j)) \setminus \{j\} \subseteq \phi^{\Gamma}(e)$; therefore it remains to show that $\nu^1(d) \subseteq \phi^{\hat{\Gamma}}(d)$. By construction, $\nu^1(d) \subset \nu^0(d) \cup \{j\}$. As $j \in \mu^{\hat{\Gamma}}(d) \subseteq \phi^{\hat{\Gamma}}(d)$, $\nu^1(d) \subseteq \phi^{\hat{\Gamma}}(d)$, as required.

We next suppose that our induction hypothesis is satisfied for some $\ell = 1, \ldots, \hat{L}$ and show that it is satisfied for $\ell + 1$. For every $e \neq d^\ell$, $\nu^\ell(e) = \nu^{\ell+1}(e)$; therefore the induction hypothesis directly implies that $\nu^{\ell+1}(e) \subseteq \phi^{\hat{\Gamma}}(e)$. By construction, $\nu^{\ell+1}(d^\ell) \subset \nu^\ell(d^\ell) \cup \{j^\ell\}$; therefore it remains to show that $j \in \phi^{\hat{\Gamma}}(d^\ell)$.

We first show that $j^\ell$ (who is unmatched at $\nu^\ell$) has a claim to $\mu^{\hat{\Gamma}}(j^\ell)$ at $\nu^\ell$. This is trivially the case if $|\nu^\ell(\mu^{\hat{\Gamma}}(j^\ell)| < q_{\mu^{\hat{\Gamma}}(j^\ell)}$. Otherwise, $\underline{\nu}^\ell(\mu^{\hat{\Gamma}}(j^\ell))$ is the $q^{\text{th}}_{\mu^{\hat{\Gamma}}(j^\ell)}$ highest-priority student in $\nu^\ell(\mu^{\hat{\Gamma}}(j^\ell))$. As $\nu^\ell(\mu^{\hat{\Gamma}}(j^\ell)) \subseteq \phi^{\hat{\Gamma}}(\mu^{\hat{\Gamma}}(j^\ell))$ by the induction hypothesis, she has a weakly lower priority than the $q^{\text{th}}_{\mu^{\hat{\Gamma}}(j^\ell)}$ highest-priority student in $\phi^{\hat{\Gamma}}(\mu^{\hat{\Gamma}}(j^\ell))$. By construction, that student has a weakly lower priority than $\underline{\mu}^{\hat{\Gamma}}(\mu^{\hat{\Gamma}}(j^\ell))$; hence $\underline{\mu}^{\hat{\Gamma}}(\mu^{\hat{\Gamma}}(j^\ell)) \succeq_{\mu^{\hat{\Gamma}}(j^\ell)} \underline{\nu}^\ell(\mu^{\hat{\Gamma}}(j^\ell))$. Combining this with the fact that $j^\ell \in \mu^{\hat{\Gamma}}(\mu^{\hat{\Gamma}}(j^\ell)) \setminus \nu^\ell(\mu^{\hat{\Gamma}}(j^\ell))$, we conclude that $j^\ell \succ_{\mu^{\hat{\Gamma}}(j^\ell)} \underline{\nu}^\ell(\mu^{\hat{\Gamma}}(j^\ell))$ so $j^\ell$ has a claim to $\mu^{\hat{\Gamma}}(j^\ell)$ at $\nu^\ell$.

We now conclude our inductive argument by showing that $j \in \phi^{\hat{\Gamma}}(d^\ell)$. By construction, $d^\ell$ is $j^\ell$'s most preferred school among those to which she has a claim at $\nu^\ell$; therefore $d^\ell R_{j^\ell} \mu^{\hat{\Gamma}}(j^\ell)$. If $d^\ell = \mu^{\hat{\Gamma}}(j^\ell)$, then $j \in \mu^{\hat{\Gamma}}(d^\ell) \subseteq \phi^{\hat{\Gamma}}(d^\ell)$; therefore we focus on the case where $d^\ell P_{j^\ell} \mu^{\hat{\Gamma}}(j^\ell)$. As $j^\ell$ is unassigned at $\nu^\ell$, $\mu(j^\ell)$ has rejected her before step $\ell$ of $\hat{\Delta}$; hence $|\nu^\ell(\mu(j^\ell))| = q_{\mu(j^\ell)}$ and $\underline{\nu}^\ell(\mu(j^\ell)) \succ_{\mu(j^\ell)} j^\ell$. By an analogous reasoning to the one above, $|\nu^\ell(\mu(j^\ell))| = q_{\mu(j^\ell)}$ combined with the induction hypothesis implies that $\underline{\mu}^{\hat{\Gamma}}(\mu(j^\ell)) \succeq_{\mu(j^\ell)} \underline{\nu}^\ell(\mu(j^\ell))$; therefore $\underline{\mu}^{\hat{\Gamma}}(\mu(j^\ell)) \succ_{\mu(j^\ell)} j^\ell$. In turn, this implies that $\mu(j^\ell) \neq \mu^{\hat{\Gamma}}(j^\ell)$ so $j^\ell$ is matched to $\mu^{\hat{\Gamma}}(j^\ell)$ somewhere along $\hat{\Gamma}$. At that point, $\mu^{\hat{\Gamma}}(j^\ell)$ is her most preferred school to which she has a claim; therefore $d^\ell$ is matched to $q_{d^\ell}$ students who all have a higher priority than her. As schools get higher-priority students throughout a rotation, it follows that that $\underline{\mu}^{\hat{\Gamma}}(d^\ell) \succ_{d^\ell} j^\ell$. This in turns implies that $j^\ell \in \phi^{\hat{\Gamma}}(d^\ell)$, which concludes our inductive argument.

It follows, by induction, that $\nu^{\hat{L}+1}(d) \subseteq \phi^{\hat{\Gamma}}(d)$. As $|\nu^{\hat{L}+1}(d)| = q_d$ by construction, this implies that $\underline{\mu}^{\hat{\Gamma}}(d) \succeq_d \underline{\nu}^{\hat{L}+1}(d)$, again by a reasoning analogous to the one above. By construction, $d$ rejects $j$ in step $\hat{L}$ of $\hat{\Delta}$ so $\underline{\nu}^{\hat{L}+1}(d) \succ_d j$. It follows that $\underline{\mu}^{\hat{\Gamma}}(d) \succ_d j$, a contradiction. □

*Proof of Lemma 2.*

We first show by contradiction that $j$ does not appear in $\hat{\Gamma}$. If she does, then there exists $k = 1, \ldots, \hat{K}$ such that $j = i^k$ and $\mu^{\hat{\Gamma}}(j) = c^k$. By definition, $\mu^{\hat{\Gamma}}(j)$ is $j$'s most preferred school among those to which she has a claim at $\mu^k$ and, by assumption, $dP_j\mu^{\hat{\Gamma}}(j)$; therefore $\mu^k(d)$ contains $q_d$ students with a higher priority than $j$. As schools receive higher priority student throughout a rotation, so does $\mu^{\hat{\Gamma}}(d)$; consequently, $j$ does not have a claim to $d$ at $\mu^{\hat{\Gamma}}$, a contradiction.

We have established that $j$ does not appear in $\hat{\Gamma}$, which implies that $\mu^{\hat{\Gamma}}(j) = \mu(j)$; hence $dP_j\mu(j)$. By assumption, $\mu^{\hat{\Gamma}}(d)$ contains at most $q_d - 1$ students with a higher priority than $j$. By Corollary 1, $\mu^{\hat{\Gamma}}(d) \succeq_d \mu(d)$; therefore $\mu(d)$ also contains at most $q_d - 1$ students with a higher priority than $j$ so $j$ has a claim to $d$ at $\mu$. □

*Proof of Lemma 3.*

Consider a claim $(j, d)$ at $\mu^{\hat{\Gamma}}$. We need to show that this claim is vacuous. By Lemma 2, $(j, d)$ is a claim at $\mu$ and, as $\mu$ is essentially stable, it is vacuous. The remainder of the proof follows a similar inductive argument to that of Lemma 1. Let

$$j = j^0 \to d = d^0 \to j^1 \to d^1 \to j^2 \to d^2 \to \cdots \to j^{\hat{L}} \to d^{\hat{L}} \to j$$

be the rotation initiated by $(j, d)$ at $\mu$, which we denote by $\hat{\Delta}$. Let $\nu^\ell$ $(\ell = 0, 1, \ldots, \hat{L}, \hat{L}+1)$ be the matching (or pseudo matching for $\ell = 1, \ldots, L$ since these leave a student unmatched) obtained at step $\ell$ of $\hat{\Delta}$. Then, $\mu^{\hat{\Delta}}(h) = \mu^{\hat{L}+1}(h)$ for all $h \neq j$ and $\mu^{\hat{\Delta}}(j) = \mu(j)$.

We denote by $\Delta^*$ the rejection chain initiated by $(j, d)$ at $\mu^{\hat{\Gamma}}$ and by $\mu^* = \mu^{\hat{\Gamma}, \Delta^*}$ the matching obtained after, starting from $\mu$, the rotation $\hat{\Gamma}$ and the reassignment chain $\Delta^*$ are successively carried out. For any matching $\tilde{\mu}$ and and any school $e \in C$, let $\underline{\tilde{\mu}}(e)$ be the lowest-priority student in $\tilde{\mu}(e)$. and, for every school $e \in C$, we define $\phi^*(e) = \mu^*(e) \cup \{h \in S : \underline{\mu}^*(e) \succ_e h\}$. That is, $\phi^*(e)$ contains the students matched to $e$ at $\mu^*$ and those who have a lower priority than all of these students.

Our proof is by induction and uses the following hypothesis:

For some $\ell = 1, \ldots, \hat{L}$ and for every school $e \in C$, $\nu^\ell(e) \subseteq \phi^*(e)$.

We begin by showing that our induction hypothesis is satisfied for $\ell = 1$. Consider any school $e \in C$ and recall that $\nu^0 = \mu$. All the students in $\mu(e) \setminus \mu^*(e)$ are rejected by $e$ along either $\hat{\Gamma}$ or $\Delta^*$, which means that they have a lower priority than the students in $\mu^*(e)$. We conclude that $\nu^0(e) = \mu(e) \subseteq \phi^*(e)$. For all $e \notin \{d, \mu(j)\}$, $\nu^0(e) = \nu^1(e)$ while $\nu^1(\mu(j)) = \nu^0(\mu(j)) \setminus \{j\} \subseteq \phi^\Gamma(e)$; therefore it remains to show that $\nu^1(d) \subseteq \phi^*(d)$. By construction, $\nu^1(d) \subset \nu^0(d) \cup \{j\}$ and $j$ is matched to $d$ at the beginning of $\Delta^*$; therefore either $j \in \mu^*(d)$ or $d$ rejects $j$ somewhere along $\Delta^*$, in which case $\underline{\mu}^*(d) \succ_d j$. We conclude that $j \in \phi^*(d)$, which implies that $\nu^1(d) \subseteq \phi^*(d)$.

We next suppose that our induction hypothesis is satisfied for some $\ell = 1, \ldots, \hat{L}$ and show that it is satisfied for $\ell + 1$. For every $e \neq d^\ell$, $\nu^\ell(e) = \nu^{\ell+1}(e)$; therefore the induction hypothesis directly implies that $\nu^{\ell+1}(e) \subseteq \phi^*(e)$. By construction, $\nu^{\ell+1}(d^t) \subset \nu^\ell(d^\ell) \cup \{j^\ell\}$; therefore it remains to show that $j \in \phi^*(d^\ell)$.

We first show that $j^\ell$ (who is unmatched at $\nu^\ell$) has a claim to $\mu^*(j^\ell)$ at $\nu^\ell$. This is trivially the case if $|\nu^\ell(\mu^*(j^\ell)| < q_{\mu^*(j^\ell)}$. Otherwise, as $\nu^\ell(\mu^*(j^\ell)) \subseteq \phi^*(\mu^*(j^\ell))$ by the induction hypothesis, $\underline{\nu}^\ell(\mu^*(j^\ell))$ has a weakly lower priority than the $q^{\text{th}}_{\mu^{\hat{\Gamma}}(j^\ell)}$ highest-priority student in $\phi^*(\mu^*(j^\ell))$. By construction, that student has a weakly lower priority than $\underline{\mu}^*(\mu^*(j^\ell))$; hence $\underline{\mu}^*(\mu^*(j^\ell)) \succeq_{\mu^*(j^\ell)} \underline{\nu}^\ell(\mu^*(j^\ell))$. Combining this with the fact that $j^\ell \in \mu^*(\mu^*(j^\ell)) \setminus \nu^\ell(\mu^*(j^\ell))$, we conclude that $j^\ell \succ_{\mu^*(j^\ell)} \underline{\nu}^\ell(\mu^*(j^\ell))$ so $j^\ell$ has a claim to $\mu^*(j^\ell)$ at $\nu^\ell$.

We now conclude our inductive argument by showing that $j \in \phi^*(d^\ell)$. By construction, $d^\ell$ is $j^\ell$'s most preferred school among those to which she has a claim at $\nu^\ell$; therefore $d^\ell R_{j^\ell} \mu^*(j^\ell)$. If $d^\ell = \mu^*(j^\ell)$, then $j \in \mu^*(d^\ell) \subseteq \phi^*(d^\ell)$; therefore we focus on the case where $d^\ell P_{j^\ell} \mu^*(j^\ell)$. As $j^\ell$ is unassigned at $\nu^\ell$, $\mu(j^\ell)$ has rejected her before step $\ell$ of $\hat{\Delta}$; hence $|\nu^\ell(\mu(j^\ell))| = q_{\mu(j^\ell)}$ and $\underline{\nu}^\ell(\mu(j^\ell)) \succ_{\mu(j^\ell)} j^\ell$. By an analogous reasoning to the one above, $|\nu^\ell(\mu(j^\ell))| = q_{\mu(j^\ell)}$ combined with the induction hypothesis implies that $\underline{\mu}^*(\mu(j^\ell)) \succeq_{\mu(j^\ell)} \underline{\nu}^\ell(\mu(j^\ell))$; therefore $\underline{\mu}^*(\mu(j^\ell)) \succ_{\mu(j^\ell)} j^\ell$. In turn, this implies that $\mu(j^\ell) \neq \mu^*(j^\ell)$ so $j^\ell$ is matched to $\mu^*(j^\ell)$ somewhere along either $\hat{\Gamma}$ or $\Delta^*$. At that point, $\mu^*(j^\ell)$ is her most preferred school to which she has a claim; therefore $d^\ell$ is matched to $q_{d^\ell}$ students who all have a higher priority than her. As schools get higher-priority students throughout a rotation and a reassignment chain, it follows that that $\underline{\mu}^*(d^\ell) \succ_{d^\ell} j^\ell$. This in turns implies that $j^\ell \in \phi^*(d^\ell)$, which concludes our inductive argument.

By induction, we conclude that $\nu^{\hat{L}+1}(d) \subseteq \phi^*(d)$. As $|\nu^{\hat{L}+1}(d)| = q_d$ by construction, this implies that $\underline{\mu}^*(d) \succeq_d \underline{\nu}^{\hat{L}+1}(d)$, again by a reasoning analogous to the one above. By construction, $d$ rejects $j$ in step $\hat{L}$ of $\hat{\Delta}$ so $\underline{\nu}^{\hat{L}+1}(d) \succ_d j$. It follows that $\underline{\mu}^*(d) \succ_d j$; therefore $d$ rejects $j$ somewhere along $\Delta^*$, which means that this reassignment chain returns to $j$; hence $j$'s claim to $d$ at $\mu^{\hat{\Gamma}}$ is vacuous. □

*Proof of Lemma 4.*

Part (i) is shown in Kojima and Manea (2010), and they refer to this property as *weak Maskin monotonicity.* For part (ii), consider a student $i$, and let $DA_i(P) = a$ and $DA_i(P') = a'$. By part (i), we have $a'R'_ia$. Since $P'_i$ is a monotonic transformation of $P_i$ at $a$, $a'R'_ia$ implies $a'R_ia$. □

*Proof of Lemma 5.*

We start by showing that $\bar{P}$ is a monotonic transformation of $P^{\hat{r}}$ at $DA(P^{\hat{r}})$. For each $j \in S$, let $DA_j(P^{\hat{r}}) = a_j$. For all $j$ removed from the market at some round $r < \hat{r}$, $\mu^r(j) = \mu^{\hat{r}}(j) = a_j$. Thus, both $P^{\hat{r}}_j$ and $\bar{P}_j$ rank school $a_j$ first, and $\bar{P}_j$ is trivially a monotonic transformation of $P^{\hat{r}}_j$ at $a_j$ for these students.

Next, consider the students who are still in the market at the beginning of round $\hat{r}$, and note that for all such students, $P^{\hat{r}}_j = P_j$. Consider some such $j \neq i$. By Lemma 2 of Tang and Yu (2014), $\mu^R(j)R_ja_j$ for all $j$. Since $P^{\hat{r}}_j = P_j$, this further implies that $\mu^R(j)R^{\hat{r}}_ja_j$. Now, consider preference profile $\bar{P}_j$. $\bar{P}_j$ simply raises $\mu^R(j)$ to the top of the ordering, without altering the relative rankings of any other seats (in particular, no schools "jump" over student $j$'s round $\hat{r}$ assignment $a_j$ in the move from $P^{\hat{r}}_j$ to $\bar{P}_j$), and so $\bar{P}_j$ is a monotonic transformation of $P^{\hat{r}}_j$ at $a_j$ for all $j \neq i$.

Last, consider student $i$. She is removed in round $\hat{r}$, and so $cP^{\hat{r}}_ia_i$ (otherwise, student $i$ would not claim a seat at $c$ at $\mu^R$).[31] By similar logic (no school $a'$ "jumps" over $a_i$ in going from $P^{\hat{r}}_i$ to $\bar{P}_i$), $\bar{P}_i$ is a monotonic transformation of $P^{\hat{r}}_i$ at $a_i$. Thus, we have shown that $\bar{P}_j$ is a monotonic transformation of $P^{\hat{r}}_j$ at $DA_j(P^{\hat{r}})$ for all $j \in S$, and so preference profile $\bar{P}$ is a monotonic transformation of preference profile $P^{\hat{r}}$ at $DA(P^{\hat{r}})$.

Finally, given a matching $\mu$, say student $i$ is **not Pareto improvable** if, for every $\nu$ that Pareto dominates $\mu$, $\nu(i) = \mu(i)$. Since $\bar{P}$ is a monotonic transformation of $P^{\hat{r}}$ at $DA(P^{\hat{r}})$,

[31] Because $i$ is removed in round $\hat{r}$, we have $\mu^R(i) = \mu^{\hat{r}}(i) = a_i$; because she claims a seat at $c$ at $\mu^R$, we have $cP_i\mu^R(i)$; again because $i$ is still in the market at round $\hat{r}$, we have $P^{\hat{r}}_i = P_i$. This all implies that $cP^{\hat{r}}_ia_i$.

Lemma 4, part (ii) gives $DA_j(\bar{P})R_j^{\hat{r}}DA_j(P^{\hat{r}})$ for all $j \in S$, i.e., the matching $DA(\bar{P})$ Pareto dominates the matching $DA(P^{\hat{r}})$ with respect to $P^{\hat{r}}$. Since $i$ is removed in round $\hat{r}$, she must be matched with an underdemanded school at $DA(P^{\hat{r}})$ which, by Lemma 1 of Tang and Yu (2014), implies that she is not Pareto improvable relative to $P^{\hat{r}}$. Since $DA(\bar{P})$ Pareto dominates $DA(P^{\hat{r}})$ and $i$ is not Pareto improvable, her matching does not change: $DA_i(\bar{P}) = DA_i(P^{\hat{r}})$. Since $i$ is removed at round $\hat{r}$, her assignment at $R > \hat{r}$ is the same as her assignment at the end of round $\hat{r}$: $\mu^R(i) = DA_i(P^{\hat{r}})$. □

*Proof of Lemma 6.*

Run $DA(\bar{P})$ with the alternative method by letting each student $j \neq i$ make applications in any arbitrary order. By construction of $\bar{P}$, each $j$ applies to $\mu^R(j)$ and is tentatively matched to $\mu^R(j)$. No rejections occur because each $j \neq i$ is assigned to the unique seat to which she is assigned at $\mu^R$. Now, again by construction of $\bar{P}$, when $i$ enters, she begins by applying to $c$. We can index the rest of the steps of DA as a chain of rejections, which we denote $\Xi$, where

$$\text{Step}\,\Xi(k): \quad \text{“student } i^k \text{ applies to school } a^k \text{ which rejects student } i^{k+1}\text{”}.$$

This chain of rejections eventually terminates at some $K$ when a student applies a school with a vacant seat. When a student $i^{k+1}$ is rejected, she goes to the next school on her list and applies. It may be the case that when a student applies to a school, she is rejected immediately, and must continue down her list. Formally, if $i^k \neq i^{k+1}$ we say step $\Xi(k)$ is **effective**. If a step is ineffective ($i^k = i^{k+1}$), then the same student who applied is also the one rejected, and nothing would change if $i^k$ simply skipped her application to $a^k$. Let $\Xi'$ be an alternative rejection chain that deletes all of the ineffective steps of $\Xi$. Deleting ineffective steps has no effect on the final outcome, and so the final matching at the end of $\Xi$ and $\Xi'$ is the same, and by construction, is $DA(\bar{P})$.

The key now is that the steps of $\Xi'$ are the same as the steps of the reassignment chain $\Gamma$. Recall from above that all students $j \neq i$ are tentatively matched to the same school when $i$ enters under $DA(\bar{P})$ as they are matched to when $\Gamma$ begins (namely, school $\mu^R(j)$). Consider step 1. In the former case, a student $j$ is rejected from her initial school $c = \mu^R(j)$.[32] The rest of her preference list $\bar{P}_j$ coincides with her true preferences $P_j$ so she goes down her

[32]Since $i$ is assumed to have a claim at $c$ at $\mu^R$ (and $\mu^R$ is nonwasteful), we have $i \succ_c j$, for some $j \in \mu^R(c)$.

true list $P_j$ until she reaches a school where she has higher priority than some tentatively matched student. This is the same as step 1 of the reassignment chain $\Gamma$. We now have a new tentative matching for DA that is the same as the $k = 1$ matching for $\Gamma$, and the same student $i^1$ who is tentatively unassigned and will make the next application. Using the same argument, the second step of $\Xi'$ leads to the same tentative matching as the $k = 2$ matching under $\Gamma$ and so on for each additional matching, until the same student $i^K$ applies to the first school $c^K$ that has an empty seat, at which point both $\Gamma$ and $\Xi'$ end at the same final matching.[33] □

# B Comparison to other weakenings of stability

In this appendix, first we show formally that our definition of essential stability is distinct from other approaches to weakening stability that have been proposed in the literature by finding matchings that satisfy each of the other definitions but are strongly unstable under our definition. Both Alcalde and Romero-Medina (2015) and Cantala and Pápai (2014) show that the DA+TTC mechanism satisfies their respective definitions of stability, while we showed in Section 4 that DA+TTC is not essentially stable. Therefore, the matching $\mu^*$ from Example 1 is $\tau$-fair, reasonably stable, and securely stable according to their respective definitions, but is strongly unstable according to the definition used in this paper.

The definitions of Morrill (2016), Tang and Zhang (2016), and Ehlers and Morrill (2017) are satisfied by the EADA mechanism and so it is less obvious that they are formally distinct. However, as we show here, they are not equivalent.

We first consider Morrill (2016) (the same argument will apply to Ehlers and Morrill (2017)), who introduces a new formal definition of fairness.[34] In contrast to stability, which is defined on a matching itself, fairness in Morrill (2016) is defined on a set of matchings; i.e.,

[33]Technically, the reassignment chain $\Gamma$ goes back to the top of $P_j$ every time $j$ needs an assignment while the rejection chain goes to the next school in $\bar{P}_j$, but they are equivalent here. This is because, as the reassignment chain progresses, the lowest priority of all the students matched to any school only increases, and so, even though $j$ keeps going back to the top of the list in the reassignment chain, once $j$ has been rejected from a school, she will continue to be rejected, and it is equivalent for her to just start with the next school down the list. Since all schools other than the top school under $\bar{P}_j$ are in the same order as $P_j$, the next (effective) school that $j$ applies to will be equivalent under both scenarios.

[34]Morrill (2016) additionally introduces the notion of a *legal set of assignments* that is defined slightly differently from the fair set. While the result below is stated in terms of fairness, Morrill (2016) shows that the fair set of assignments is equivalent to the legal set of assignments, and so the result applies to both the fair set and the legal set.

an individual matching $\mu$ cannot be deemed "fair" or "unfair" independently, but is only fair in relation to other matchings. More formally, say that a matching $\mu$ **blocks** a matching $\nu$ if there exists some $i$ such that $\mu(i) = aP_i\nu(i)$ and $i \succ_a j$ for some $j \in \nu(a)$. Given a set of matchings $M$, a matching $\mu$ is **possible** for $M$ if $\mu$ is not blocked by any $\nu \in M$. Denote the set of possible matchings for a set $M$ by $\pi(M)$. Then, a set of matchings $F$ is **fair** if

1. For all $\mu \in F$, $\mu$ is not blocked by any $\nu \in F \cup \pi(F)$
2. For all $\mu \notin F$, $\mu$ is blocked by some $\nu \in \pi(F)$.

Example 1 can be used to show that essential stability is different from fairness as defined in Morrill (2016) (and, by extension, from the analogous definitions of Ehlers and Morrill (2017) and Tang and Zhang (2016)). More precisely, we exhibit a matching $\mu$ that must be included in any fair set of matchings $F$, but is not essentially stable. To shorten notation, we refer to a matching by a string of letters representing the school assigned to each student in order of their indices. For example, $\mu = ABCDE$ means that $i_1$ is assigned to $A$, $i_2$ to $B$, $i_3$ to $C$, and so forth.

**Proposition 4.** *Let $F$ be a fair set of matchings, and let $\mu = BACDE$. Then, $\mu \in F$, but $\mu$ is not essentially stable.*

*Proof.* Showing $\mu$ is not essentially stable is simple. Note that $i_3$ claims the seat at school $B$, and the reassignment chain that follows is $(i_3 \to B \to i_1 \to A \to i_2 \to C)$. Since this does not return to $i_3$, the claim $(i_3, B)$ is non-vacuous and so $\mu$ is not essentially stable.

Next, we show that if $F$ is a fair set of matchings, then $\mu \in F$. Let $\pi(F)$ be the set of possible matchings for $F$. First, note that the DA outcome is $\mu^{DA} = ABCDE$, and $\mu^{DA} \in F$ for any $F$ (because it is not blocked by anything). Next, observe that each student $i$ has the highest priority at her DA school. So, $i$ can use the DA matching to block any other matching $\nu$ that gives her a school she disprefers to her DA school. This implies that for all $\nu \in \pi(F)$, $\nu$ Pareto dominates $\mu^{DA}$.[35]

Now, assume that $\mu = BACDE \notin F$. By part (2) of the definition of fairness, there exists some $\nu \in \pi(F)$ that blocks it. The only potential student who can block $\mu$ is $i_3$, who can block with $B$. Let $\nu$ be some $\nu \in \pi(F)$ at which $\nu(i_3) = B$. Since $\nu$ must Pareto dominate $\mu^{DA}$, there is only one possibility: $\nu = ACBDE$.[36] Thus, $\nu = ACBDE \in \pi(F)$.

[35] If $\nu$ does not Pareto dominate $\mu^{DA}$, then there is some $i$ such that $\mu^{DA}(i)P_i\nu(i)$. Then, $\nu$ is not possible for $F$, because $i$ would block $\nu$ using $\mu^{DA}$, which is always included in any $F$.

[36] Since $\nu$ must Pareto dominate $\mu^{DA}$, $i_1$ must get $A$ (because $i_3$ is assigned $B$). Then, since $A$ and $B$ are taken, $\nu(i_2) = C$, which further implies that $\nu(i_4) = D$. The only school left is $E$, and so $\nu(i_5) = E$.

Since $\nu \in \pi(F)$, there is no $\rho \in F$ that blocks it. Since $\nu$ can be blocked by any matching $\rho$ such that $\rho(i_4) = C$, we have $\rho(i_4) = C$ implies that $\rho \notin F$; in particular, $\rho = ABDCE \notin F$.

Since $\rho \notin F$, there must be some $\sigma \in \pi(F)$ that blocks $\rho$. The only student who can block $\rho$ is $i_5$, who can block with any $\sigma$ such that $\sigma(i_5) = D$. However, any such $\sigma$ has some student who is assigned to a school worse than her DA assignment,[37] which contradicts that every $\sigma \in \pi(F)$ Pareto dominates $\mu^{DA}$. $\square$

The above proposition shows that our definition is not equivalent to that of Morrill (2016), and, by extension, Ehlers and Morrill (2017). As far as the definition of weak stability from Tang and Zhang (2016), a result of Morrill (2016) shows that the set of fair matchings $F$ is equivalent to the vNM stable set. Tang and Zhang (2016) show that every matching that is in the vNM stable set is weakly stable in their sense. Thus, the same matching $\mu$ from the above proposition is weakly stable in the sense of Tang and Zhang (2016), but is not essentially stable.

# C Semilattice

As mentioned in the main text, the existence of a minimal element may intuitively suggest that the set of essentially stable matchings forms a semilattice. We show in this appendix that this is in fact not the case. The set of essentially stable matchings forms a **meet-semilattice** with respect to the partial ordering $R$ if for any two essentially stable matchings $\mu_1$ and $\mu_2$, there exists a *greatest lower bound* (also called *infimum* or *meet*) $\overline{\mu}$ such that (i) $\overline{\mu}$ is an essentially stable matching, (ii) $\mu_1 R \overline{\mu}$, $\mu_2 R \overline{\mu}$, and (iii) for any essentially stable matching $\mu$: $\mu_1 R \mu$ and $\mu_2 R \mu$ imply $\overline{\mu} R \mu$.

**Proposition 5.** *The set of essentially stable matchings may not form a meet-semilattice with respect to the partial ordering $R$.*

*Proof.* The proof is by counterexample, which we present below.

**Example 3.** Let there be 7 students $S = \{i_1, i_2, i_3, i_4, i_5, i_6, i_7\}$ and 7 schools $C = \{A, B, C, D, E, F, G\}$, each with capacity 1. The priorities and preferences are given in the following tables.

[37] For each student $i_1, i_2, i_3$, and $i_4$, the schools weakly preferred to her DA assignment are some subset of $\{A, B, C\}$. Since there are only 3 seats at these schools and 4 students, some student must be assigned to a school worse than her DA assignment.

| $\succ_A$ | $\succ_B$ | $\succ_C$ | $\succ_D$ | $\succ_E$ | $\succ_F$ | $\succ_G$ |
|---|---|---|---|---|---|---|
| $i_7$ | $i_1$ | $i_2$ | $i_6$ | $i_4$ | $i_5$ | $i_3$ |
| $i_5$ | $i_2$ | $i_3$ | $i_3$ | $i_5$ | $i_6$ | $i_7$ |
| $i_3$ | $\vdots$ | $\vdots$ | $i_4$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $i_1$ | | | $i_1$ | | | |
| $\vdots$ | | | $\vdots$ | | | |

| $P_{i_1}$ | $P_{i_2}$ | $P_{i_3}$ | $P_{i_4}$ | $P_{i_5}$ | $P_{i_6}$ | $P_{i_7}$ |
|---|---|---|---|---|---|---|
| $A$ | $B$ | $C$ | $D$ | $E$ | $F$ | $G$ |
| $D$ | $C$ | $A$ | $E$ | $A$ | $D$ | $A$ |
| $B$ | $\vdots$ | $D$ | $\vdots$ | F | $\vdots$ | $\vdots$ |
| $\vdots$ | | G | | $\vdots$ | | |
| | | $\vdots$ | | | | |

The following matchings are essentially stable:[38]

$$\mu_1 = \begin{pmatrix} A & B & C & D & E & F & G \\ i_3 & i_1 & i_2 & i_4 & i_5 & i_6 & i_7 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} A & B & C & D & E & F & G \\ i_5 & i_2 & i_3 & i_1 & i_4 & i_6 & i_7 \end{pmatrix}$$

$$\mu_3 = \begin{pmatrix} A & B & C & D & E & F & G \\ i_5 & i_1 & i_2 & i_3 & i_4 & i_6 & i_7 \end{pmatrix} \qquad \mu_4 = \begin{pmatrix} A & B & C & D & E & F & G \\ i_3 & i_1 & i_2 & i_6 & i_4 & i_5 & i_7 \end{pmatrix}$$

It is easy to verify that $\mu_1$ and $\mu_3$ are stable. At $\mu_2$, the only claim is $i_4$'s claim to $D$. The reassignment chain initiated by that claim is

$$i_4 \to D \to i_1 \to B \to i_2 \to C \to i_3 \to D \to i_4 \to \cdots;$$

therefore the claim is vacuous and $\mu_2$ is essentially stable. At $\mu_4$, $i_5$'s claim to $A$ is the only one. The reassignment chain it initiates is

$$i_5 \to A \to i_3 \to G \to i_7 \to A \to i_5 \to \cdots;$$

therefore the claim is vacuous and $\mu_4$ is essentially stable.

It is easy to verify that neither one of $\mu_1$ and $\mu_2$ Pareto dominates the other and that the same holds for $\mu_3$ and $\mu_4$; however, $\mu_1$ and $\mu_2$ both Pareto dominate $\mu_3$ as well as $\mu_4$. To conclude the proof, suppose towards a contradiction that the set of essentially stable matchings forms a meet-semilattice with respect to the partial ordering $R$. Then, $\mu_1$ and $\mu_2$ have a greatest lower bound $\overline{\mu}$. By definition, $\mu_1 R \overline{\mu}$, $\mu_2 R \overline{\mu}$, $\overline{\mu} R \mu_3$, and $\overline{\mu} R \mu_4$; therefore

$$A = \mu_1(i_3) R_{i_3} \overline{\mu}(i_3) R_{i_3} \mu_4(i_3) = A \quad \text{and} \quad A = \mu_2(i_5) R_{i_5} \overline{\mu}(i_5) R_{i_5} \mu_3(i_5) = A.$$

[38] There are also a student-optimal (essentially) stable matching that assigns each student to her favorite school and a student-pessimal (essentially) stable matching that assigns to each school the student with the top priority.

It follows that $\overline{\mu}(i_3) = \overline{\mu}(i_5) = A$, a contradiction since each school has capacity 1. □

# D Robust Essential Stability

A possible concern is that vacuous claims are implemented within a reassignment chain. If we argue that vacuous claims are not as serious as non-vacuous one, then perhaps they should be discarded when constructing reassignment chains. In this appendix, we show that this does not affect our results.

**Definition 3.** The **robust reassignment chain initiated by claim** $(i, c)$ at matching $\mu$ is the list

$$i^0 \to c^0 \to i^1 \to c^1 \to \cdots \to i^K \to c^K$$

where,

- $i^0 = i$, $\mu^0 = \mu$, $c^0 = c$ and for each $k \geq 1$:
- $i^k$ is the lowest-priority student in $\mu^{k-1}(c^{k-1})$,
- $\mu^k$ is defined as: $\mu^k(j) = \mu^{k-1}(j)$, for all $j \neq i^{k-1}, i^k$, $\mu^k(i^{k-1}) = c^{k-1}$, and student $i^k$ is unassigned,
- $c^k$ is student $i^k$'s most preferred school for which she has a **non-vacuous** claim at $\mu^k$,
- and terminates at the first $K$ such that $|\mu^K(c^K)| < q_{c^K}$.

A robust reassignment chain is identical to a reassignment chain except for one difference: at any step $k$, student $i^k$ is matched to the school she prefers among those to which she has a *non-vacuous* claim. We say that $i$'s claim to $c$ is **robustly vacuous** if the robust reassignment chain it initiates returns to $i$ and that $\mu$ is **robustly essentially stable** if all claims at $\mu$ are robustly vacuous. To show that the possible presence of vacuous counter-claims does not affect any of our results, we prove the following equivalence:

**Proposition 6.** *A matching is robustly essentially stable if and only it is essentially stable.*

*Proof.* Let

$$i^0 = i \to c^0 = c \to i^1 \to c^1 \to \cdots \to i^K \to c^K$$

be the *robust* reassignment chain initiated by $(i, c)$ at $\mu$ and for every $k = 0, 1, \ldots, K, K+1$ and let $\mu^k$ be the matching (or pseudo matching for $k = 1, \ldots, K$ since these leave a student unmatched) considered at step $k$ of that chain. We denote that chain by $\overline{\Gamma}$ and the matching obtained once that chain is carried out by $\mu^{\overline{\Gamma}} = \mu^{K+1}$. We denote the reassignment chain initiated by $(i, c)$ at $\mu$ by $\Gamma$ and the matching obtained when this chain is carried out by $\mu^{\Gamma}$. For any matching $\tilde{\mu}$ and any school $e \in C$, let $\underline{\tilde{\mu}}(e)$ be the lowest-priority student in $\tilde{\mu}(e)$. Finally, we define, for every school $e \in C$, $\phi^{\Gamma}(e) = \mu^{\Gamma}(e) \cup \{h \in S : \underline{\mu}^{\Gamma}(e) \succ_e h\}$ to be the set of students who either are matched to $e$ at $\mu^{\Gamma}$ or have a lower priority than all students matched to $e$ at $\mu^{\Gamma}$. We make use of the following result, which we prove below. (The proof follows an inductive argument similar to those of Lemmas 1 and 3.)

**Lemma 7.** *For all $e \in C$, $\mu^{\overline{\Gamma}}(e) \subseteq \phi^{\Gamma}(e)$ and, for all $h \in S$, $\mu^{\overline{\Gamma}}(h) R_h \mu^{\Gamma}(h)$.*

*(RES $\Rightarrow$ ES)* Suppose towards a contradiction that $(i, c)$ is robustly vacuous but not vacuous at $\mu$. On the one hand, $(i, c)$ is robustly vacuous; therefore $c$ rejects $i$ somewhere along $\overline{\Gamma}$ and $\mu^{\overline{\Gamma}}(c)$ contains $q_c$ students with a higher priority than $i$. By Lemma 7, $\mu^{\overline{\Gamma}}(c) \subseteq \phi^{\Gamma}(c)$ so $\phi^{\Gamma}(c)$ contains at least $q_c$ students with a higher priority than $i$. On the other hand, $(i, c)$ is not vacuous so $i \in \mu^{\Gamma}(c)$, which means that $\phi^{\Gamma}(c)$ contains at most $q_c - 1$ students with a higher priority than $i$, a contradiction. We conclude that every robustly vacuous claim is vacuous, which directly implies that every robustly essentially stable matching is essentially stable.

*(ES $\Rightarrow$ RES)* Suppose that $\mu$ is essentially stable. Then, $\Gamma$ ends with $i$ matched to $\mu(i)$ by Proposition 1. Consider $i^K$ and $c^K$, the last student and school to appear in $\overline{\Gamma}$. If $i^K = i$, she appears in $\Gamma$. Otherwise, by construction, $\mu(i)$ rejects $i^K$ somewhere along $\overline{\Gamma}$. This means that $\mu(i^K)$ is matched at $\mu^{\overline{\Gamma}}$ to $q_{\mu(i^K)}$ students with a higher priority than $i^K$. By Lemma 7, $\phi^{\Gamma}(\mu(i^K))$ contains $q_{\mu(i^K)}$ students with a higher priority than $i^K$. In turn, this implies that $\underline{\mu}^{\Gamma}(\mu(i^K)) \succ_{\mu(i^K)} i^K$ so $\mu(i^K)$ rejects $i^K$ somewhere along $\Gamma$, which means that $i^K$ appears in $\Gamma$. By assumption, $c^K$ has a free seat at $\mu^K$ so either $c^K = \mu(i)$ or $|\mu(c^K)| < q_{c^K}$. In either case, $c^K$ has a free seat the last time $i^K$ appears in $\Gamma$; therefore $\mu^{\Gamma}(i^K) R_{i^K} c^K$. By Lemma 7, $c^K = \mu^{\overline{\Gamma}}(i^K) R_{i^K} \mu^{\Gamma}(i^K)$; therefore $\mu^{\Gamma}(i^K) = c^K$. Since $c^K$ has a free seat, $\Gamma$ ends when $i^K$ is matched to $c^K$; therefore $i^K = i$ and $c^K = c$. This means that $\overline{\Gamma}$ returns to $i$ and $(i, c)$ is robustly vacuous at $\mu$. The same reasoning is valid for all claims at $\mu$; therefore that matching is robustly essentially stable. $\square$

*Proof of Lemma 7.*

We first show that $\mu^1(e) \subseteq \phi^{\Gamma}(e)$ for all $e \in C$. For every $e$, $\mu(e) \subseteq \phi^{\Gamma}(e)$ since schools only receive higher priority students throughout a reassignment chain. For all $e \notin \{c, \mu(i)\}$,

$\mu^1(e) = \mu(e) \subseteq \phi^\Gamma(e)$ while $\mu^1(\mu(i)) = \mu(\mu(i)) \setminus \{i\} \subseteq \phi^\Gamma(e)$. Finally, $\mu^1(c) \subseteq \mu(c) \cup \{i\}$ and, therefore, it remains to show that $i \in \phi^\Gamma(c)$. If $(i, c)$ is not vacuous at $\mu$, $i \in \mu^\Gamma(c)$ so $i \in \phi^\Gamma(c)$. If $(i, c)$ is vacuous, $c$ rejects $i$ somewhere along $\Gamma$ so all students in $\mu^\Gamma(c)$ have a higher priority than $i$ and $i \in \phi^\Gamma(c)$.

We next proceed with the following induction hypothesis:

For some $k = 1, \ldots, K$ and for every school $e \in C$, $\mu^k(e) \subseteq \phi^\Gamma(e)$.

We show that $\mu^{k+1}(e) \subseteq \phi^\Gamma(e)$. This holds trivially for all $e \neq c^k$ since $\mu^{k+1}(e) = \mu^k(e)$. In addition, $\mu^{k+1}(c^k) \subseteq \mu^k(c^k) \cup \{i^k\}$; therefore, it remains to show that $i^k \in \phi^\Gamma(c^k)$. The following result is proved separately (again, the proof makes use of an inductive argument).

**Lemma 8.** *$i^k$ has a non-vacuous claim to $\mu^\Gamma(i^k)$ at $\mu^k$.*

As $c^k$ is by definition $i^k$'s most preferred school to which she has a non-vacuous claim at $\mu^k$, Lemma 8 implies $c^k R_{i^k} \tilde{c} = \mu^\Gamma(i^k)$. If $c^k = \tilde{c}$, then $i^k \in \mu^\Gamma(c^k) \subseteq \phi^\Gamma(c^k)$. If $c^k P_{i^k} \tilde{c}$, then $i^k$ does not have a claim to $c^k$ the last time she appears in $\Gamma$. Since schools only get higher–priority students throughout a reassignment chain, this means that $\mu^\Gamma(c^k)$ contains $q_{c^k}$ students who all have a higher priority than $i^k$. Therefore, $\underline{\mu}^\Gamma(c^k) \succ_{c^k} i^k$ so $i^k \in \phi^\Gamma(c^k)$, which concludes the proof of the first part of the statement.

We next turn to the second part of the statement, considering any student $h \in S$. By Lemma 1, $\mu(h) R_h \mu^\Gamma(h)$. Each time she appears in $\overline{\Gamma}$ (if any), by Lemma 8, $h$ has a non-vacuous claim to $\mu^\Gamma(h)$. Therefore, she remains matched to a school she weakly prefer to $\mu^\Gamma(h)$ throughout $\overline{\Gamma}$, which means that $\mu^{\overline{\Gamma}}(h) R_h \mu^\Gamma(h)$. □

*Proof of Lemma 8.*

For ease of notation, let $\tilde{c} = \mu^\Gamma(i^k)$. We first show that $i^k$ has a claim to $\tilde{c}$ at $\mu^k$. This is trivially the case if $|\mu^k(\tilde{c})| < q_{\tilde{c}}$. Otherwise, the facts that $i^k \in \mu^\Gamma(\tilde{c}) \setminus \mu^k(\tilde{c})$ and $\mu^k(\tilde{c}) \subseteq \phi^\Gamma(\tilde{c})$ imply that $i^k \succ_{\tilde{c}} \underline{\mu}^k(\tilde{c})$.

We next show that $i^k$'s claim to $\tilde{c}$ at $\mu^k$ is not vacuous. Let

$$j^0 = i^k \to d^0 = \tilde{c} \to j^1 \to d^1 \to \cdots \to j^L \to d^L$$

be the reassignment chain initiated by $(i^k, c^k)$ at $\mu^k$, which we denote by $\Delta$. Let $\mu^\Delta$ be the matching obtained after, starting from $\mu^k$, this reassignment chain is carried out and, for all $\ell = 0, 1, \ldots, L, L+1$, let $\nu^\ell$ denote the matching (or pseudo matching for $\ell = 1, \ldots, L$ since these leave a student unmatched) obtained at step $\ell$.

By assumption, $\nu^0(e) = \mu^k(e) \subseteq \phi^\Gamma(e)$ for all $e \in C$. We next proceed with the following induction hypothesis:

For some $\ell = 0, 1, \ldots, L$ and for every school $e \in C$, $\nu^\ell(e) \subseteq \phi^\Gamma(e)$.

As $\nu^\ell(e) = \nu^{\ell+1}(e)$ for all $e \neq d^\ell$ and $\nu^{\ell+1}(d^\ell) \subseteq \nu^\ell(d^\ell) \cup \{j^\ell\}$, it is sufficient to show that $j^\ell \in \phi^\Gamma(d^\ell)$. By a reasoning analogous to the one above, $j^\ell$ has a claim to $d^\ell$ at $\nu^\ell$; therefore $d^\ell R_{j^\ell} \mu^\Gamma(j^\ell)$. If $d^\ell = \mu^\Gamma(j^\ell)$, then $j^\ell \in \mu^\Gamma(d^\ell) \subseteq \phi^\Gamma(d^\ell)$. If $d^\ell P_{j^\ell} \mu^\Gamma(j^\ell)$, then $j^\ell$ would not have been matched to $\mu^\Gamma(j^\ell)$ the last time she appeared in $\Gamma$ unless $\underline{\mu}^\Gamma(d^\ell) \succ_{d^\ell} j^\ell$. Therefore, $j^\ell \in \phi^\Gamma(d^\ell)$. We conclude that $\mu^\Delta(e) \subseteq \phi^\Gamma(e)$ for all $e \in C$.

Towards a contradiction, suppose that $i^k$'s claim to $\tilde{c}$ at $\mu^k$ is vacuous. Then, $\tilde{c}$ rejects $i^k$ somewhere along $\Delta$ so $\mu^\Delta(\tilde{c})$ contains $q_{\tilde{c}}$ students with a higher priority than $\tilde{c}$. Since $i^k \in \mu^\Gamma(\tilde{c})$, this contradicts the fact that $\mu^\Delta(\tilde{c}) \subseteq \phi^\Gamma(\tilde{c})$. $\square$