# Responsive Affirmative Action in School Choice[*]

Battal Doğan[†]

January 24, 2015

**Abstract**

School choice programs aim to give students the option to choose their school. At the same time, underrepresented minority students should be favored to close the opportunity gap. A common way to achieve this is to have a majority quota at each school, and to require that no school be assigned more majority students than its majority quota. An alternative way is to reserve some seats at each school for the minority students, and to require that a reserve seat at a school be assigned to a majority student only if no minority student prefers that school to her assignment. However, fair rules based on either type of affirmative action suffer from a common problem: a stronger affirmative action may not benefit any minority student and hurt some minority students. First, we show that this problem is pervasive: the problem disappears only if the minority students "mostly" have priority over the majority students. Then, we uncover the root of this problem: for some minority students, treating them as minority students does not benefit them, but possibly hurts other minority students. We propose a new assignment rule (Modified deferred acceptance with minority reserves), which treats such minority students as majority students, achieves affirmative action, and never hurts a minority student without benefiting another minority student.

*JEL* Classification Numbers: C78, D47, D71, D78

Keywords: School choice, affirmative action, minimal responsiveness

## 1 Introduction

Due to historical discrimination based on criteria such as color and ethnicity, some groups are not represented in education as much as they are in the general population. The term "affirmative action" refers to regulations that aim to remedy the situation by favoring the underrepresented minority groups. Currently in many school districts, students are placed in public schools through

[*] I am grateful to William Thomson for his guidance and many helpful comments. I thank Paulo Barelli, Lars Ehlers, Srihari Govindan, Isa Hafalir, Onur Kesten, Bettina Klaus, Asen Kochov, Romans Pancs, Christoph Schlegel, Bumin Yenmez, Kemal Yildiz, and seminar participants at Carnegie Mellon University, University of Lausanne, University of Rochester, the Midwest Economic Theory Conference in 2013, and the 8th Matching in Practice Workshop for helpful comments.

[†] battaldogan@gmail.com; Faculty of Business and Economics, University of Lausanne, 1015 Lausanne, Switzerland.

public school choice programs that allow students to choose their schools instead of being assigned to a school solely based on where they live, and many school districts have affirmative action policies to favor minority students and help them attend more preferred schools.[1]

Since school capacities are limited and some schools may be over-demanded, schools are given priority orderings over students.[2] Given the preferences of the students and the priority orderings, how to assign students to schools in a good way calls for the design of assignment rules. Assignment rules without affirmative action policies have been extensively studied since the pioneering work of Abdulkadiroğlu and Sönmez (2003).[3] A central objective has been fairness: there should be no student-school pair such that the student prefers the school to her assigned school, and the school has a vacant seat or it is assigned another student who has lower priority at that school. That is, no students' priority at any school should be violated. Having no priority violation is desirable also from the stability perspective. An unfair assignment is unstable in the sense that there is a pair of agents who have the incentive to circumvent the assignment. There is extensive empirical evidence that agents do indeed circumvent an unstable assignment and the clearinghouses that use stable assignment rules have succeeded while the ones using unstable assignment rules tend to fail (Kagel and Roth (2000); Roth (2002); Ünver (2000)).

For school districts without an affirmative action policy, the design program for fair assignment rules has been successful (Abdulkadiroğlu (2013); Roth (2008)). However, the case with affirmative action policies is controversial. To highlight a controversy, consider for instance the most common affirmative action policy, the quota-type of affirmative action: at each school, there is a quota for the majority students, and no school can be assigned more majority students than its quota. This policy has two implications. First, it is a constraint and the fairness requirement should be modified to be compatible with this constraint (in particular, some priority violations will now be allowed since otherwise the quota requirement will be violated). Second, the purpose of this policy is to favor the minority students and the assignment rule should be responsive to that. A minimal requirement for an assignment rule to be responsive is the following: consider any school choice problem with majority quotas; consider the assignment recommended by the rule; consider another problem obtained by lowering the quotas at some schools; consider the new assignment recommended by the rule; if a minority student is worse off, then at least one other minority student should be better off. If a rule satisfies this property, then we say that it is minimally responsive to affirmative action, or simply *minimally responsive*.[4] Otherwise, we

[1] Some school districts in the US that have affirmative action policies are Boston (MA), Jefferson County (KY), Kansas City (MO), Louisville (KY), Minneapolis (MN), and St. Louis (MO).

[2] In most of the school districts, priorities are determined by criteria such as whether the student lives in the attendance area of that school and whether the student has a sibling attending that school.

[3] For a survey on school choice, see Abdulkadiroğlu (2013).

[4] *Minimal responsiveness* is introduced by Kojima (2012). Kojima (2012) calls it "respect for the spirit of affirmative action".

say that it is perversely responsive.

For any affirmative action policy, as long as a stronger affirmative action is well-defined, minimal responsiveness is also well-defined and requires that a stronger affirmative action should not hurt some minority students without benefiting another minority student. In fact, one may argue that a rule should respond to a stronger affirmative action by favoring all the minority students or at least a significant proportion of them. Yet, all that minimal responsiveness requires is that a stronger affirmative action does not hurt all the minority students, hence the name minimal responsiveness.

Kojima (2012) shows that under the quota-type affirmative action policy, there is no fair assignment rule that is *minimally responsive*. The first contribution of our paper is to better understand this impossibility, in particular whether it is due to some exceptional school choice environments. We give a complete answer to this question by offering a characterization of the priority structures for which the incompatibility disappears (Theorem 1).[5] It turns out that such priority structures are very restricted. Specifically, suppose that the number of minority students is greater than the capacity of each school, which is very likely to be met in practice. Then, there is a fair and *minimally responsive* rule that achieves the quota-type affirmative action if and only if each majority student has lower priority than each minority student at each school (Corollary 1).

Recently, Hafalir et al. (2013) proposed another type of affirmative action, in which each school reserves a certain number of seats for the minority students, although majority students can also be assigned to those seats provided that no minority student prefers that school to her assigned school. We call this the reserve-type affirmative action. Hafalir et al. (2013) also propose a fair assignment rule, "deferred acceptance with minority reserves" ($DA^m$), that achieves the reserve-type affirmative action and brings considerable efficiency gains over fair rules that achieve the quota-type affirmative action. Yet, $DA^m$ is not *minimally responsive*. Our second contribution is twofold: First, we show that no fair and *minimally responsive* assignment rule achieves the reserve-type affirmative action; second, we show the scope of the impossibility (Theorem 2). Interestingly, it is almost the same as the scope of the impossibility for the quota-type affirmative action. Specifically, suppose that the number of minority students is greater than the total capacity of any two schools, which is also very likely to be met in practice. Then, there is a fair and *minimally responsive* rule that achieves the reserve-type affirmative action if and only if each majority student has lower priority than each minority student (except for at most one minority student) at each school (except for at most one school) (Proposition 2).

Hafalir et al. (2013), after noting that $DA^m$ is not *minimally responsive*, propose the fol-

[5] The priority structures for which the incompatibility of fairness and *Pareto efficiency* disappears have been characterized before (Ergin (2002) and Ehlers and Erdil (2010)).

lowing solution. They show that if the minority reserves are chosen "carefully", which they refer to as smart reserves, then $DA^m$ is *minimally responsive*. However, to design smart reserves, one needs to know the number of minority students that would be assigned to each school at a fair assignment when there is no affirmative action, which obviously requires preference information (Hafalir et al. (2013)). They propose to use the past data to figure out what the assignment would be without affirmative action, which does not necessarily guarantee *minimal responsiveness*.

Instead of checking what happens to the minority students when we move from a weaker affirmative action to a stronger affirmative action, we could have focused on what happens when we move from no affirmative action to affirmative action. This would give us a requirement weaker then *minimal responsiveness*. In all the proofs of the results relating to the impossibilities with the current policies, we start with a no affirmative action problem and show that when we move to (a stronger) affirmative action, *minimal responsiveness* has to be violated (This is the case also for the proofs in Kojima (2012)). Thus, the rules based on the current affirmative action policies do not even satisfy this weaker requirement.

All these disappointing news about the two types of affirmative action expose the need for an affirmative action rule that is *minimally responsive*. Our third contribution is to propose a fair and *minimally responsive* rule that achieves a certain type of affirmative action. Our proposal is a modification of $DA^m$. We call it "modified deferred acceptance with minority reserves" and denote it by $MDA^m$. Our modification is based on the following observation about why $DA^m$ is *perversely responsive*: in the $DA^m$ algorithm, a minority student who has a lower priority than a majority student at a school, is temporarily accepted by that school while the majority student is rejected; however, the majority student being rejected initiates a sequence of rejections that may end with the minority student being rejected by the same school; in a sense, the minority student "interferes" with the admission process of the school without getting anything out of it. The key idea behind $MDA^m$ is to first detect these interferers and treat them as majority students at the schools at which they interfere. The way $MDA^m$ is defined departing from $DA^m$ is inspired by the way $EADAM$ in Kesten (2010) is defined departing from the deferred acceptance rule.

$MDA^m$ is *minimally responsive* but $DA^m$ is not. Moreover, we show that when $DA^m$ is used and we switch to a stronger affirmative action, if no minority student benefits, then also no majority student benefits (Proposition 3). That is, whenever $DA^m$ perversely responses to affirmative action, there is an efficiency loss in the Pareto sense. Moreover, we show that when $DA^m$ is used, the "efficiency loss" due to a stronger affirmative action may be severe (Proposition 4). Hafalir et al. (2013), through simulations, show that affirmative action under $DA^m$ brings considerable welfare gains to the minority students on average. We show that at each problem, $MDA^m$ either Pareto dominates $DA^m$ or gives the same assignment (Theorem 4), implying that all the welfare gains for the minority students are preserved.

$MDA^m$ achieves a type of affirmative action where the only difference from the reserve-type affirmative action is that a minority student may be excluded from a school although the reserve is not exhausted, provided that there is no fair assignment at which at least one minority student is better off and no minority student is worse off. We call this "conditional-reserve–type affirmative action". We show that $MDA^m$ is not Pareto dominated (either for the minority or for all the agents) by any fair rule that achieves this type of affirmative action (Theorem 3).

One important deficiency of $MDA^m$ is that it is not strategy-proof, although $DA^m$ is. Yet, we show that no fair rule is *minimally responsive*, *strategy-proof*, and achieves any type of affirmative action we have mentioned (Theorem 5). Moreoever, we propose the notion of *minimal fairness*, which is a requirement that, we argue, should be satisfied by a fair rule achieving a minimum level of affirmative action. We show that no rule is *minimally fair*, *minimally responsive*, and *strategy-proof* (Theorem 6). This result suggests that the previous impossibilities are not due to the particular choice of affirmative action policies but rather to a general incompatibility of *minimal responsiveness* and *strategy-proofness* under affirmative action.

## 2 Related literature

The literature on school choice with affirmative action has so far been silent on *minimal responsiveness* except for Kojima (2012). To our knowledge, our paper is the first to propose a fair and *minimally responsive* affirmative action rule. However, several theoretical studies have proposed assignment rules intended for affirmative action with desirable fairness and strategic properties. Abdulkadiroğlu and Sönmez (2003) introduce a version of the deferred acceptance algorithm to the affirmative action setting and discuss such properties. Ehlers et al. (2014) study quota type affirmative action policies by incorporating both upper and lower type-specific bounds, and also allowing for more than two types of students. A part of the literature, including Echenique and Yenmez (2014), Kominers and Sönmez (2013), Westkamp (2013), and Erdil and Kumano (2012), approaches the affirmative action problem from the perspective of designing priorities or choice functions for the schools. In particular, another type of affirmative action that is not analyzed in this paper, is implemented by raising the priorities of the minority students relative to the majority students, while maintaining the priority orderings within each type. In this case, a stronger affirmative action means that for each minority student and for each school, each student who used to have a lower priority than her at that school still has a lower priority, and possibly a student who did not have a lower priority now has a lower priority. This type of affirmative action is also known to suffer from the same problem: fairness is not compatible with *minimal responsiveness* (Kojima (2012)). On the experimental side, Klijn et al. (2014) provide comparative analysis of quota-type and reserve-type affirmative action policies, the main focus being on the strategic behavior of students.

As we have noted before, the way $MDA^m$ is defined is very similar to the way $EADAM$ in Kesten (2010) is defined. The deferred acceptance rule, in case of no affirmative action, is not Pareto efficient and $EADAM$ is proposed to remove its inefficiency by identifying the students that cause the inefficiency and treating them differently. On the other hand, $DA^m$ is not minimally responsive and we propose $MDA^m$ to remove this deficiency by identifying the students that cause it and treating them differently. Yet, there are several important differences. The setup in which $EADAM$ is defined does not incorporate different types of students and affirmative action. An interferer in that setup is treated by removing from her preference ordering the school at which she interferes. By contrast in our setup, an interferer, who is necessarily a minority student, is treated as a majority student at the school she interferes. Although minimal responsiveness is not completely symmetric to Pareto efficiency in that setup, and the way we treat the interferers is also not completely symmetric, most of the results in Kesten (2010) have counterparts in our setup.

## 3 School choice problems with affirmative action

Let $\boldsymbol{S}$ be a finite set of students. There are two types of students, minority students and majority students. Let $S^m$ and $S^M$ denote the sets of minority and majority students, respectively. They are nonempty sets such that $S^m \cup S^M = S$ and $S^m \cap S^M = \emptyset$. Let $\boldsymbol{C}$ be a finite set of schools. Suppose that $|S|, |C| \geq 2$. For each $s \in S$, student $s$ has a strict (i.e. complete, transitive, and anti-symmetric) preference relation $\boldsymbol{R_s}$ over $C \cup \{s\}$, where $s$ denotes her outside option, which can be attending a private school or homeschooling. Let $\mathcal{R}$ be the set of all such preference relations. Let $\boldsymbol{P_s}$ be the strict relation associated with $R_s$. Let $\boldsymbol{R} \equiv (R_s)_{s\in S} \in \mathcal{R}^S$ be a preference profile.

Each school $c \in C$ has a strict priority relation $\succeq_c$ over $S$. Let $\boldsymbol{\succ_c}$ be the strict relation associated with $\succeq_c$. Let $\boldsymbol{\succeq} \equiv (\succeq_c)_{c\in C}$ be the priority profile.

School $c$ can admit up to a certain number of students, its capacity. Let $\boldsymbol{q_c} \in \mathbb{N}$ be the capacity of school $c$. Let $\boldsymbol{q} \equiv (q_c)_{c\in C} \in \mathbb{N}^C$ be the capacity profile.

For each school, there is an affirmative action parameter, denoted by $r_c$, such that $r_c \in \mathbb{N}$ and $r_c \leq q_c$. We incorporate the parameter to analyze affirmative action policies in a unified framework. The use of $r$ will be clarified shortly when we discuss different types of affirmative action. Yet, to have an idea, $r_c$ is the number of seats at $c$ at which the minority students are favored.

In short, a school choice problem with affirmative action, or simply a **problem**, is a list $(S^m, S^M, C, R, \succeq, q, r)$ such that $r \leq q$. Since $(S^m, S^M, C, \succeq, q)$ will be fixed, unless otherwise noted, a problem is simply a pair $(R, r)$.

A matching is an assignment of students to schools such that each student is assigned to a school or to her outside option, and no more students are assigned to a school than its capacity. Formally, a **matching $\boldsymbol{\mu}$** is a mapping from $S \cup C$ to the subsets of $S \cup C$ such that

i. for each $s \in S$, $\mu(s) \in C \cup \{s\}$,
ii. for each $c \in C$, $\mu(c) \subseteq S$ and $|\mu(c)| \leq q_c$, and
iii. for each $s \in S$ and $c \in C$, $\mu(s) = c$ if and only if $s \in \mu(c)$.

For each matching $\mu$ and school $c$, let **$\boldsymbol{\mu^m(c)}$ and $\boldsymbol{\mu^M(c)}$ denote the sets of minority and majority students assigned to $\boldsymbol{c}$ at $\boldsymbol{\mu}$**.

An affirmative action rule, or simply a **rule**, chooses a matching for each problem.

Let $\mu$ and $\mu'$ be matchings. Let $R \in \mathcal{R}^S$. The matching **$\boldsymbol{\mu}$ Pareto dominates $\boldsymbol{\mu'}$ at $\boldsymbol{R}$** if for each $s \in S$, $\mu(s)\ R_s\ \mu'(s)$, strict preference holding for at least one $s \in S$. A matching is **Pareto efficient** if it is not *Pareto dominated* by any other matching. Due to affirmative action concerns, we are also interested in comparing two matchings with respect to the minority students' welfare. The matching **$\boldsymbol{\mu}$ Pareto dominates $\boldsymbol{\mu'}$ for the minority at $\boldsymbol{R}$** if for each $m \in S^m$, $\mu(m)\ R_m\ \mu'(m)$, strict preference holding for at least one $s \in S^m$.

A central objective in design for school choice is fairness. When there is no affirmative action, that is when the student types and the affirmative action parameter are ignored, fairness corresponds to the following. Given a problem $R$, a matching $\mu$, and a student-school pair $(s, c)$, the **priority of $\boldsymbol{s}$ is violated at $\boldsymbol{c}$** if $s$ prefers $c$ to $\mu(s)$ and a student $s'$ with a lower priority is assigned to $c$ at $\mu$. We also say that the priority of $s$ is violated by $s'$ at $c$. A matching **$\boldsymbol{\mu}$ is fair** if the following conditions are satisfied.[6]

1. No priority violation: No students' priority at any school is violated.
2. Outside option lower bound: No student prefers her outside option to her assignment.
3. Non-wastefulness: If a student prefers a school $c$ to her assignment, then the capacity of $c$ is exhausted, i.e. $|\mu(c)| = q_c$.

A **rule is fair** if it chooses a *fair* matching at each problem.

We now introduce two types of affirmative action policies. The first policy, which is very common in practice and known as quota-type affirmative action policy, considers, for each school $c$, the parameter $r_c$ as the number of seats at school $c$ that can be assigned only to the minority students. The second policy, which we call reserve-type affirmative action policy, considers the affirmative action parameter at each school as the number of seats that are reserved for the minority students, but can be assigned to majority students provided that no minority student

[6] In the matching literature, an $(s, c)$ pair such that the priority of $s$ is violated at $c$ is usually called a blocking pair, and a fair matching is also called stable.

prefers that school to her assigned school.[7] When there is an affirmative action policy, an objective is to achieve affirmative action in a fair way. Note that each affirmative action policy imposes constraints and hence the fairness requirement for the no affirmative action case should be modified accordingly. Next, we introduce two fairness requirements, namely *fairness with quota* and *fairness with reserve*, that formalize the modifications.

### 3.1 Fairness with quota

In the quota-type affirmative action policy, for each school $c$, the parameter $r_c$ represents the number of seats at school $c$ that can be assigned only to the minority students. In other words, $c$ is allowed to admit only up to $q_c - r_c$ majority students; the difference $q_c - r_c$ is called its majority-quota.

Let $(R, r)$ be a problem. A matching $\mu$ is fair with respect to quota-type affirmative action, or simply ***fair with quota***, if the following conditions are satisfied.[8]

1. No school admits more majority students than its majority-quota.
2. If there are $s, s' \in S$ and $c \in C$ such that the priority of $s$ is violated by $s'$ at $c$, then $s \in S^M$, $s' \in S^m$, and the majority quota of $c$ is met at $\mu$, i.e. $|\mu^M(c)| = q_c - r_c$.
3. No student prefers her outside option to her assignment.
4. If a student prefers a school $c$ to her assignment, then either $s \in S^m$ and $|\mu(c)| = q_c$ or $s \in S^M$ and $|\mu^M(c)| = q_c - r_c$.

A **rule is fair with quota** if it chooses, at each problem, a matching that is *fair with quota*. Note that if a rule is fair with quota, then it achieves the quota-type affirmative action in a fair way.

### 3.2 Fairness with reserve

The reserve-type affirmative action policy considers the affirmative action parameter at each school as the number of seats that are reserved for the minority students. Here, a school is allowed to assign some of its reserved seats to majority students provided that no minority student prefers that school to her assigned school.

Let $(R, r)$ be a problem. A matching $\mu$ is fair with respect to the reserve-type affirmative action, or simply ***fair with reserve*** if the following conditions are satisfied.[9]

[7] Hafalir et al. (2013) introduces an assignment rule that achieves reserve-type affirmative action.

[8] In Kojima (2012), quota requirement and modified fairness requirement are considered separately and he refers to fairness as stability. A rule being *fair with quota* in our setup is technically equivalent to saying that a rule is *stable* in Kojima (2012) setup under quota-type affirmative action.

[9] *Fairness with reserve* is equivalent to the *stability* requirement in Hafalir et al. (2013).

1. If there are $s, s' \in S$ and $c \in C$ such that the priority of $s$ is violated by $s'$ at $c$, then $s \in S^M$, $s' \in S^m$, and the minority reserve at $c$ is unexceeded at $\mu$, i.e. $|\mu^m(c)| \leq r_c$.
2. There are no $m \in S^m$ and $c \in C$ such that $m$ prefers $c$ to $\mu(m)$ and the minority reserve at $c$ is not exhausted, i.e. $|\mu^m(c)| < r_c$.
3. No student prefers her outside option to her assignment.
4. If a student prefers a school $c$ to her assignment, then the capacity of $c$ is exhausted, i.e. $|\mu(c)| = q_c$.

A **rule is fair with reserve** if it chooses, at each problem, a matching that is *fair with reserve*. Note that if a rule is fair with reserve, then it achieves the reserve-type affirmative action in a fair way.

# 4 Impossibilities with quota or reserve

The following requirement is essential for a rule that is operating in line with the intention of affirmative action: consider any school choice problem with affirmative action; consider the assignment recommended by the rule; consider another problem obtained by weakly increasing the affirmative action parameter at each school; consider the new assignment recommended by the rule; if a minority student is worse off, then at least one other minority student should be better off. If a rule satisfies this requirement, then we say that it is minimally responsive to affirmative action, or simply *minimally responsive*. Formally, a rule $\boldsymbol{\varphi}$ is **minimally responsive** if there are no $R \in \mathcal{R}^S$, and $r, r' \in \mathbb{N}$ such that $r' \geq r$ and $\varphi(R, r)$ Pareto dominates $\varphi(R, r')$ for the minority at $R$. It is **perversely responsive** if it is not *minimally responsive*.

It turns out that no rule is *fair with quota* and *minimally responsive* (Kojima (2012)). Of course, this impossibility result is on the full domain of school choice problems. On the other hand, there are restricted domains of problems where we have possibility. For instance, on a domain of problems such that for each problem $S^m = \emptyset$, then any rule that is *fair with quota* is also trivially *minimally responsive*. Note that the policy makers in school districts can observe $S^m$, $S^M$, $C$, $\succ$, and $q$, although they can not observe the preferences of the students. Then, a natural question is the following: for which $(S^m, S^M, C, \succ, q)$ lists is there a rule that is *fair with quota* and *minimally responsive*? Providing an answer to this question serves two purposes:

1. Understanding the degree and the structure of incompatibility: if such lists are "common" among all possible lists, the incompatibility is arguably not severe, and that can justify using a rule which is *fair with quota*. Also, understanding the structure of such lists may yield new types of affirmative action that do not suffer from the problem;
2. Policy recommendation: the central authority does not know the preferences of the students. However, it has the information and also, to some extent, control over:

(a) which student is of which type: in some universities, the academic achievements of a student are examined in light of income, and "income level thresholds" can be controlled;

(b) the priority profile: in some states, the school priorities depend on the "attendance area" of the schools, which can be controlled;

(c) capacity profile: the capacity of a school usually does not represent the exact number of students that it can accommodate, so that can be controlled too.

Based on the answer to the aforementioned question, we may recommend ways to control such parameters in a way that permits us to define rules that are fair and *minimally responsive*.

The following notion is essential for the answer. For each $s \in S$, each $c \in C$, and each priority ordering $\succeq_c$, let $U_c^{\succeq}(s) \equiv \{s' \in S : s' \succeq_c s\}$ denote the set of students consisting of $s$ and all students with higher priority. A list $\boldsymbol{(S^m, S^M, C, \succ, q)}$ **gives full priority to the minority** if there are no $m \in S^m$, $M \in S^M$, and $c \in C$ such that $[M \succ_c m$ and $|U_c^{\succeq}(m)| \geq q_c+1]$. Observe that if $(S^m, S^M, C, \succ, q)$ *gives full priority to the minority*, then at each school $c$, either each minority student is ranked above each majority student, or each minority student is one of the $q_c$ highest-priority students.

**Lemma 1.** *Suppose that $(S^m, S^M, C, \succ, q)$ gives full priority to the minority. Let $R \in \mathcal{R}^S$ and $r, r' \leq q$. For each matching $\mu$ which is fair with quota at $(R, r)$, there is a matching $\mu'$ which is fair with quota at $(R, r')$ such that each minority student is assigned to the same school at $\mu$ and $\mu'$.*

*Proof.* See Appendix 8.1. □

The following theorem states that the existence of a rule that is *fair with quota* and *minimally responsive* requires that $(S^m, S^M, C, \succ, q)$ give full priority to the minority.

**Theorem 1.** *There is a rule that is fair with quota and minimally responsive if and only if $(S^m, S^M, C, \succ, q)$ gives full priority to the minority.*

*Proof.* See Appendix 8.2. □

Corollary 1, which directly follows from Theorem 1, shows that with a mild assumption, the impossibility disappears only when the minority students already have priority over majority students.

**Corollary 1.** *Suppose that the number of minority students is greater than the capacity of each school. Then, there is a rule which is fair with quota and minimally responsive if and only if each minority student has priority over each majority student at each school.*

It turns out that also no rule is *fair with reserve* and *minimal responsiveness.*

**Proposition 1.** *No rule is fair with reserve and minimally responsive.*

*Proof.* See Appendix 8.3. □

Next, we characterize the lists $(S^m, S^M, C, \succ, q)$ for which *fairness with reserve* is compatible with *minimal responsiveness.* The following notions are essential for this purpose. Let $s_1, s_2 \in S$, $c \in C$, $N \subseteq S$. The pair $\boldsymbol{(s_1, N)}$ **is a threat for** $\boldsymbol{s_2}$ **at** $\boldsymbol{c}$ **and** $\boldsymbol{(q, \succeq)}$ if $|N| = q_c - 1$ and

**either:** $s_2 \in S^m$, $s_1 \succ_c s_2$, and $N \subseteq U_c^{\succeq}(s_2) \setminus \{s_1, s_2\}$.

**or:** $s_2 \in S^M$, $N \cap S^M \subseteq U_c^{\succeq}(s_2) \setminus \{s_1, s_2\}$, and if $s_1 \in S^M$ then $s_1 \succ_c s_2$.

The intuition is that, if $(s_1, N)$ is a threat for $s_2$ at $c$ and $(q, \succeq)$, then there is $(R, r)$ such that the students $s_2$, $s_1$, and $N$ compete for a seat at $c$ and $s_2$ does not get a seat at $c$. Let $\boldsymbol{T(s, c)}$ **denote the set of threats for** $\boldsymbol{s}$ **at** $c$ **and** $\boldsymbol{(q, \succeq)}$.

A list $\boldsymbol{(S^m, S^M, C, \succ, q)}$ **has a cycle** if there is a pair of minority students $m, m' \in S^m$, a majority student $M \in S^M$, a list of students $s_1, \ldots, s_{k-2} \in S$, a list of schools $c_1, \ldots, c_k \in C$, and a list of disjoint sets of students $N_1, \ldots, N_k \subseteq S$ such that

1) $m \succ_{c_1} m'$, $M \succ_{c_1} m'$, $M \succ_{c_2} s_1$, $m \notin N_1$, and

2) $(m', N_1) \in T(M, c_1)$, $(M, N_2) \in T(s_1, c_2)$, $(s_k, N_k) \in T(m, c_k)$, $(m, N_1) \in T(m', c_1)$, and for each $t \in \{3, \ldots, k-1\}$, $(s_{t-2}, N_t) \in T(s_{t-1}, c_t)$.

A list $(m, m', M, s_1, \ldots, s_{k-2}, c_1, \ldots, c_k, N_1, \ldots, N_k)$ satisfying 1 and 2 is a cycle of length $k$. A list $\boldsymbol{(S^m, S^M, C, \succ, q)}$ **is acyclic** if it has no cycle.

**Theorem 2.** *There is a rule that is fair with reserve and minimally responsive if and only if* $(S^m, S^M, C, \succ, q)$ *is acyclic.*

*Proof.* See Appendix 8.6. □

Proposition 2 shows how restrictive *acyclicity* is.

**Proposition 2.** *Suppose that the number of minority students is greater than the total capacity of the two smallest schools, i.e. for each pair* $c, c' \in C$, $|S^m| > q_c + q_{c'}$. *Let* $M \in S^M$. *If* $(S^m, S^M, C, \succ, q)$ *is acyclic, then there are at least* $|S^m| - 1$ *minority students each of whom has higher priority than* $M$ *at at least* $|C| - 1$ *schools.*

*Proof.* See Appendix 8.7. □

In practice, it is very unlikely that the minority students mostly have priority over the majority students. Therefore, the incompatibilities are not exceptional, but pervasive. The above results also reveal two striking facts:

1. The school choice environments in which the incompatibility disappears are almost the same for the two types of affirmative action.
2. The environments at which the minority students mostly have priority over the majority students are the environments at which we don't really need affirmative action, since any reasonable assignment rule respecting the priorities will favor them. Therefore, for either type of affirmative action, the only situations at which the incompatibility disappears, are the ones at which we don't need affirmative action.

In what follows, we will look for a rule that is *minimally responsive*. Our departure point is a prominent rule that is *fair with reserve*, namely the **deferred acceptance rule with minority reserves, $\boldsymbol{DA^m}$**, (Hafalir et al. (2013)). It is defined through the following algorithm.

$\boldsymbol{DA^m}$ algorithm runs as follows. Let $(R, r)$ be given.

**Step 1.** Each student $s$ applies to her top-ranked school. Each student applying to her outside options is assigned to her outside option. Each school $c$ considers its applicants. Among these, it tentatively accepts up to $r_c$ minority students, those with the highest $\succeq_c$-priorities. Then, among the remaining applicants it tentatively accepts those with the highest $\succeq_c$-priorities until its capacity is filled or it runs out of applicants. It rejects all other applicants. If there is no rejection by any school at this step, then stop.

**Step $\boldsymbol{k \geq 2}$.** Each student $s$ who is rejected at Step $k-1$ applies to her top-ranked school from among the ones that have not rejected her, possibly her outside option $s$. Each student applying to her outside options is assigned to her outside option. Each school $c$ considers the students it tentatively accepted at Step $k-1$ and its new applicants at Step $k$. Among these, it tentatively accepts up to $r_c$ minority students, those with the highest $\succeq_c$-priorities. Then, it tentatively accepts applicants with the highest $\succeq_c$-priorities until its capacity is filled or it runs out of applicants. It rejects all other applicants. If there is no rejection by any school at this step, then stop.

The algorithm stops in finite time. For each problem $(R, r)$, the outcome of the above algorithm is the outcome chosen by $DA^m$. We denote it by $DA^m(R, r)$.

Hafalir et al. (2013) considers a weakening of *minimal responsiveness* which requires that not all the minority students are worse off when the minority reserve at each school weakly increases. $DA^m$ satisfies this requirement (Hafalir et al. (2013)). However, by Proposition 1, it may be that when the minority reserve at each school weakly increases, then $DA^m$ chooses a matching that is Pareto inferior for the minority. In other words, under a stronger affirmative action, $DA^m$ may hurt some minority students without benefiting any minority student. Proposition 3 shows that the perverse effects of a stronger affirmative action are even more severe: when $DA^m$ is used, whenever no minority student benefits from a stronger affirmative action, also no majority student does.

**Proposition 3.** *Let $R \in \mathcal{R}^S$ and $r, r' \in \mathbb{N}^{\mathbb{C}}$. If $r \leq r'$ and $DA^m(R, r)$ Pareto dominates $DA^m(R, r')$ at $R$ for the minority, then $DA^m(R, r)$ Pareto dominates $DA^m(R, r')$ at $R$.*

*Proof.* See Appendix 8.4. □

The following example shows that not only might a stronger affirmative action result in a Pareto inferior matching, but the associated efficiency loss might be severe.

**Example 1.** *Let $S^m \equiv \{m_1, m_2, \dots, m_7\}$, $S^M \equiv \{M_1, M_2, \dots, M_8\}$, $C \equiv \{c_1, c_2, \dots, c_8\}$. Let $(q_{c_1}, \dots, q_{c_8}) \equiv (2, 4, 2, 2, 2, 2, 1, 1)$, $(r_{c_1}, \dots, r_{c_8}) \equiv (0, 0, \dots, 0)$, and $(r'_{c_1}, \dots, r'_{c_8}) \equiv (0, 2, 0, \dots, 0)$. Let $\succeq$ and $R$ be as depicted below. Each student prefers each school to her outside option. We only depict the preferences over schools.*

| $\succ_{c_1}$ | $\succ_{c_2}$ | $\succ_{c_3}$ | $\succ_{c_4}$ | $\succ_{c_5}$ | $\succ_{c_6}$ | $\succ_{c_7}$ | $\succ_{c_8}$ |
|---|---|---|---|---|---|---|---|
| $M_1$ | $M_8$ | $M_1$ | $M_3$ | $M_5$ | $M_6$ | $M_7$ | $m_5$ |
| $M_2$ | $M_1$ | $M_2$ | $M_4$ | $m_3$ | $m_4$ | $m_7$ | $m_6$ |
| $M_3$ | $m_6$ | $M_3$ | $M_5$ | $M_6$ | $M_7$ | $m_5$ | $m_7$ |
| $m_1$ | $M_2$ | $M_4$ | $m_3$ | $m_4$ | $m_7$ | $M_1$ | $M_1$ |
| $M_4$ | $m_7$ | $M_5$ | $M_1$ | $M_1$ | $M_1$ | $M_2$ | $M_2$ |
| $M_5$ | $m_1$ | $M_6$ | $M_2$ | $M_2$ | $M_2$ | $M_3$ | $M_3$ |
| $M_6$ | $m_2$ | $M_7$ | $M_6$ | $M_3$ | $M_3$ | $M_4$ | $M_4$ |
| $m_2$ | $M_3$ | $M_8$ | $M_7$ | $M_4$ | $M_4$ | $M_5$ | $M_5$ |
| $M_7$ | $M_4$ | $m_1$ | $M_8$ | $M_7$ | $M_5$ | $M_6$ | $M_6$ |
| $M_8$ | $M_5$ | $m_2$ | $m_1$ | $M_8$ | $m_1$ | $m_1$ | $M_7$ |
| $m_3$ | $M_6$ | $m_3$ | $m_2$ | $m_1$ | $m_2$ | $m_2$ | $M_8$ |
| $m_4$ | $M_7$ | $m_4$ | $m_4$ | $m_2$ | $m_3$ | $m_3$ | $m_1$ |
| $m_5$ | $m_3$ | $m_5$ | $m_5$ | $m_5$ | $m_5$ | $m_4$ | $m_2$ |
| $m_6$ | $m_4$ | $m_6$ | $m_6$ | $m_6$ | $m_6$ | $m_4$ | $m_3$ |
| $m_7$ | $m_5$ | $m_7$ | $m_7$ | $m_7$ | $M_8$ | $M_8$ | $m_4$ |

| $R_{m_1}$ | $R_{m_2}$ | $R_{m_3}$ | $R_{m_4}$ | $R_{m_5}$ | $R_{m_6}$ | $R_{m_7}$ | $R_{M_1}$ | $R_{M_2}$ | $R_{M_3}$ | $R_{M_4}$ | $R_{M_5}$ | $R_{M_6}$ | $R_{M_7}$ | $R_{M_8}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $c_2$ | $c_2$ | $\underline{c_4}$ | $\underline{c_5}$ | $\underline{c_7}$ | $\underline{c_8}$ | $\underline{c_6}$ | $\underline{c_2}$ | $\underline{c_2}$ | $\underline{c_3}$ | $\underline{c_3}$ | $\underline{c_4}$ | $\underline{c_5}$ | $\underline{c_6}$ | $c_7$ |
| $c_3$ | $c_3$ | $c_6$ | $c_7$ | $c_2$ | $c_3$ | $c_3$ | $c_4$ | $c_4$ | $c_5$ | $c_5$ | $c_6$ | $c_7$ | $c_8$ | $c_8$ |
| $c_4$ | $c_4$ | $c_7$ | $c_8$ | $c_3$ | $c_4$ | $c_4$ | $c_5$ | $c_5$ | $c_6$ | $c_6$ | $c_7$ | $c_8$ | $c_2$ | $c_6$ |
| $c_5$ | $c_5$ | $c_8$ | $c_2$ | $c_4$ | $c_5$ | $c_5$ | $c_6$ | $c_6$ | $c_7$ | $c_7$ | $c_8$ | $c_2$ | $c_3$ | $c_3$ |
| $c_6$ | $c_6$ | $c_2$ | $c_3$ | $c_5$ | $c_6$ | $c_8$ | $c_7$ | $c_7$ | $c_8$ | $c_8$ | $c_2$ | $c_3$ | $c_4$ | $c_4$ |
| $c_7$ | $c_7$ | $c_3$ | $c_4$ | $c_6$ | $c_7$ | $c_7$ | $c_8$ | $c_8$ | $c_2$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_5$ |
| $c_8$ | $c_8$ | $\boxed{c_5}$ | $\boxed{c_6}$ | $\boxed{c_8}$ | $\boxed{c_2}$ | $\boxed{c_2}$ | $\boxed{c_3}$ | $\boxed{c_3}$ | $\boxed{c_4}$ | $\boxed{c_4}$ | $\boxed{c_5}$ | $\boxed{c_6}$ | $\boxed{c_7}$ | $\underline{\boxed{c_2}}$ |
| $\underline{\underline{\boxed{c_1}}}$ | $\underline{\underline{\boxed{c_1}}}$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ |

*$DA^m(R, r)$ and $DA^m(R, r')$ are represented by the underlined and boxed matchings, respectively. At $DA^m(R, r)$, $m_1$ and $m_2$ are assigned to their least-preferred schools, $M_8$ is assigned to*

*her second least-preferred school, and all the other students are assigned to their most-preferred schools. At $DA^m(R, r')$, $m_1$ and $m_2$ are still assigned to their least-preferred schools, $M_8$ is still assigned to her second least-preferred school. However, all the other students are assigned to their second least–preferred schools.*

Proposition 4 shows that for an arbitrary set of schools and an arbitrary capacity profile, there are school choice problems where the efficiency loss due to a stronger affirmative action under $DA^m$ is severe. Moreover, the number of minority students who are hurt may be high.

**Proposition 4.** *Let $C = \{c_1, \ldots, c_K\}$ be a set of schools and $q$ be a capacity profile. Let $K \geq 3$ and $q_1 \geq q_2 \geq \cdots \geq q_K$. Let $k \in \mathbb{N}$ be such that $q_3 \leq k \leq q_3 + q_4 + \cdots + q_K$. There is a set of minority students $S^m$, a set of majority students $S^M$, a preference profile $R$, a priority profile $\succeq$, and a pair of minority reserve profiles $r \leq r' \leq q$, with the following properties.*

*(1) Each student prefers each school to her outside option. Each student is assigned to a school at both $DA^m(R, r)$ and $DA^m(R, r')$.*

*(2) There are $k$ minority students and $(2q_{c_3} + q_{c_4} + \cdots + q_{c_K}) - k$ majority students who are assigned to their most-preferred schools at $DA^m(R, r)$ and second least-preferred schools at $DA^m(R, r')$. All the remaining students are assigned to the same school in both matchings.*

*Proof.* See Appendix 8.5. □

## 5 A minimally responsive affirmative action rule

Theorem 2 provide insights into why no rule is *fair with reserve* and *minimally responsive*, and in particular why $DA^m$ is *perversely responsive*: a minority student who has a lower priority than a majority student at a school, is temporarily accepted by that school while the majority student is rejected. However, the majority student being rejected initiates a sequence of further rejections that may end with the minority student being rejected by the same school. In a sense, a minority student "interferes" with the admission process of a school. In the light of these observations, one would think that "not favoring" a minority student at a school at which she is an interferer may remove some of the deficiencies of $DA^m$. In what follows, we implement this idea. We propose a modification of $DA^m$ such that some minority students are treated as majority students at the schools they interfere.

Let $m \in S^m$, $c \in C$. Let $(R, r)$ be a problem. The minority student $\boldsymbol{m}$ **is an interferer for $\boldsymbol{c}$ at $\boldsymbol{(R, r)}$**[10] if there is a majority student $M \in S^M$ with $M \succ_c m$ and a pair of steps $t, t'$

[10] The notion of an "interferer" is very similar to the notion of an "interrupter" due to Kesten (2010). In Kesten (2010), an interrupter is a student who initiates a rejection cycle and possibly causes the Deferred acceptance rule to be inefficient.

of the $DA^m(R, r)$ algorithm such that at Step $t$, $m$ is tentatively accepted and $M$ is rejected by $c$, and at Step $t'$, $m$ is rejected by $c$. We call $(m, c)$ an **interfering pair** of Step $t'$ at $(R, r)$. Let **$\boldsymbol{IP(R, r)}$ denote the set of interfering pairs at $\boldsymbol{(R, r)}$**. Given $k \in \mathbb{N}$, let **$\boldsymbol{IP_k(R, r)}$ denote the set of interfering pairs of Step $\boldsymbol{k}$ at $\boldsymbol{(R, r)}$**.

We modify the $DA^m$ algorithm in the following way. Let $k$ be the last step in $DA^m(R, r)$ algorithm at which there is an interfering pair. We run $DA^m$ algorithm such that for each $(m, c) \in IP(R, r)$ which is interfering at Step $k$, the interfering minority $m$ is treated as a majority student at school $c$. Then, we determine the last step at which there is an interferer in this new algorithm, and treating those minorities as majorities at the schools they interfere we run $DA^m$ algorithm again. We repeat this procedure until there is no new interfering pair.[11]

Before we provide the formal description of our algorithm, we introduce it by means of an example. Let $(S, C, \succeq, q, R, r)$ be the problem defined as follows. Let $S^m \equiv \{m_1, m_2, m_3, m_4\}$, $S^M \equiv \{M_1, M_2, M_3\}$, $C \equiv \{c_1, c_2, c_3, c_4, c_5, c_6\}$. Let $(q_{c_1}, q_{c_2}, q_{c_3}, q_{c_4}, q_{c_5}, q_{c_6}, q_{c_7}, q_{c_8}) = (1, 1, 1, 2, 1, 1, 1, 2)$ and $r = (1, 0, 1, 0, 1, 0, 1, 0)$. Let $\succeq$ and $R$ be as depicted as follows. Only the top parts of the profiles that are relevant to the example are depicted.

| $\succ_{c_1}$ | $\succ_{c_2}$ | $\succ_{c_3}$ | $\succ_{c_4}$ | $\succ_{c_5}$ | $\succ_{c_6}$ | $\succ_{c_7}$ | $\succ_{c_8}$ |
|---|---|---|---|---|---|---|---|
| $M_1$ | $M_1$ | $M_2$ | $M_2$ | $M_3$ | $M_3$ | $M_4$ | $M_4$ |
| $m_3$ | $m_2$ | $m_2$ | $m_1$ | $m_5$ | $m_5$ | $m_6$ | $m_4$ |
| $m_1$ |  | $m_3$ |  | $m_4$ |  |  |  |

| $R_{m_1}$ | $R_{m_2}$ | $R_{m_3}$ | $R_{m_4}$ | $R_{m_5}$ | $R_{m_6}$ | $R_{M_1}$ | $R_{M_2}$ | $R_{M_3}$ | $R_{M_4}$ |
|---|---|---|---|---|---|---|---|---|---|
| $c_1$ | $c_2$ | $c_3$ | $c_5$ | $c_6$ | $c_7$ | $c_1$ | $c_3$ | $c_5$ | $c_7$ |
| $c_4$ | $c_3$ | $c_1$ | $c_8$ | $c_5$ | $c_6$ | $c_2$ | $c_4$ | $c_6$ | $c_8$ |

The $DA^m(R, r)$ algorithm is depicted below.

| Step | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $c_8$ |
|---|---|---|---|---|---|---|---|---|
| 1 | $\boxed{m_1}$,$M_1$ | $\boxed{m_2}$ | $\boxed{m_3}$, $M_2$ |  | $\boxed{m_4}$, $M_3$ | $\boxed{m_5}$ | $\boxed{m_6}$,$M_4$ |  |
| 2 |  | $\boxed{M_1}$,$m_2$ |  | $\boxed{M_2}$ |  | $\boxed{M_3}$,$m_5$ |  | $\boxed{M_4}$ |
| 3 |  |  | $\boxed{m_2}$,$m_3$ |  | $\boxed{m_5}$,$m_4$ |  |  |  |
| 4 | $\boxed{m_3}$,$m_1$ |  |  |  |  |  |  | $\boxed{m_4}$, $\boxed{M_4}$ |
| 5 |  |  |  | $\boxed{m_1}$,$\boxed{M_2}$ |  |  |  |  |

There are three interfering pairs at $(R, r)$: $(m_3, c_3)$ and $(m_4, c_5)$ are interfering pairs at Step 3, and $(m_1, c_1)$ is an interfering pair of Step 4. So, the last step of the $DA^m(R, r)$ algorithm at

[11] Detecting only the last step interferers is crucial. In Section 8.12, we show that two other natural ways of removing the interferers do not work.

which there is an interfering pair is Step 4, with $(m_1, c_1)$ as the interfering pair . We then rerun the $DA^m$ algorithm by treating $m_1$ as a majority student at $c_1$.

| Step | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $c_8$ |
|---|---|---|---|---|---|---|---|---|
| 1 | $m_1, \boxed{M_1}$ | $\boxed{m_2}$ | $\boxed{m_3}, M_2$ | | $\boxed{m_4}, M_3$ | $\boxed{m_5}$ | $\boxed{m_6}, M_4$ | |
| 2 | | | | $\boxed{m_1}, \boxed{M_2}$ | | $\boxed{M_3}, m_5$ | | $\boxed{M_4}$ |
| 3 | | | | | $\boxed{m_5}, m_4$ | | | |
| 4 | | | | | | | | $\boxed{m_4}, \boxed{M_4}$ |

Now, there is only one interfering pair: $(m_4, c_5)$ is the interfering pair of Step 3. So, we run $DA^m$ algorithm by treating $m_4$ as a majority student at $c_5$, and still treating $m_1$ as a majority student at $c_1$.

| Step | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $c_8$ |
|---|---|---|---|---|---|---|---|---|
| 1 | $m_1, \boxed{M_1}$ | $\boxed{m_2}$ | $\boxed{m_3}, M_2$ | | $m_4, \boxed{M_3}$ | $\boxed{m_5}$ | $\boxed{m_6}, M_4$ | |
| 2 | | | | $\boxed{m_1}, \boxed{M_2}$ | | | | $\boxed{m_4}, \boxed{M_4}$ |

Now, there is no more interfering pair. Therefore, we stop, and the final outcome is the matching chosen by our modified $DA^m$.

Before we introduce the formal description of this modified $DA^m$, we define a few auxiliary notions. An **affirmative action problem with school specific criteria** is a triple $(f^m, R, r)$ such that $R \in \mathcal{R}^S$, $r \leq q$, and $f^m : C \to S$ is a correspondence that associates a (possibly empty) set of students with each school. For each $c \in C$, $f^m(c)$ indicates the set of minority students at $c$. Note that unlike an affirmative action problem, the students who are considered as minority students at two different schools may differ. Given such a problem $(f^m, R, r)$, let the $\boldsymbol{DA^m(f^m, R, r)}$ algorithm be defined as follows.

**Step 1.** Each student $s$ applies to her top-ranked school. Each student applying to her outside options is assigned to her outside option. Each school $c$ considers its applicants. It tentatively accepts up to $q_c^m$ students from among the $f^m(c)$ applicants with the highest $\succeq_c$-priorities. Then, among its remaining applicants it tentatively accepts those with the highest $\succeq_c$-priorities until its capacity is filled or it runs out of applicants. It rejects the others. If there is no rejection by any school at this step, then stop.

**Step $\boldsymbol{k \geq 2}$.** Each student $s$ who is rejected at Step $k-1$ applies to her top-ranked school from among the ones that have not rejected her. Each student applying to her outside options is assigned to her outside option. Each school $c$ considers the tentatively accepted students at Step $k-1$ and its new applicants at Step $k$. Each scool $c$ first tentatively accepts up to $q_c^m$

students from among the $f^m(c)$ applicants with the highest $\succeq_c$-priorities. Then, it tentatively accepts applicants with the highest $\succeq_c$-priorities until its capacity is filled or the applicants are exhausted. Those who are not tentatively accepted are rejected. If there is no rejection by any school at this step, then stop.

The above algorithm stops in finite time. Let $DA^m(f^m, R, r)$ be the outcome of the $DA^m(f^m, R, r)$ algorithm.

Student ***s* is an interferer for *c* at $(f^m, R, r)$** if $s \in f^m(c)$, there is a majority student, say $M \in S^M$, with $M \succ_c s$, and there are two steps of the $DA^m(f^m, r, r)$ algorithm, say $t$ and $t'$, such that at Step $t$, $s$ is tentatively accepted and $M$ is rejected by $c$, and at Step $t'$, $s$ is rejected by $c$. We call $(s, c)$ an interfering pair of Step $t'$ at $(f^m, R, r)$. Let **$IP(f^m, R, r)$ denote the set of interfering pairs at $(f^m, R, r)$**. Given $k \in \mathbb{N}$, let **$IP_k(f^m, R, r)$ denote the set of interfering pairs of Step *k* at $(f^m, R, r)$**.

**Modified Deferred Acceptance with Minority Reserves, $MDA^m$.** For each problem $(R, r)$, $MDA^m(R, r)$ algorithm runs as follows.

**Round 0** Set for each $c \in C$, $f_1^m(c) = S^m$.

**Round 1.** Run $DA^m(f_1^m, R, r)$. If there is no interfering pair, i.e. if $IP(f_1^m, R, r) = \emptyset$, then stop. Otherwise, let $k$ be the last step at which there is an interfering pair, i.e. $IP_k(f_1^m, R, r) \neq \emptyset$. For each $c \in C$, let $T(c, k)$ denote the set of students who are interfering at school $c$ at Step $k$, i.e. $T(c, k) = \{m \in S^m : (m, c) \in IP_k(f_1^m, R, r)\}$. For each $c \in C$, change the status of the students in $T(c, k)$ to majority students, i.e. set $f_2^m(c) = f_1^m(c) \setminus T(c, k)$, and move to round 2.

$\vdots$

**Round t.** Run $DA^m(f_t^m, R, r)$. If there is no interfering pair, i.e. if $IP(f_t^m, R, r) = \emptyset$, then stop. Otherwise, let $k$ be the last step at which there is an interfering pair, i.e. $IP_k(f_t^m, R, r) \neq \emptyset$. For each $c \in C$, let $T(c, k)$ denote the set of students who are interfering at school $c$ at Step $k$, i.e. $T(c, k) = \{m \in S^m : (m, c) \in IP_k(f_t^m, R, r)\}$. For each $c \in C$, change the status of the students in $T(c, k)$ to majority students, i.e. set $f_t^m(c) = f_{t-1}^m(c) \setminus T(c, k)$ and move to Step $t + 1$.

Note that whenever the algorithm moves to a next round, say from round $t$ to round $t + 1$, for at least one school, say $c$, we have $|f_{t+1}^m(c)| < |f_t^m(c)|$. Since we have finitely many students and schools, eventually at a round $T$ we will have for each $c \in C$, $f_T^m(c) = \emptyset$ unless the algorithm terminates at an earlier round. But then at Round $T$ there can not be any interfering pairs. Thus the algorithm stops in finitely many rounds.

The $MDA^m$, in a fair way, achieves the following type of affirmative action. This type of affirmative action, which we call conditional reserve-type affirmative action, considers the affirmative action parameter at each school as the number of seats that are reserved for the minority students. Here, a school is allowed to assign some of its reserved seats to majority students if no minority student prefers that school to her assigned school or there is no assignment that is *fair with reserve* at which at least one minority student is better off and no minority student is worse off.[12] So, the only difference from the reserve-type affirmative action is that a minority student may be excluded from a school although the reserve is not exhausted, provided that there is no assignment that is *fair with reserve* and Pareto dominates the former assignment for the minority.

A matching $\mu$ is fair with respect to the conditional-reserve–type affirmative action, or simply ***fair with conditional reserve***, if the following conditions are satisfied.[13]

1. If there are $m \in S^m$ and $M \in S^M$ such that $m$ prefers $c$ to $\mu(m)$, $M$ has a higher priority at $c$, and the minority reserve at $c$ is not exhausted, i.e. $|\mu^m(c)| < r_c$, then there is no matching that is *fair with reserve* and Pareto dominates $\mu$ for the minority.
2. if there are $s, s' \in S$ and $c \in C$ such that the priority of $s$ is violated by $s'$ at $c$, then $s \in S^M$, $s' \in S^m$, and the minority reserve at $c$ is not exceeded at $\mu$, i.e. $|\mu^m(c)| \leq r_c$.
3. No student prefers her outside option to her assignment.
4. If a student prefers a school $c$ to her assignment, then the capacity of $c$ is exhausted, i.e. $|\mu(c)| = q_c$.

A **rule is fair with conditional reserve** if it chooses, at each problem, a matching that is *fair with conditional reserve*. Note that if a rule is fair with conditional reserve, then it achieves the conditional-reserve–type affirmative action in a fair way. Theorem 3 shows that $MDA^m$ is *fair with conditional reserve*, and also lists some other properties of $MDA^m$.

**Theorem 3.** ***3-a.*** *$MDA^m$ is fair with conditional reserve and minimally responsive.*

***3-b.*** *Let $(R, r)$ be a problem. No matching is fair with conditional reserve and Pareto dominates $MDA^m(R, r)$ for the minority. Also, no matching is fair with conditional reserve and Pareto dominates $MDA^m(R, r)$.*

***3-c.*** *Let $(R, r)$ be a problem. No matching is fair and Pareto dominates $MDA^m(R, r)$ for the minority. Also, no matching is fair and Pareto dominates $MDA^m(R, r)$.*

*Proof.* See Appendix 8.8. □

Also, at each problem $(R, r)$, $MDA^m(R, r)$ either Pareto dominates or is equal to $DA^m(R, r)$. In fact, at each round of the $MDA^m$ algorithm, no student is made worse off.

**Proposition 5.** *For each problem, and for each round $r \geq 1$ of the $MDA^m$ algorithm, the matching obtained at the end of round $t$ matches each student with a school that is at least as*

[12] The definition here is not self-consistent in the sense that it refers to another fairness requirement, *fair with reserve*. In Section 8.11, we propose a self-consistent notion and show that all the results we provide are still valid.

[13] For the case with no affirmative action, in the literature, there are versions of the fairness requirement that are very similar to *fairness with conditional reserve*. Two examples are $\tau$-fairness in Alcalde and Romero-Medina (2014) and reasonable fairness in Kesten (2010) (The notion of reasonable fairness appears in a working paper version of Kesten (2010), but not in the published version).

*desirable for her as the school she is matched at the end of round $t-1$.*

*Proof.* See Appendix 8.9. □

**Theorem 4.** *At each problem $(R, r)$, $MDA^m(R, r)$ either Pareto dominates or is equal to $DA^m(R, r)$.*

*Proof.* Follows from Proposition 5. □

Hafalir et al. (2013) provides simulation results revealing that on average when $DA^m$ is used, reserve-type affirmative action brings significant welfare gains for the minority students. Theorem 4 shows that all those welfare gains, and even more, are present with $MDA^m$.

# 6 Strategic properties

A rule $\varphi$ **is strategy-proof** if no student is ever made better off by misreporting her preferences, i.e. there are no problem $(R, r)$, student $s$, and preferences $R'_s$ such that $\varphi_s(R'_s, R_{-s}, r)\ P_s\ \varphi_s(R, r)$.

$DA^m$ is *strategy-proof* (Hafalir et al. (2013)). However, Example 2 shows that when $MDA^m$ is used, a majority student can be better off by misreporting her preferences.

**Example 2.** *Let $S^m \equiv \{m, m'\}$, $S^M \equiv \{M\}$, $C \equiv \{c_1, c_2, c_3\}$, $q \equiv (1, 1, 1)$, and $(r_{c_1}, r_{c_2}, r_{c_3}) \equiv (1, 0, 0)$. Let the preference profile $R$ and the priority profile $\succeq$ be as depicted below. Also, a preference relation for $M$, namely $R'_M$, which is different from $R_M$, is depicted.*

| $R_m$ | $R_{m'}$ | $R_M$ | $R'_M$ | $\succ_{c_1}$ | $\succ_{c_2}$ | $\succ_{c_3}$ |
|---|---|---|---|---|---|---|
| $\underline{c_1}$ | $\underline{c_3}$ | $\boxed{c_1}$ | $c_1$ | $M$ | $M$ | $M$ |
| $\boxed{c_2}$ | $c_1$ | $\underline{c_2}$ | $c_3$ | $m'$ | $m'$ | $m'$ |
| $c_3$ | $c_2$ | $c_3$ | $c_2$ | $m$ | $m$ | $m$ |

*The underlined matching represents $MDA^m(R, r)$. The boxed matching represents $MDA^m($ $R'_M, R_{-M}, r)$. Note that, when the true preferences of $M$ is $R_M$, if she misreports her preferences as $R'_M$, she is assigned to $c_1$ instead of $c_2$, which makes her better off.*

Example 3 shows that, also a minority student can be better off by misreporting her preferences.

**Example 3.** *Let $S^m \equiv \{m, m', m''\}$, $S^M \equiv \{M\}$, $C \equiv \{c_1, c_2, c_3, c_4\}$, $q \equiv (1, 1, 1, 1)$, and $(r_{c_1}, r_{c_2}, r_{c_3}, r_{c_4}) \equiv (1, 0, 0, 0)$. Let the preference profile $R$ and the priority profile $\succeq$ be as depicted below. Also, a preference relation for $m'$, namely $R'_{m'}$, which is different from $R_{m'}$, is depicted.*

| $R_m$ | $R_{m'}$ | $R'_{m'}$ | $R_{m''}$ | $R_M$ | | $\succ_{c_1}$ | $\succ_{c_2}$ | $\succ_{c_3}$ | $\succ_{c_4}$ |
|---|---|---|---|---|---|---|---|---|---|
| $\underline{c_1}$ | $\boxed{c_2}$ | $c_2$ | $\underline{\underline{\boxed{c_4}}}$ | $\boxed{c_1}$ | | $M$ | $M$ | $m'$ | $m'$ |
| $\boxed{c_3}$ | $\underline{c_3}$ | $c_4$ | $c_1$ | $\underline{c_2}$ | | $m''$ | $m''$ | $M$ | $M$ |
| $c_2$ | $c_4$ | $c_3$ | $c_2$ | $c_3$ | | $m$ | $m'$ | $m$ | $m$ |
| $c_4$ | $c_1$ | $c_1$ | $c_3$ | $c_4$ | | $m'$ | $m$ | $m''$ | $m''$ |

*The underlined matching represents $MDA^m(R, r)$. The boxed matching represents $MDA^m($ $R'_{m'}, R_{-m'}, r)$. Note that, when the true preferences of $m'$ is $R_{m'}$, if she misreports her preferences as $R'_{m'}$, she is assigned to $c_2$ instead of $c_3$, which makes her better off.*

Each of the two examples above shows that $MDA^m$ is not strategy-proof. Although this is bad news for the $MDA^m$, Theorem 5 shows that it is not possible to satisfy *strategy-proofness* if we insist on *fairness with conditional reserve* and *minimal responsiveness.*

**Theorem 5.** *No rule is fair with conditional reserve, minimally responsive, and strategy-proof.*

*Proof.* Follows from Theorem 6.

□

In fact, this impossibility result is not due to our particular fairness notion. We will show that a "minimal fairness under affirmative action" is enough to have an impossibility.

Let $(R, r)$ be a problem. A matching $\mu$ is **fair on the no–affirmative-action side** if the following conditions are satisfied.

1. No students' priority is violated at any school by any student of the same type,
2. No students' priority is violated at any school $c$ such that $r_c = 0$.

Note that *fairness on the no–affirmative-action side* requires that on the side of the problem where the affirmative action constraints are not effective, the fairness requirement is the same as the usual fairness requirement for the no–affirmative-action case. One way to incorporate a minimal degree of affirmative action is to require the following: there should be no pair of a minority student and a majority student such that the minority student prefers the school the majority student is assigned, the reserve at that school is not exhausted yet, and exchanging their seats results in an assignment that is *fair on the no–affirmative-action side.* To formalize it, given a pair of students $s, s' \in S$, let $\boldsymbol{\mu_{s \leftrightarrow s'}}$ denote the matching obtained by only switching the seats of $s$ and $s'$ at $\mu$. A matching $\mu$ is minimally fair under affirmative action, or simply **minimally fair**, if the following conditions are satisfied.

1. $\mu$ is fair on the no–affirmative-action side,
2. there are no $m \in S^m$, $M \in S^M$, and $c, c' \in C$ such that
   (a) $\mu(m) = c$ and $\mu(M) = c'$,

(b) $c' \; P_m \; c$, $|\mu^m(c')| < r_{c'}$, $r_c = 0$, and $M \succ_c m$,

(c) $\mu_{m \leftrightarrow M}$ is fair on the no–affirmative-action side.

**Theorem 6.** *No rule is minimally fair, minimally responsive, and strategy-proof.*

*Proof.* See Appendix 8.10. □

## 7 Conclusion

This paper deals with a common problem with current affirmative action policies: a stronger affirmative action may hurt some minority students without benefiting any minority student. First, we showed that the problem is pervasive: it disappears only when the minority students mostly have priority over the majority students. Then, we proposed a new affirmative action rule which never hurts a minority student without benefiting another minority student.

Given that now we have a rule which is *minimally responsive*, a natural question is whether there are rules that are "more responsive" to affirmative action. An ideal affirmative action rule would make no minority student worse off when we move to a stronger affirmative action. In fact, the affirmative action parameter can be thought of as resources made available to the minority students, and an increase in the parameter can be interpreted as an increase in these resources. A natural requirement is that when the resources available to a group increases, no member of the group should be hurt. A requirement similar to this requirement, namely resource monotonicity, has been studied for problems where finitely many objects are to be allocated among agents.[14],[15] However, we know that in such models it is not possible to satisfy resource monotonicity along with some other requirements (See for example Ehlers and Klaus (2003) and Thomson (2003)). Looking for rules that are more responsive to affirmative action, and that make, in case of a stronger affirmative action, a "sufficient" proportion of the minority students better off, is an interesting research direction.

[14] In some models, agents are allowed to consume at most one object, in others they can consume more than one. Considering each seat at each school as an object, a school choice problem fits in the model of allocating objects when each agent can receive at most one. Yet, in most of these models, objects are not assigned priority orderings.

[15] Resource monotonicity in these models is not exactly the counterpart of the requirement we consider here, since the reserves for the minority students are not the only resources available for the minority students, and the reserves for minority students are not resources only for the minority students, except for the quota-type affirmative action.

# 8 Appendix

## 8.1 Proof of Lemma 1

*Proof.* Let $\mu$ be *fair with quota* at $(R, r)$. We will construct the matching $\mu'$ in the following way. Each school keeps the minority students that it was assigned at $\mu$. To allocate the remaining seats (if any), we choose a *fair* matching at the problem where the minority students are absent, and each school's capacity is set equal to the minimum of the number of remaining seats and its majority quota at $(R, r')$. Formally, for each $c \in C$, let $q'_c = \min\{q_c - |\mu^m(c)|, q_c - r'_c\}$. Let $R^M = (R_s)_{s \in S^M}$. Let $\succeq^M$ be the restriction of $\succeq$ to the majority students. Consider the problem $(S^M, C, R^M, \succeq^M, q')$. Let $\mu''$ be a matching that is *fair* at this problem.[16] Let $\mu'$ be defined as follows: each school is assigned the minority students that are assigned to it at $\mu$ and the majority students that are assigned to it at $\mu''$, i.e. for each $c \in C$, $\mu'(c) = \mu^m(c) \cup \mu''(c)$.

We claim that $\mu'$ is *fair with quota* at $(R, r')$. Since $\mu$ is *fair with quota* at $(R, r)$ and $\mu''$ is *fair* at $(S^M, C, R^M, \succeq, q')$, for each $s \in S$, then $\mu'(s)\ R_s\ s$. Suppose that there are $s \in S$ and $c \in C$ such that $c\ P_s\ \mu(s)$. Suppose that $s \in S^m$. Since $\mu$ is *fair with quota* at $(R, r)$, for each $s' \in \mu(c)$, $s' \succ_c s$. So, $U_c^{\succeq}(s) \geq q_c + 1$. Since $(S^m, S^M, C, \succ, q)$ *gives full priority to the minority*, for each $s' \in S^M$, $s \succ_c s'$. Thus, $\mu'(c) = \mu(c) = \mu^m(c)$, and for each $s' \in \mu'(c)$, $s' \succ_c s$.

Suppose that $s \in S^M$. Since $\mu''$ is *fair* at $(S^M, C, R^M, \succeq, q')$, $\mu'(c) = q_c$, and for each $s' \in \mu'^M(c)$, $s' \succ_c s$. Suppose that there is $s' \in \mu'^m(c)$ such that $s \succ_c s'$. Now, since $\mu'(c) = q_c$ and for each $s' \in \mu'^M(c)$, $s' \succ_c s$, there is $s'' \in S^m$ such that $s \succ_c s''$ and $U_c^{\succeq}(s'') \geq q_c + 1$, which contradicts the assumption that $(S^m, S^M, C, \succ, q)$ *gives full priority to the minority*. Hence, $\mu'$ is *fair with quota* at $(R, r')$.

□

## 8.2 Proof of Theorem 1

*Proof.* **Only if part.** Suppose that $(S^m, S^M, C, \succ, q)$ *does not give priority to the minority*. Then there are a pair of students $m \in S^m, M \in S^M$, and a school $c \in C$ such that $M \succ m$ and $|U_c^{\succeq}(m)| \geq q_c + 1$. Let $c' \in C, c' \neq c$. Let $S' \subseteq U_c^{\succeq}(m)$ be such that $m, M \in S'$ and $|S'| = q_c + 1$. Let $R \in \mathcal{R}^S$ be as depicted below: $c'\ P_M\ c\ P_M\ M$, and for each $c'' \in C \setminus \{c, c'\}$, $M\ P_M\ c''$; for each $s \in S' \setminus \{M\}$, $c\ P_s\ s$, and for each $c' \in C \setminus \{c\}$, $s\ P_s\ c'$; for each $s \in S \setminus S'$ and for each $c \in C$, $s\ P_s\ c$.

[16] There is a *fair* matching at this problem. One such matching can be found by the deferred acceptance algorithm.

| $R_M$ | $(R_s)_{s\in S'\setminus\{M\}}$ | $(R_s)_{s\in S\setminus S'}$ |
|---|---|---|
| $c'$ | $c$ | $s$ |
| $c$ | $s$ | $\vdots$ |
| $M$ | $\vdots$ | |
| $\vdots$ | | |

Let for each $c \in C$, $r_c = 0$. Let $r'_{c'} = q_{c'}$, and for each $c \neq c'$, $r'_c = r_c$. Let $\mu$ be the matching defined as follows: $\mu(c) = S' \setminus \{M\}$, $\mu(c') = M$, and for each $c'' \in C \setminus \{c, c'\}$, $\mu(c'') = \emptyset$. Let $\mu'$ be the matching defined as follows: $\mu(c) = S' \setminus \{m\}$, and for each $c'' \neq c$, $\mu(c'') = \emptyset$. Let $\varphi$ be a rule that is *fair with quota*. Since $\varphi$ is *fair with quota*, $\varphi(R, r) = \mu$ and $\varphi(R, r') = \mu'$. Moreover, $\mu$ Pareto dominates $\mu'$ for the minority. Since $r' \geq r$, $\varphi$ is *perversely responsive*.

**If part.** Suppose that $(S^m, S^M, C, \succ, q)$ *gives full priority to the minority*. Let $\mu$ be *fair with quota* at $R$. By Lemma 1, for each $r \leq q$, there is a matching, say $\mu(r)$, such that $\mu(r)$ is *fair with quota* at $(R, r)$ and for each $s \in S^m$, $\mu(s) = \mu(r)(s)$. Let $\varphi$ be the rule defined as follows: for each $(R, r)$, $\varphi(R, r) = \mu(r)$. Note that $\varphi$ is *fair with quota*. Moreover, given two problems $(R, r), (R, r')$, and for each $s \in S^m$, $\varphi_s(R, r) = \varphi_s(R, r')$. Hence, $\varphi$ is *minimally responsive*.

□

### 8.3 Proof of Proposition 1

*Proof.* Let $(S^m, S^M, C, q, R, \succeq, r)$ be the problem defined as follows. Let $S^m \equiv \{m_1, m_2\}$, $S^M \equiv \{M\}$, $C \equiv \{c_1, c_2\}$. Let $q \equiv (1, 1)$, $r \equiv (0, 0)$, and $r' \equiv (1, 0)$. Let $\succeq$ and $R$ be as depicted below.

| $\succ_{c_1}$ | $\succ_{c_2}$ | $R_{m_1}$ | $R_{m_2}$ | $R_M$ |
|---|---|---|---|---|
| $M$ | $M$ | $c_1$ | $\underline{c_2}$ | $\underline{c_1}$ |
| $m_2$ | $m_2$ | $\underline{\boxed{m_1}}$ | $\boxed{c_1}$ | $\boxed{c_2}$ |
| $m_1$ | $m_1$ | $c_2$ | $m_2$ | $M$ |

Only one matching is *fair with reserve* at $(R, r)$: the underlined matching. Let us call it $\mu$. Only one matching is *fair with reserve* at $(R, r')$: the boxed matching. Let us call it $\mu'$. Now, observe that $\mu$ Pareto dominates $\mu'$ for the minority at $R$, although $r' \geq r$.

□

### 8.4 Proof of Proposition 3

*Proof.* Let $\mu \equiv DA^m(R, r)$ and $\mu' \equiv DA^m(R, r')$. We argue by contradiction. Suppose that there is $M \in S^M$ such that $\mu'(M)\ P_M\ \mu(M)$. Let $c \equiv \mu(M)$, $c_1 \equiv \mu'(M)$.

**Step 1.** Since $c_1 \; P_M \; c$ and $\mu(M) = c$, there is a step of the $DA^m(R, r)$ algorithm, say $k_1$, at which $M$ is rejected by $c_1$. Note that at this step the capacity of $c_1$ is exhausted. Since $M \in \mu'(c_1)$, there is a student $s \in S$ who is temporarily accepted by $c_1$ at Step $k_1$ of the $DA^m(R, r)$ algorithm and $s \notin \mu'(c_1)$. We claim that there is such a student who is a majority student. To see this, suppose that $s \in S^m$. Note that since $\mu$ Pareto dominates $\mu'$ for the minority at $R$, $\mu(s) \; R_s \; \mu'(s)$. Also, since $s$ is temporarily accepted by $c_1$ at a step of the $DA^m(R, r)$ algorithm, then $c_1 \; R_s \; \mu(s)$. These statements together imply that $c_1 \; P_s \; \mu'(s)$. Since $\mu'$ is *fair with reserve* at $(R, r')$, $M \succ_{c_1} s$. Moreover, since $s$ is temporarily accepted by $c_1$ at Step $k$ of the $DA^m(P_1)$ algorithm while $M$ is rejected at the same step, the number of minority students temporarily accepted by $c_1$ at that step is at most $r_{c_1}$. Also, since $\mu'$ is *fair with reserve* at $(R, r')$ and $c_1 \; P_s \; \mu'(s)$, then $|\mu'^m(c_1)| = r'$. These statements, together with $r' \geq r$, imply that there is a majority student, say $M_1$, who is temporarily accepted by $c_1$ at Step $k_1$ of the $DA^m(R, r)$ algorithm and $M_1 \notin \mu'(c_1)$.

**Step 2.** Since $M_1 \succ_{c_1} M$ and $\mu'$ is *fair with reserve* at $(R, r')$, then $\mu'(M_1) \; P_{M_1} \; \mu(M_1) = c_1$. Let $c_2 \equiv \mu'(M_1)$. By the same arguments as in Step 1, there is $M_2 \in S^M$ such that there is a step of the $DA^m(R, r)$ algorithm, say $k_2$, at which $M_1$ is rejected by $c_2$, $M_2$ is temporarily accepted by $c_2$ at the same step, and $M_2 \notin \mu'(c_2)$. Now, since $c_2 \; P_{M_1} \; c_1$, then $k_2 < k_1$. Continuing in this fashion, eventually we have $c_t \in C$ and $M_{t-1}, M_t \in S^M$ such that at Step 1 of the $DA^m(R, r)$ algorithm, $M_{t-1}$ is rejected by $c_t$, $M_t$ is temporarily accepted by $c_t$, and $M_t \notin \mu'(c_t)$, $M_{t-1} \in \mu'(c_t)$. Note that $c_t$ is the top-choice of $M_t$, and $M_t \succ_{c_t} M_{t-1}$. Yet $M_t \notin \mu'(c_t)$, $M_{t-1} \in \mu'(c_t)$, contradicting the assumption that $\mu'$ is *fair with reserve* at $(R, r')$. □

## 8.5 Proof of Proposition 4

*Proof.* First suppose that $k = q_3$. Let $S_1, S_2, S'_2, S_3, S'_3, \ldots, S_{K-1}, S'_{K-1}, S_K$ be sets of students such that $|S_1| = q_3$, $S_K = q_K$, and for each $t \in 2, \ldots, K-1$, $|S_t| = q_{t+1}$, $|S'_t| = q_t - q_{t+1}$. Let $S^m \equiv S'_3 \cup S'_4 \cup \cdots \cup S'_{K-1} \cup S_K$ and $S^M \equiv S_1 \cup S_2 \cup \cdots \cup S_{K-1}$.

Let $\succeq$ be a priority profile as partially depicted below. Only the depicted part of the priority profile is relevant for the proof. Each student in $S_1$ has higher priority at $c_1$ than each student who is not in $S_1$; each student in $S'_2$ has higher priority at $c_2$ than each student who is not in $S'_2$; each student in $S_2$ has higher priority at $c_2$ than each student who is not in $S_2 \cup S'_2$, and so on.

| $\succ_{c_1}$ | $\succ_{c_2}$ | $\succ_{c_3}$ | $\succ_{c_4}$ | $\succ_{c_5}$ | $\cdots$ | $\succ_{c_K}$ |
|---|---|---|---|---|---|---|
| $S_1$ | $S'_2$ | $S_2$ | $S_3$ | $S_4$ | $\cdots$ | $S_{K-1}$ |
| $\vdots$ | $S'_3$ | $S'_3$ | $S_4$ | $S_5$ | $\cdots$ | $S_K$ |
| | $\vdots$ | $S_3$ | $S'_3$ | $S'_4$ | $\cdots$ | $S'_{K-1}$ |
| | $S'_{K-1}$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $S_K$ | | | | | |
| | $S_2$ | | | | | |
| | $S_1$ | | | | | |
| | $\vdots$ | | | | | |

Let $R$ be a preference profile as partially depicted below. Only the depicted part of the preference profile is relevant for the proof. Note that for each set in $S_1, S_2, S'_2, S_3, S'_3, \ldots, S_{K-1}, S'_{K-1}, S_K$, students in that set have the same preferences.

| $S_1$ | $S_2$ | $S'_2$ | $S_3$ | $S'_3$ | $\cdots$ | $S_{K-1}$ | $S'_{K-1}$ | $S_K$ |
|---|---|---|---|---|---|---|---|---|
| $c_2$ | $\underline{c_2}$ | $\underline{\underline{\boxed{c_2}}}$ | $\underline{c_3}$ | $\underline{c_3}$ | $\cdots$ | $\underline{c_{K-1}}$ | $\underline{c_{K-1}}$ | $\underline{c_K}$ |
| $c_3$ | $c_4$ | $c_4$ | $c_5$ | $c_5$ | $\cdots$ | $c_3$ | $c_3$ | $c_3$ |
| | $c_5$ | $c_5$ | $c_6$ | $c_6$ | $\cdots$ | $c_4$ | $c_4$ | $c_4$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | | | $c_K$ | $c_K$ | | $c_{K-2}$ | $c_{K-2}$ | $c_{K-2}$ |
| | $c_K$ | $c_K$ | $c_2$ | $c_4$ | | $c_2$ | $c_K$ | $c_{K-1}$ |
| $c_K$ | $\boxed{c_3}$ | $c_3$ | $\boxed{c_4}$ | $\boxed{c_2}$ | $\vdots$ | $\boxed{c_K}$ | $\boxed{c_2}$ | $\boxed{c_2}$ |
| $\underline{\underline{\boxed{c_1}}}$ | $c_1$ | $c_1$ | $c_1$ | $c_1$ | | $c_1$ | $c_1$ | $c_1$ |

Let $r \equiv (0, 0, \ldots, 0)$ and $r' \equiv (0, q_3, 0, 0, \ldots, 0)$. In the first step of the $DA^m(R, r)$ algorithm, "each student in $S_1$" (from now on, we say $S_k$ instead of "each student in $S_k$") is rejected by $c_2$, and all the other students are tentatively accepted by their most-preferred schools. In the following steps, $S_1$ is rejected by all the other schools but $c_1$, until eventually accepted by $c_1$. Thus, $DA^m(R, r)$ is the underlined matching.

In the first step of the $DA^m(R, r')$ algorithm, $S_2$ is rejected by $c_2$, and all the other students are tentatively accepted by their most-preferred schools. Note that due to the affirmative action parameter, $S_2$ is rejected by $c_2$ in favor of the students in $S_1$ who have lower priority at $c_2$. In the following steps, $S_2$ is rejected by schools $c_4, c_5, \ldots, c_K$, and then applies to $c_3$. At this step, $S_2$ is tentatively accepted by $c_3$. At the same step, $S_3 \cup S'_3$ is tentatively rejected by $c_3$. In the following steps, $S_3 \cup S'_3$ is rejected by schools $c_5, c_6, \ldots, c_K$, and $c_4$, and then $S_3$ is rejected by $c_2$, $S'_3$ is rejected by $c_4$. Then, $S_3$ applies to $c_4$, is tentatively accepted, and $S_4 \cup S'_4$ is rejected by $c_4$. At the same step, $S'_3$ applies to $c_2$, is tentatively accepted, and the $|S'_3|$ lowest-priority students in $S_1$ are rejected by $c_2$. These rejected students, after being rejected by all the other schools but $c_1$, are eventually accepted by $c_1$.

Continuing similarly, we reach a step at which $S_{K-1}$ is tentatively accepted, and $S_K$ is rejected, by $c_K$. Then, after being rejected by schools $c_3, c_4, \ldots, c_{K-1}$, eventually $S_K$ is tentatively accepted by $c_2$ and at the same step the $|S_K|$ lowest-priority students in $S_1$ are rejected by $c_2$. At this point no student in $S_1$ is tentatively accepted by $c_2$ because $|S'_3| + |S'_4| + \cdots + |S'_{K-1}| + |S_K| = q_3 = |S_1|$. These rejected students, after being rejected by all the other school but $c_1$, are eventually accepted by $c_1$. Thus, $DA^m(R, r)$ is the matching in boxes.

To cover the case of an arbitrary $k$ such that $q_3 \le k \le q_3 + q_4 + \cdots + q_K$, all we need to do is to switch $k - q_3$ of the majority students in $S_3 \cup S_4 \cup \cdots \cup S_{K-1}$ to minority students. Everything else is as above.

□

## 8.6 Proof of Theorem 2

*Proof.* **If part.** Suppose that $(S^m, S^M, C, \succ, q)$ has a cycle, say $(m, m', M, s_1, \ldots, s_{k-2}, c_1, \ldots, c_k, N_1, \ldots, N_k)$. Let $R \in \mathcal{R}^S$ be defined as follows:

$c_k \; R_m \; c_1 \; R_m \; m$, and for each $c \in C \setminus \{c_k, c_1\}$, $m \; R_m \; c$;
$c_1 \; R_{m'} \; m'$, and for each $c \in C \setminus \{c_1\}$, $m' \; R_{m'} \; c$;
$c_1 \; R_M \; c_2 \; R_M \; M$, and for each $c \in C \setminus \{c_1, c_2\}$, $M \; R_M \; c$;
for each $t \in \{1, \ldots, k-2\}$, $c_{t+1} \; R_{s_t} \; c_{t+2} \; R_{s_t} \; s_t$, and for each $c \in C \setminus \{c_{t+1}, c_{t+2}\}$, $s_t \; R_{s_t} \; c$.

| $R_m$ | $R_{m'}$ | $R_M$ | $R_{s_1}$ | $\cdots$ | $R_{s_t}$ | $\cdots$ | $R_{s_{k-2}}$ |
|---|---|---|---|---|---|---|---|
| $c_k$ | $c_1$ | $c_1$ | $c_2$ | $\cdots$ | $c_{t+1}$ | $\cdots$ | $c_{k-1}$ |
| $c_1$ | $m'$ | $c_2$ | $c_3$ | $\cdots$ | $c_{t+2}$ | $\cdots$ | $c_k$ |
| $m$ | $\vdots$ | $M$ | $s_1$ | $\cdots$ | $s_t$ | $\cdots$ | $s_{k-2}$ |
| $\vdots$ |  | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ | $\vdots$ |

Let $r$ be defined as follows: for each $t \in \{1, \ldots, k\}$, $q_{c_t} \equiv |N_t \cap S^m|$. Let $r'$ be defined as follows: for each $c \in C \setminus c_1$, $r'_c \equiv r_c$ and $r'_{c_1} \equiv r_{c_1} + 1$. Note that $N_{c_1} = q_{c_1} - 1$, so $r'_{c_1} \le q_{c_1}$. Now, there is only one matching that is *fair with reserve* at $(R, r)$, say $\mu$, defined as follows: $\mu(c_1) = N_1 \cup \{M\}$; for each $t \in \{2, \ldots, k-1\}$, $\mu(c_t) = N_t \cup \{s_{t-1}\}$; $\mu(c_k) = N_k \cup \{m\}$. Also, there is only one matching that is *fair with reserve* at $(R, r')$, say $\mu'$, defined as follows: $\mu'(c_1) = N_1 \cup \{m\}$; $\mu'(c_2) = N_2 \cup \{M\}$; for each $t \in \{3, \ldots, k\}$, $\mu(c_t) = N_t \cup \{s_{t-2}\}$. Note that $\mu(m) \; P_m \; \mu'(m)$ and for each $s \in S^m$, $\mu(s) \; R_s \; \mu'(s)$. Thus $\mu$ Pareto dominates $\mu'$ at $R$ for the minority.

**Only if part.** Suppose that $(S^m, S^M, C, \succ, q)$ is acyclic. We will show that $DA^m$, which is *fair with reserve*, is *minimally responsive*. Suppose that $DA^m$ is *perversely responsive*. Then there are $P_1 = (R, r)$ and $P_2 = (R, r')$ such that $r' \geq r$ and $DA^m(P_1) = \mu$ Pareto dominates $DA^m(P_2) = \mu'$ for the minority at $R$. By Proposition 3, $DA^m(P_1) = \mu$ Pareto dominates $DA^m(P_2) = \mu'$ at $R$.

**Step 1.** For each $c \in C$, if $\mu(c) \neq \mu'(c)$, then $|\mu(c)| = |\mu'(c)| = q_c$. To see this, first note that since $\mu'$ is *fair with reserve* at $(R, r')$ and $\mu$ Pareto dominates $\mu'$ at $R$, if there is $s \in \mu(c) \setminus \mu'(c)$, then $|\mu'(c)| = q_c$. Thus for each $c \in C$, if $\mu(c) \setminus \mu'(c) \neq \emptyset$ then $|\mu'(c) \setminus \mu(c)| \geq |\mu(c) \setminus \mu'(c)|$. Since obviously, for each $c \in C$ such that $\mu(c) \setminus \mu'(c) = \emptyset$, $|\mu'(c) \setminus \mu(c)| \geq |\mu(c) \setminus \mu'(c)|$, we have that for each $c \in C$, $|\mu'(c) \setminus \mu(c)| \geq |\mu(c) \setminus \mu'(c)|$. Now, suppose that there is $c \in C$ such that $|\mu'(c) \setminus \mu(c)| > |\mu(c) \setminus \mu'(c)|$. Then there is $s \in S$ such that $\mu(s) = s, \mu'(s) \neq s$, and $\mu(s) \; P_s \; \mu(s) = s$, which contradicts the assumption that $\mu'$ is *fair with reserve* at $(R, r')$. Hence, for each $c \in C$, $|\mu'(c) \setminus \mu(c)| = |\mu(c) \setminus \mu'(c)|$. Thus, for each $c \in C$, if $\mu(c) \neq \mu'(c)$ then $|\mu(c)| = |\mu'(c)| = q_c$.

**Step 2.** There is a list $(m, m', M, s_1, \ldots, s_{k-2}, c_1, \ldots, c_k)$ consisting of a pair of minority students $m, m' \in S^m$, a majority student $M \in S^M$, a list of students $s_1, s_2, \ldots, s_{k-1} \in S$ and a list of schools $c_1, c_2, \ldots, c_k \in C$ such that

1) $M \succ_c m'$, $m \succ_c m'$, and

2) $\mu(M) = c$, $\mu'(M) = c_1$, for each $t \in \{1, \ldots, k-1\}$, $\mu(s_t) = c_t$, $\mu'(s_t) = c_{t+1}$, and $\mu(m) = c_k$, $\mu'(m) = c$.

To see this, since $\mu \neq \mu'$, there are a step of $DA^m(P_1)$ algorithm, say Step $k$, $m' \in S^m$, and $c_1 \in C$ such that up to Step $k$, the student applications and acceptance-rejection decisions of the schools are the same in $DA^m(P_1)$ algorithm and $DA^m(P_2)$ algorithm, but in Step $k$ of $DA^m(P_2)$ algorithm the minority student $m'$, who was rejected by $c_1$ at Step $k$ of $DA^m(P_1)$ algorithm, is temporarily accepted by $c_1$. Since $m'$ is rejected by $c_1$ in $DA^m(P_1)$ algorithm, $c_1 \; P_{m'} \; \mu(m')$. Then $\mu$ being *fair with reserve* at $(R, r)$ implies that $|\mu(c_1)| = q_{c_1}$ and $|\mu^m(c_1)| = q_{c_1}$. Moreover, given that the set of students applying to $c_1$ at Step $k$ of $DA^m(P_1)$ algorithm and $DA^m(P_2)$ algorithm are the same, $m'$ being accepted at Step $k$ of $DA^m(P_2)$ algorithm while being rejected at the same step of $DA^m(P_1)$ algorithm implies that the number of minority students temporarily accepted by $c_1$ at Step $k$ of $DA^m(P_2)$ algorithm is greater than $r_{c_1}$. Hence, $|\mu'^m(c_1)| > r_{c_1}$, and in particular $|\mu'^m(c_1)| > |\mu^m(c_1)|$ and $|\mu'^M(c_1)| < |\mu^M(c_1)|$. Now, note that $m' \notin \mu'(c_1)$ because otherwise, we would have $\mu'(m) \; P_m \; \mu(m)$, contradicting the assumption that $\mu$ Pareto dominates $\mu'$ at $R$. So, let $m \in S^m$ be such that $m \notin \mu(c_1)$, $m \in \mu'(c_1)$. Let $M \in S^M$ be such that $m \in \mu(c_1)$, $m \notin \mu'(c_1)$. Note that $m \succ_{c_1} m'$ and $M \succ_{c_1} m'$. Since for each $c \in C$ such that $\mu(c) \neq \mu'(c)$, $|\mu(c)| = |\mu'(c)| = q_c$, there is a sequence of students $(s_1, s_2, \ldots, s_{k-2})$ and a sequence of schools $(c_1, c_2, \ldots, c_k)$ such that $\mu'(M) = c_2$, $\mu(m) = c_k$, and for each $t \in \{1, \ldots, k-1\}$, $\mu(s_t) = c_{t+1}$,

$\mu'(s_t) = c_{t+2}$. So we have the desired list.

**Step 3.** There is a list of subsets of students $(N_1, \ldots, N_k)$ such that $(m, m', M, s_1, \ldots, s_{k-2}, c_1, \ldots, c_k, N_1, \ldots, N_k)$ forms a cycle. To see this, since $c_1 \; P_M \; \mu'(M)$ and $c_1 \; P_{m'} \; \mu'(m')$, stability of $\mu'$ at $(R, r')$ implies that $|\mu'(c_1)| = q_{c_1}$ and for each $s \in \mu'(c_1) \cap S^M$, $s \succ_{c_1} M$. Note that $m \in \mu'(c_1)$. Let $N_1 \equiv |\mu'(c_1) \setminus \{m\}|$. Note that $(m', N_1) \in T(M, c_1)$.

Since $c_2 \; P_{s_1} \; \mu'(s_1)$, stability of $\mu'$ at $(R, r')$ implies that $|\mu'(c_2)| = q_{c_2}$. Moreover, if $s_1 \in S^m$, then for each $s \in \mu'(c_2)$, $s \succ_{c_2} s_1$. Also if $s_1 \in S^M$, then for each $s \in \mu'(c_2) \cap S^M$, $s \succ_{c_2} s_1$, $M \succ_c s_2$. Let $N_2 \equiv |\mu'(c_2) \setminus \{M\}|$. Note that $(M, N_2) \in T(s_1, c_2)$.

Let $t \in \{3, \ldots, k-1\}$. Since $c_t \; P_{s_{t-1}} \; \mu'(s_{t-1})$, stability of $\mu'$ at $(R, r')$ implies that $|\mu'(c_t)| = q_{c_t}$. In case $s_{t-1} \in S^m$, for each $s \in \mu'(c_t)$, $s \succ_{c_t} s_{t-1}$. In case $s_{t-1} \in S^M$, if $s_t \in S^m$ then for each $s \in \mu'(c_t) \cap S^M$, $s \succ_{c_t} s_{t-1}$, and if $s_{t-2} \in S^M$, then $s_{t-2} \succ_{c_t} s_{t-1}$. Let $N_t \equiv |\mu'(c_t) \setminus \{s_{t-2}\}|$. Note that $(s_{t-2}, N_t) \in T(s_{t-1}, c_t)$.

Since $c_k \; P_m \; \mu'(m)$, stability of $\mu'$ at $(R, r')$ implies that $|\mu'(c_k)| = q_{c_k}$ and for each $s \in \mu'(c_k)$, $s \succ_{c_k} m$. Moreover $s_{k-2} \succ_{c_k} m$. Let $N_k \equiv |\mu'(c_k) \setminus \{s_{k-2}\}|$. Note that $(s_{k-2}, N_k) \in T(m, c_k)$. Thus $(N_1, \ldots, N_k)$ is the desired list of subsets of students.

□

## 8.7 Proof of Proposition 2

*Proof.* Suppose that $(S^m, S^M, C, \succ, q)$ is acyclic. If $M$ has lower priority than each minority student at each school, then the statement obviously holds. So suppose that there are $c \in C$ and $m \in S^m$ such that $M$ has higher priority than $m$ at $c$. If there is no other school at which $M$ has a higher priority than a minority student different from $m$, again the statement holds. So, suppose also that there are $c' \in C$ and $m' \in S^m$ such that $c \neq c'$, $m \neq m'$, and $M$ has higher priority than $m'$ at $c'$. Let $m_1$ and $m_2$ denote the minority students with the lowest and second-lowest priorities at $c$, respectively. Similarly, let $m'_1$ and $m'_2$ denote the minority students with the lowest and second-lowest priorities at $c'$, respectively.

Suppose that $m_1 \neq m'_1$. Then $m'_1 \succ_c m_1$. Moreover, since $|S^m| > q_c + q_{c'}$, there are two disjoint sets of minority students $N_1 \subseteq U_c^{\succ}(m_1) \setminus \{m_1, m'_1, M\}$ and $N_2 \subseteq U_{c'}^{\succ}(m'_1) \setminus \{m'_1, M\}$ such that $N_1 = q_c - 1$ and $N_2 = q_{c'} - 1$. Then, $(m'_1, m_1, M, c, c', N_1, N_2)$ constitutes a cycle, contradicting the assumption that $(S^m, S^M, C, \succ, q)$ is acyclic.

Suppose that $m_1 = m'_1$. Then, either $M \succ_c m_2$ or $M \succ_{c'} m'_2$. Without loss of generality, suppose that $M \succ_c m_2$. Now, by the same arguments as in the above paragraph by letting $m_2$ play the role of $m'_1$, we construct a cycle, contradicting the assumption that $(S^m, S^M, C, \succ, q)$ is acyclic.

□

## 8.8 Proof of Theorem 3

*Proof.* We first prove $3-b$ and $3-c$, and then prove $3-a$.

$\boldsymbol{3-b, 3-c:}$ We will prove a stronger result, which will in turn imply these two results. Consider the following auxiliary fairness notion. A matching is **weakly fair** if the following conditions are satisfied.

1. NLet $s, s' \in S$ and $c \in C$. Then, if the priority of $s$ is violated by $s'$ at $c$, then $s \in S^M$, $s' \in S^m$, and the priority of $s$ is violated by $s'$ at $c$ at unexceeded minority reserve.
2. No student prefers her outside option to her assignment.
3. If a student prefers a school $c$ to her assignment, then the capacity of $c$ is exhausted, i.e. $|\mu(c)| = q_c$.

Note that this new fairness notion is obtained by removing Part $a$ of the "conditional priority requirement" in the definition of *fairness w.r.t. weak-reserve.* Observe that, if a matching is *fair with conditional reserve*, then it is also *weakly fair.* Also, if a matching is *fair*, then it is also *weakly fair.* Hence, proving that, at each problem $(R, r)$, no matching is *weakly fair* and *Pareto dominates* $MDA^m(R,r)$ *for the minority* and also no matching is *weakly fair* and *Pareto dominates* $MDA^m(R,r)$, implies $3-b$ and $3-c$.

*Step 1.* Let $(R, r)$ be a problem. We first prove that no matching is *weakly fair* and *Pareto dominates* $MDA^m(R,r)$ *for the minority.* We argue by contradiction. Suppose that there is a *weakly fair* matching $\mu$ which Pareto dominates $MDA^m(R,r)$ for the minority at $R$. Then, there is a minority student $m$ who is rejected at a step of the last round of $MDA^m(R,r)$ algorithm by the school she is assigned to at $\mu$. Let $T$ denote the last round of $MDA^m(R,r)$ algorithm. Let $k$ be the first step of round $T$ at which a student (not necessarily a minority student), say $s$, is rejected by $\mu(m)$. Note that for each student $s'$ who is tentatively accepted by $c$ at Step $k$ of round $T$, $c\ R_{s'}\ \mu(s')$, since otherwise $s'$ would have applied to (and be rejected by) $\mu(s')$ in round $T$ before Step $k$, contradicting the assumption that $k$ is the first step at which a student is rejected by the school she is assigned to at $\mu$.

First, suppose that $s$ is treated as a minority at $c$ in round $T$, i.e. $s \in f^m_T(c)$. Note that the minority reserve at $c$ is exhausted at step $k$. Let $m \in f^m_T(c)$ be a minority student who is tentatively assigned to $c$ at Step $k$. Note that $m \succ_c s$. Since $c\ R_m\ \mu(m)$ and $\mu$ is *weakly fair*, $m \in \mu(c)$. Thus all the students in $f^m_T(c)$ who are tentatively assigned to $c$ at Step $k$ are assigned to $c$ at $\mu$. Also $s$, who is rejected at Step $k$, is assigned to $c$ at $\mu$. Hence $|\mu^m(c)| > r_c$. Moreover, there is a student $s'$ who is treated as a majority in round $T$ and tentatively accepted by $c$ at Step $k$ but $s' \notin \mu(c)$. Moreover, $s' \succ_c s$, and $s'$ is not assigned to a school better than $c$ at $\mu$, contradicting the assumption that $\mu$ is *weakly fair.*

Now, suppose that $s$ is treated as a majority student at $c$ in round $T$, i.e. $s \notin f^m_T(c)$, and also suppose that the minority reserve capacity is exhausted at Step $k$. Then, each student who is treated as a minority student in round $T$ and who is tentatively accepted at Step $k$ by $c$ has a higher priority than $s$ at $c$, and moreover is assigned to $c$ at $\mu$ due to $\mu$ being *weakly fair*. But then, there is a student $s'$, who is treated as a majority student by $c$ at round $T$, is tentatively accepted at Step $k$ by $c$, and has a higher priority than $s$ at $c$ but is not assigned to $c$ at $\mu$. Note that if the minority reserve of $c$ is exhausted also at $\mu$, and if $s$ is a minority student, then $\mu^m(c) > r_c$. But since $\mu$ is *weakly fair*, $s'$ must be assigned to a more preferred school at $\mu$ to which she must have applied before Step $k$.

Lastly, suppose that $s$ is treated as a majority at $c$ in round $T$, i.e. $s \notin f_T^m(c)$, and also suppose that the minority reserve capacity is not exhausted at Step $k$. Each student who is treated as a minority in round $T$ is tentatively accepted by $c$ at Step $k$, and has higher priority than $s$ at $c$, is assigned to $c$ at $\mu$ by the same arguments as in the previous case. So suppose that there is a student $s'$ who is treated as a minority in round $T$ such that $s \succ_c s'$, and $s'$ is tentatively accepted by $c$ at Step $k$ but is not assigned to $c$ at $\mu$. Observe that $s$ is a majority student and $s'$ is a minority student. Remember that $s'$ prefers $c$ to her assignment at $\mu$. But since $\mu$ Pareto dominates $MDA^m(R, r)$ for the minority, $s'$ must be rejected by $c$ in the last round after Step $k$. But then, $s'$ is an interfering student at $c$, contradicting the assumption that this round is the last round. Hence, each student who is treated as a minority in round $T$ and tentatively accepted at Step $k$ by $c$ is assigned to $c$ at $\mu$. But then, there is a student, say $s'$, who is treated as a majority student in round $T$ is tentatively accepted at Step $k$ by $c$, and has a higher priority than $s$ at $c$, but is not assigned to $c$ at $\mu$. But since $\mu$ is *weakly fair*, $s'$ must be assigned to a better school at $\mu$ to which she should have applied before Step $k$.

*Step 2.* The following modification of the proof at Step 1 proves that no matching is *weakly fair* and *Pareto dominates* $MDA^m(R, r)$. Similarly, suppose that there is a *weakly fair* matching $\mu$ which Pareto dominates $MDA^m(R, r)$ at $R$. To obtain a contradiction in the proof of at Step 1, there are two parts where we use the fact that "$\mu$ Pareto dominates $MDA^m(R, r)$ for the minority at $R$." The first one is that in the first paragraph we claim that there is a student who is rejected at a step of the last round of $MDA^m(R, r)$ algorithm by the school she is assigned to at $\mu$. This claim holds also when $\mu$ Pareto dominates $MDA^m(R, r)$ at $R$. The second one is that in the last paragraph, a minority student is claimed not to be worse off at $\mu$ compared to $MDA^m(R, r)$ at $R$. This claim also holds when $\mu$ Pareto dominates $MDA^m(R, r)$ at $R$. With these modifications, the same arguments yield to a contradiction, which proves the assertion.

$\boldsymbol{3 - a:}$

Let $(R, r)$ be a problem. By $3 - b$, no matching is *fair with conditional reserve* and Pareto dominates $MDA^m(R, r)$ for the minority. Thus, no matching is *fair with reserve* and Pareto dominates $MDA^m(R, r)$ for the minority. Given this, it directly follows that $MDA^m$ is *fair with conditional reserve.*

To prove that $MDA^m$ is *minimally responsive*, we argue by contradiction. Suppose that there are two problems $(R, r)$ and $(R, r')$ such that $r' \geq r$ and $MDA^m(R, r)$ Pareto dominates $MDA^m(R, r')$ for the minority at $R$. Let $\mu \equiv MDA^m(R, r)$ and $\mu' \equiv MDA^m(R, r')$. By Part 2, $\mu$ is not *fair with conditional reserve* at $(R, r')$. Note that $\mu$ is *fair with conditional reserve* at $(R, r)$ since $MDA^m$ is *fair with conditional reserve*. So, for each $s \in S$, $\mu(s)\ R_s\ s$. Observe that at $\mu$, if the priority of a minority student is violated at $(R, r)$, it is also violated at $(R, r')$.

Suppose that the priority of a majority student $M$ is violated at a school $c$ at $(R, r')$. If $|\mu^m(c)| \leq r'_c$ and there is a majority student $M' \in \mu(c)$ such that $M \succ_c M'$, then the priority of $M$ is violated at $c$ also at $(R, r)$. Suppose that $|\mu^m(c)| > r'_c$ and there is $s' \in \mu(c)$ such that $s \succ_c s'$. Since $|\mu^m(c)| \geq r'_c \geq r_c$, again $\mu$ is not *weakly fair* at $(R, r)$ contradicting the assumption that $MDA^m$ is *fair with conditional reserve*.

□

## 8.9 Proof of Proposition 5

*Proof.* Let $(R, r)$ be a problem. Let $\mu$ and $\mu'$ be the outcomes of the $MDA^m(R, r)$ algorithm at round $t - 1$ and $t$, respectively. We argue by contradiction. Suppose that there is $s \in S$ such that $\mu(s)\ P_s\ \mu'(s)$. Let $c_1 \equiv \mu(s)$ and $c \equiv \mu'(s)$. Observe that there is a step of $DA^m(f^m_t, R, r)$ algorithm, say Step $k_1$, at which $s$ is rejected by $c_1$. Also note that at this step the capacity of $c_1$ must be exhausted.

**Step 1.** We claim that there is a student $s_1 \in S$ such that $s_1$ is temporarily accepted by $c_1$ at Step $k_1$ of $DA^m(f^m_t, R, r)$ algorithm, $s_1 \notin \mu(c_1)$, and $\mu(s_1)\ P_{s_1}\ c_1$.

*Case 1.* Suppose that $s$ is treated as a minority at round $t - 1$ by $c_1$, i.e. $s \in f^m_{t-1}(c_1)$. First note that, since $s \in \mu(c_1)$, $(s, c_1)$ is not an interfering pair of round $t - 1$ of $MDA^m(R, r)$ algorithm. So, $s \in f^m_t(c_1)$. Moreover, at Step $k_1$ of $DA^m(f^m_t, R, r)$ algorithm, at least $r$ minority students are tentatively accepted by $c_1$. So, for each student $s''$ which is tentatively accepted by $c_1$ at that step, $s'' \succ_{c_1} s$. If $|\mu^m(c)| \leq q^m_{c_1}$, then there is a minority student $s' \in S^m$ such that such that $s'$ is temporarily accepted by $c_1$ at Step $k_1$ of $DA^m(f^m_t, R, r)$ algorithm and $s' \notin \mu(c_1)$. Then, also $\mu(s')\ P_{s'}\ c_1$. If $|\mu^m(c)| > q^m_{c_1}$, then there is a student $s' \in S$ such that such that $s'$ is temporarily accepted by $c_1$ at Step $k_1$ of $DA^m(f^m_t, R, r)$ algorithm and $s' \notin \mu(c_1)$. Then, also $\mu(s')\ P_{s'}\ c_1$.

*Case 2.* Suppose that $s$ is treated as a majority at round $t - 1$ by $c_1$, i.e. $s \notin f^m_{t-1}(c_1)$. Then $s \notin f^m_t(c_1)$. If the number of minority students tentatively accepted by $c_1$ at Step $k_1$ of $DA^m(f^m_t, R, r)$ algorithm is greater than $|\mu^m(c_1)|$, then there is a minority student $s' \in S^m$ such that $s'$ is temporarily accepted by $c_1$ at Step $k_1$ of $DA^m(f^m_t, R, r)$ algorithm and $s' \notin \mu(c_1)$. Then, also $\mu(s')\ P_{s'}\ c_1$. If the number of minority students tentatively accepted by $c_1$ at Step $k_1$ of $DA^m(f^m_t, R, r)$ algorithm is not greater than $|\mu^m(c_1)|$, then there is a majority student

$s' \in S^M$ such that $s'$ is temporarily accepted by $c_1$ at Step $k_1$ of $DA^m(f_t^m, R, r)$ algorithm and $s' \notin \mu(c_1)$. Note that $s' \succ_{c_1} s$. Then, also $\mu(s')\ P_{s'}\ c_1$.

**Step 2.** Let $c_2 \equiv \mu(s_1)$. Since $\mu(s_1)\ P_{s_1}\ \mu'(s_1)$, by the same arguments as in Step 1, there is $s_2 \in S$ such that there is a step of $DA^m(f_t^m, R, r)$ algorithm, say $k_2$, at which $s_1$ is rejected by $c_2$, $s_2$ is temporarily accepted by $c_2$ at the same step, $s_2 \notin \mu(c_2)$, and $\mu(s_2)\ P_{s_2}\ c_2$. Now, since $c_2\ P_{s_1}\ c_1$, then $k_2 < k_1$. The fact that we can continue in this fashion contradicts the first step of $DA^m(f_t^m, R, r)$ algorithm being Step 1. □

### 8.10 Proof of Theorem 6

*Proof.* Let $\varphi$ be a rule which is *minimally fair* and *minimally responsive*. Let $S^m \equiv \{m, m'\}$, $S^M \equiv \{M\}$, $C \equiv \{c_1, c_2, c_3\}$, and $q \equiv (1,1,1,1)$. Let $(r_{c_1}, r_{c_2}, r_{c_3}) \equiv (0,0,0)$ and $r' = (1,0,0)$. Let the preference profile $R$ and the priority profile $\succeq$ be as depicted below. The preference profile is depicted twice to indicate the matchings that are *fair with conditional reserve* at $(R, r)$ and $(R, r')$, respectively.

| $R_m$ | $R_{m'}$ | $R_M$ | $R_m$ | $R_{m'}$ | $R_M$ | $\succ_{c_1}$ | $\succ_{c_2}$ | $\succ_{c_3}$ |
|---|---|---|---|---|---|---|---|---|
| $c_1$ | $\underline{c_3}$ | $\underline{c_1}$ | $\boxed{c_1}$ | $\boxed{c_3}$ | $c_1$ | $M$ | $M$ | $M$ |
| $\underline{c_2}$ | $c_1$ | $c_2$ | $c_2$ | $c_1$ | $\boxed{c_2}$ | $m'$ | $m'$ | $m'$ |
| $c_3$ | $c_2$ | $c_3$ | $c_3$ | $c_2$ | $c_3$ | $m$ | $m$ | $m$ |

Only one matching is *minimally fair* at $(R, r)$: the underlined matching in the leftmost profile. Let us call it $\mu_1$. Then, $\varphi(R, r) = \mu_1$. Only one matching is *minimally fair* at $(R, r')$: the boxed matching in the middle profile. Let us call it $\mu_2$. Then, $\varphi(R, r') = \mu_2$.

Now, let the preference relation $R'_M$ be as depicted below. Let $R' = (R_m, R_{m'}, R'_M)$. Below, the preference profile $R'$ is depicted twice to indicate the matchings that are *fair with conditional reserve* at $(R', r)$ and $(R', r')$, respectively.

| $R_m$ | $R_{m'}$ | $R'_M$ | $R_m$ | $R_{m'}$ | $R'_M$ | $\succ_{c_1}$ | $\succ_{c_2}$ | $\succ_{c_3}$ |
|---|---|---|---|---|---|---|---|---|
| $c_1$ | $\underline{c_3}$ | $\underline{c_1}$ | $c_1$ | $\boxed{c_3}$ | $\boxed{c_1}$ | $M$ | $M$ | $M$ |
| $\underline{c_2}$ | $c_1$ | $c_3$ | $\underline{\boxed{c_2}}$ | $\underline{c_1}$ | $\underline{c_3}$ | $m'$ | $m'$ | $m'$ |
| $c_3$ | $c_2$ | $c_2$ | $c_3$ | $c_2$ | $c_2$ | $m$ | $m$ | $m$ |

Only one matching is *minimally fair* at $(R', r')$: the underlined matching on the leftmost profile. Let us call it $\mu'_1$. Then, $\varphi(R', r) = \mu'_1$. Exactly two matchings are *fair with conditional reserve* at $(R', r')$: the underlined and boxed matchings on the middle profile. Since $\varphi$ is *minimally responsive*, it can not choose the underlined matching at $(R', r')$. Let us call the boxed matching $\mu_2'$. Then, $\varphi(R', r') = \mu_2'$. But then, at problem $(R, r')$, the majority student $M$ is better off if she reports $R_M'$ instead of $R_M$. Thus, $\varphi$ is not *strategy-proof.*

□

### 8.11 A Digression on fairness with conditional reserve

The only difference between *fairness w.r.t. the reserve* and *fairness w.r.t. the conditional reserve* is that with the latter, a priority violation at unexhausted reserve may be justified on the following grounds: the matching is not Pareto dominated for the minority by another matching that is *fair with reserve.* However, a matching that is *fair with conditional reserve* and exhibits a priority violation at unexhausted reserve may be Pareto dominated by a matching that is *fair with conditional reserve.* In that sense, *fairness w.r.t. the conditional reserve* is not a self-consistent fairness notion. A self-consistent fairness notion is the following:

Let $M' \subseteq M$ be the set of matchings such that for each $\mu \in M'$, the followings hold.

1. If there are $s, s' \in S$ and $c \in C$ such that the priority of $s$ is violated by $s'$ at $c$, then $s \in S^M$, $s' \in S^m$, and the minority reserve at $c$ is unexceeded at $\mu$, i.e. $|\mu^m(c)| \leq r_c$.
2. No student prefers her outside option to her assignment.
3. If a student prefers a school $c$ to her assignment, then the capacity of $c$ is exhausted, i.e. $|\mu(c)| = q_c$.

Let $(R, r)$ be a problem. A set of assignments $F \subseteq M'$ is a **consistently fair set** if the following conditions are met.

1. Internal consistency: for each $\mu \in F$ such that there is a priority violation at unexhausted reserve, there is no $\mu' \in F$ which Pareto dominates $\mu$ for the minority.
2. External consistency: for each $\mu \in M' \setminus F$, there is a priority violation at unexhausted reserve and there is $\mu' \in F$ which Pareto dominates $\mu$ for the minority.

**Proposition 6.** *At each problem, there is a unique consistently fair set.*

*Proof.* Suppose that there are two consistently fair sets, say $F_1$ and $F_2$, such that $F_1 \neq F_2$. Without loss of generality, suppose that there is a matching, say $\mu$, such that $\mu \in F_1 \setminus F_2$. Since $\mu \notin F_2$, there is a priority violation at unexhausted reserve at $\mu$. Since $\mu \in F_1$ and $F_1$ is internally consistent, there is no matching in $F_1$ which Pareto dominates $\mu$ for the minority. Since $\mu \notin F_2$ and $F_2$ is externally consistent, there is a matching in $F_2$ which Pareto dominates $\mu$ for the minority. Then, there is $\mu' \in F_2 \setminus F_1$ that Pareto dominates $\mu$ for the minority. But then, by symmetric arguments, there is $\mu'' \in F_1 \setminus F_2$ which Pareto dominates $\mu'$ for the minority. Now, $\mu''$ Pareto dominates $\mu$ for the minority, which contradicts $F_1$ being internally consistent. □

Let $(R, r)$ be a problem. Let $F$ be the consistently fair set at this problem. A matching $\mu$ is **consistently fair** if $\mu \in F$.

The notion of a consistently fair set is analogous to the notion of a von-Neumann–Morgenstern stable set for cooperative games (Neumann and Morgenstern (1944)).[17]

**Proposition 7.** *$MDA^m$ is consistently fair.*

*Proof.* Let $(R, r)$ be a problem. By Theorem 3, there is no matching that is fair with conditional reserve and Pareto dominates $MDA^m(R, r)$ for the minority and there is no matching that is fair and Pareto dominates $MDA^m(R, r)$ for the minority. It is straightforward to show that $MDA^m(R, r) \in M'$. Hence, by external consistency of $F$, we have $MDA^m(R, r) \in F$. □

**Proposition 8.** *No rule is consistently fair, minimally responsive, and strategy-proof.*

*Proof.* In the proof for Theorem 5, four problems are considered: $(R, r)$, $(R, r')$, $(R', r)$, and $(R', r')$. One can easily check that at each problem, the set of matchings that are *fair with conditional reserve* is equal to the set of *consistently fair* matchings. Thus, the same arguments in that proof also proves this assertion. □

### 8.12 Two alternative ways to remove interferers

In the $MDA^m$ algorithm, at each round, only the interfering minority students of the last step at which there is an interferer, are treated as majority students. Actually, there are two other natural ways to proceed. we will show that these two other methods do not work.

1. Method 1: At each round, all the interfering minority students are treated as majority students at each school they interfere.
2. Method 2: At each round, only the interfering minority students of the first step at which there is an interferer, are treated as majority students.

We will show that neither of these two methods work: each of the methods is *perversely responsive*. Consider the problem $(S, C, \succeq, q, R, r)$ defined as follows. Let $S^m \equiv \{m_1, m_2, m_3\}$, $S^M \equiv \{M_1, M_2\}$, $C \equiv \{c_1, c_2, c_3, c_4\}$. Let $(q_{c_1}, q_{c_2}, q_{c_3}, q_{c_4}) = (1, 1, 1, 2)$, $r = (0, 0, 1, 0)$, and $r' = (1, 0, 1, 0)$. Let$\succeq$ and $R$ be as depicted as follows. Only the top parts of the profiles that are relevant to the example is depicted.

| $\succ_{c_1}$ | $\succ_{c_2}$ | $\succ_{c_3}$ | $\succ_{c_4}$ |
|---|---|---|---|
| $M_1$ | $M_1$ | $M_2$ | $M_2$ |
| $m_3$ | $m_2$ | $m_2$ | $m_1$ |
| $m_1$ |  | $m_3$ | $m_4$ |

| $R_{m_1}$ | $R_{m_2}$ | $R_{m_3}$ | $R_{M_1}$ | $R_{M_2}$ |
|---|---|---|---|---|
| $c_1$ | $\underline{c_2}$ | $\underline{c_3}$ | $\underline{c_1}$ | $c_3$ |
| $\underline{\underline{\boxed{c_4}}}$ | $\boxed{c_3}$ | $\boxed{c_1}$ | $\boxed{c_2}$ | $\underline{\underline{\boxed{c_4}}}$ |
|  |  | $c_4$ |  |  |

[17] Ehlers and Klaus (2007) study von-Neumann–Morgenstern stable sets in matching problems.

$DA^m(R,r)$ is the underlined matching and $DA^m(R,r')$ is the boxed matching. Note that $DA^m(R,r)$ Pareto dominates $DA^m(R,r')$ for the minority, although $r' \geq r$. This is in line with the fact that $DA^m$ is *perversely responsive.*

$DA^m(R,r)$ algorithm is depicted below.

| Step | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|
| 1 | $m_1, \boxed{M_1}$ | $m_2$ | $\boxed{m_3}, M_2$ | |
| 2 | | | | $\boxed{m_1}, \boxed{M_2}$ |

Note that there is no interfering pair at $(R,r)$. $DA^m(R,r')$ algorithm is depicted below.

| Step | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|
| 1 | $\boxed{m_1}, M_1$ | $m_2$ | $\boxed{m_3}, M_2$ | |
| 2 | | $\boxed{M_1}, m_2$ | | $\boxed{M_2}$ |
| 3 | | | $\boxed{m_2}, m_3$ | |
| 4 | $\boxed{m_3}, m_1$ | | | |
| 5 | | | | $\boxed{m_1}, \boxed{M_2}$ |

There are two interfering pairs at $(R,r')$: $(m_3,c_3)$ is an interfering pair of Step 3 and $(m_1,c_1)$ is an interfering pair of Step 4.

Let us use Method 1 to modify $DA^m$ algorithm. Then, for each $(m,c) \in IP(R,r')$, the interfering minority $m$ is treated as a majority student at school $c$.

| Step | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|
| 1 | $m_1, \boxed{M_1}$ | $m_2$ | $m_3, \boxed{M_2}$ | |
| 2 | $m_3, \boxed{M_1}$ | | | $\boxed{m_1}$ |
| 3 | | | | $\boxed{m_1}, \boxed{m_3}$ |

Now, there is no new interfering pair. However, the outcome at $(R,r)$ Pareto dominates the outcome at $(R,r')$ for the minority. Thus Method 1 is not *minimally responsive.*

Instead, let us use Method 2. Remember that the first step of $(R,r')$ at which there is an interfering pair is Step 3 and the only interfering pair of that step is $(m_3,c_3)$.

| Step | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|
| 1 | $\boxed{m_1}$,$M_1$ | $m_2$ | $m_3$, $\boxed{M_2}$ | |
| 2 | $\boxed{m_3}$,$m_1$ | $\boxed{M_1}$,$m_2$ | | |
| 3 | | | $\boxed{m_2}$,$M_2$ | $\boxed{m_1}$ |
| 4 | | | | $\boxed{m_1}$,$\boxed{M_2}$ |

Now, there is no new interfering pair. However, the outcome at $(R, r)$ Pareto dominates the outcome at $(R, r')$ for the minority. Thus Method 2 is also not *minimally responsive.*

So, the order we treat the interfering minority students is crucial for the result. Note that in the above example, when we move to a stronger affirmative action, a minority student, namely $m_1$, initiates a chain of rejections, and these rejections cause another minority student, namely $m_3$, to become an interferer, which eventually results with $m_1$ becoming an interferer. In other words, we have nested rejection cycles. It turns out that, to achive *minimal responsiveness*, first the outer rejection cycle initiated by the last interferer should be dissolved.

# References

Abdulkadiroğlu, A. (2013). School choice. *Oxford Handbook of Market Design (Z. Neeman, M. Niederle, A. E. Roth, and N. Vulkan, eds), Oxford University Press forthcoming.*

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

Alcalde, J. and A. Romero-Medina (2014). Strategy-proof fair school placement. *Working paper*.

Echenique, F. and B. Yenmez (2014). How to control controlled school choice. *American Economic Review*, forthcoming.

Ehlers, L. and A. Erdil (2010). Efficient assignment respecting priorities. *Journal of Economic Theory 145*, 1269–1282.

Ehlers, L., I. Hafalir, and M. Yildirim (2014). School choice with controlled choice constraints: Hard bounds versus soft bounds. *Journal of Economic Theory 153*, 648–683.

Ehlers, L. and B. Klaus (2003). Coalitional strategy-proof and resource-monotonic solutions for multiple assignment problems. *Social Choice and Welfare 21*, 265–280.

Ehlers, L. and B. Klaus (2007). Von neumann-morgenstern stable sets in matching problems. *Journal of Economic Theory 134*, 537–547.

Erdil, A. and T. Kumano (2012). Prioritizing diversity in school choice. *Working paper*.

Ergin, H. (2002). Efficient resource allocation on the basis of priorities. *Econometrica 70*, 2489–2497.

Hafalir, I., B. Yenmez, and M. Yildirim (2013). Effective affirmative action in school choice. *Theoretical Economics 8*, 325–363.

Kagel, J. H. and A. E. Roth (2000). The dynamics of reorganization in matching markets: A laboratory experiment motivated by a natural experiment. *Quarterly Journal of Economics 115*, 201–235.

Kesten, O. (2010). School choice with consent. *Quarterly Journal of Economics 125*, 1297–1348.

Klijn, F., J. Pais, and M. Vorsatz (2014). Affirmative action through minority reserves: An experimental study on school choice. *Working paper.*

Kojima, F. (2012). School choice: Impossibilities for affirmative action. *Games and Economic Behavior 75*, 685–693.

Kominers, S. D. and T. Sönmez (2013). Designing for diversity in matching. *Working paper.*

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

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

Roth, A. (2008). Deferred acceptance algorithms: History, theory, practice, and open questions. *International Journal of Game Theory 36*, 537–569.

Thomson, W. (2003). On monotonicity in economies with indivisible goods. *Social Choice Welfare 21*, 195–205.

Ünver, U. (2000). On the survival of some unstable two-sided matching mechanisms. *International Journal of Game Theory 33*, 239–254.

Westkamp, A. (2013). An analysis of the german university admissions system. *Economic Theory 53*, 561–589.