



# Exploring the Impact of Unilateral Affirmative Action on the School Choice Problem

## Citation

Zhang, Katherine Feifei. 2023. Exploring the Impact of Unilateral Affirmative Action on the School Choice Problem. Bachelor's thesis, Harvard College.

## Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37376414>

## Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

## Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#)

# Exploring the Impact of Unilateral Affirmative Action on the School Choice Problem

A THESIS PRESENTED

BY

KATHERINE F. ZHANG

TO

THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

BACHELOR OF THE ARTS

IN THE SUBJECT OF

COMPUTER SCIENCE

HARVARD UNIVERSITY

CAMBRIDGE, MASSACHUSETTS

MAY 2023

©2023 – KATHERINE F. ZHANG  
ALL RIGHTS RESERVED.

## Exploring the Impact of Unilateral Affirmative Action on the School Choice Problem

### ABSTRACT

Affirmative action in school choice has become a burgeoning area of interest for policymakers and researchers alike. Numerous efforts have sought to characterize the impact of affirmative action on common matching mechanisms for school admissions, such as student-proposing deferred acceptance. Particularly, much work has focused on priority-based affirmative action, or a change in school rankings that uplifts the minority student group while keeping group orderings constant. It has been theoretically shown that deferred acceptance may not necessarily produce outcomes consistent with the intentions of priority-based affirmative action. However, with many elite magnet and exam schools in the U.S. implementing such affirmative action policies independently from the rest of their school district, this popular policy action begs the question of whether such a unilateral change truly improves the outcomes of minority students. Thus, we examine the less well-studied case in which a singular school unilaterally implements affirmative action by changing their preference ordering over students. In order to determine whether and how often a unilateral affirmative action policy produces better admissions outcomes for the intended students, we investigate the impacts of three types of affirmative action ranking changes of different magnitudes – arbitrary, single-swap, and priority-based affirmative action. Through theoretical and simulative analysis, we find that worst cases for several types of student welfare are achievable even through a single school swapping a single pair of students, showing that student-proposing deferred acceptance is highly volatile even under minute changes. We do find that some guarantees of welfare are possible for priority-based affirmative action – however, our simulations show that on average, only a couple minority students will achieve better matches, and that there is high variance in potential outcomes even for students who undergo the same type of change in the affirmative action policy. As such, we cautiously conclude that unilateral affirmative action is not sufficiently beneficial in the average case and is capable of creating severe negative impacts on all students.

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
1.1	Thesis Contributions . . . . .	4
<b>2</b>	<b>RELATED WORK</b>	<b>7</b>
2.1	Deferred Acceptance in School Choice . . . . .	7
2.2	Affirmative Action in School Choice . . . . .	9
2.3	Responsive Affirmative Action . . . . .	10
<b>3</b>	<b>MODEL</b>	<b>12</b>
3.1	Setting . . . . .	12
3.2	Deferred Acceptance Algorithm . . . . .	15
3.3	Metrics . . . . .	16
<b>4</b>	<b>THEORETICAL RESULTS</b>	<b>20</b>
4.1	Impact of a Unilateral Preference Change . . . . .	21
4.2	Worst-Case Student Welfare for Specific Preference Changes . . . . .	27
4.3	Guarantees for Specific Preference Changes . . . . .	33
4.4	Lack of Guarantee for an Arbitrary Change . . . . .	35
<b>5</b>	<b>SIMULATION RESULTS</b>	<b>37</b>
5.1	Simulation Methods . . . . .	38
5.2	Arbitrary Change . . . . .	40
5.3	Single-Swap Change . . . . .	46
5.4	Priority-Based Affirmative Action . . . . .	51
<b>6</b>	<b>CONCLUSION AND FUTURE WORK</b>	<b>61</b>
	<b>APPENDIX A APPENDIX</b>	<b>69</b>
A.1	Proofs . . . . .	69
A.2	Additional Results for Arbitrary Preference Change . . . . .	75
A.3	Additional Results for Single-Swap Preference Change . . . . .	77

A.4 Additional Results for Priority-Based Affirmative Action . . . . . 77

# Acknowledgments

I am so incredibly fortunate to have had Professor Yiling Chen as my teacher, mentor, and advisor for this thesis. Thank you for helping me to build this piece of work from the ground up – it simply could not have been done without your endless guidance and feedback. And of course, thank you for your time, your energy, your positivity, and for being one of the best professors I’ve ever had. Thanks to your kindness and mentorship, I will always cherish this research experience. I would also like to thank Safwan Hossain for his advice and feedback, and especially for helping me to begin researching at all. Thank you for supporting me throughout the entire thesis-writing process, from formulating an initial model to final draft comments. Thank you to my thesis readers, Professor Ariel Procaccia and Professor Yannai Gonczarowski, for so generously agreeing to read my thesis. Finally, this thesis would not have been possible without the support of my friends and family. Thank you particularly to the WiCS community, my lovely roommates, and Mom, Dad, William, and Sophia.

# 1

## Introduction

Affirmative action in school admissions is a popular yet controversial fairness intervention that intends to rectify gaps in socioeconomic welfare resulting from systemic discrimination by affording more opportunities for minority students to receive better education. Policy discourse surrounding the benefits and drawbacks of affirmative action traditionally revolves around whether it is morally correct to introduce race and socioeconomic status into the consideration of applicants, as well as whether increased minority representation in secondary schools and colleges actually leads to im-

proved downstream outcomes for these students [29] [9] [17] [21]. However, as argued in [30], schools and school districts often introduce affirmative action policies without necessarily considering their immediate consequences and without consensus on the criteria for such policies. Thus, a more fundamental but less-discussed concern that this paper hopes to explore is whether current implementations of affirmative action even accomplish their immediate goal of enrolling more minority students in the schools they prefer.

The specific form of affirmative action that will be explored in this work is priority-based affirmative action, which simply improves the rankings, or priorities, of minority students at schools. For schools that involve admissions by criteria such as exams, grades, or artificial student-scoring systems, this may entail inflating a student's numerical score or, in a more fluid system, accounting for a student's race or socioeconomic status when comparing applicants to each other. Previous empirical work has focused on how school assignments can change when entire school systems adopts such an affirmative action policy. Practical examples of this often arise from more centralized systems for determining school admissions. For instance, China's university admissions system gives minority students bonus points on its National Higher Education Entrance Examination. This policy allows minority students to be at least as well off, if not better, as compared to their original standings. Cities like Chicago have also mandated a consistent affirmative action policy across multiple public school districts, and the city of Boston plans to eventually phase into a similar system, where every school utilizes the same scoring method for ranking students. However, a less well-studied scenario is one in which a single institution in a system implements priority-based affirmative action. Such cases are often reflected in real life, due to the existence of particular elite magnet and exam schools that have developed a reputation for competitive admissions and are often neither racially nor socioeconomically representative of their surrounding areas [6] [26] [10]. Some examples of this include the recent announcement that Boston's three exam schools would stop requiring their exam to counteract the disproportionate effects of the COVID-19 pandemic on disadvantaged students

[8]. Likewise, in 2020, Thomas Jefferson High School in Fairfax County, Virginia removed its exam in a highly controversial decision that came about as a result of the unrepresentative demographics of this particular school rather than the entire district. [24]. In fact, it is the topical – and controversial – nature of these individual changes and pilot programs that motivates analyzing the case wherein a single school changes modifies its methodology for ranking students. We call this individual “pioneering” of priority-based affirmative action a *unilateral affirmative action change*, as such a change is performed without coordination or consistency with other schools in the district.

In order to shed more light on the impacts of unilateral affirmative action, we utilize the widely-used model of school choice. In the school choice problem, models of affirmative action rest on the simplifying assumption that students are either of the “minority” or the “majority,” and that affirmative action is meant to increase the enrollment and/or improve admissions prospects of the minority class relative to the majority. School districts in U.S. cities such as Boston, Chicago, New York, Cambridge, Seattle, etc. utilize a system of school choice in which students and parents indicate a preference ordering over schools, submit grades, and take entrance exams so that schools can determine who their preferred admits are. Although schools may consider a myriad of factors in admitting one student over another, schools’ decisions can essentially be considered as forming a preference ordering over students. As such, the school choice problem has frequently been formalized as a two-sided matching market, with  $n$  students being matched to  $m$  schools. In this model, we incorporate unilateral affirmative action by fixing one school that changes its preference ordering, while others’ orderings remain constant. We measure the impact of unilateral affirmative action by comparing the admissions results for students with and without such a policy.

Many mechanisms have been devised to actually perform the matching: some include zone- and neighborhood-specific constraints or have been adjusted to include factors such as prioritizing those with siblings at the school [5]. One popular method was the Boston mechanism, in which schools give higher priority to students who also ranked them higher. This mechanism was used in counties

such as Cambridge, Tampa-St. Petersburg, Miami-Dade, Denver, and White Plains [3]. However, this mechanism can lead to perverse incentives for students to misreport their preferences [2]. In 2005, the city of Boston replaced the Boston mechanism with the ubiquitous Gale-Shapley student-proposing deferred acceptance algorithm, which has become the standard in school choice thanks to its lack of manipulability [14]. As such, this work studies the impact of unilateral affirmative action on student-proposing deferred acceptance.

## 1.1 THESIS CONTRIBUTIONS

This thesis seeks to investigate the impact of a single school changing its preference ranking in a school choice problem, a case that has not been sufficiently explored, despite the fact that it is reflective of real-world affirmative action policies. In this work, we consider the school choice problem with  $n$  students and  $n$  schools when student-proposing deferred acceptance is used to determine the matching outcome both before and after the affirmative action change. We consider this special case sufficiently representative, as [28] demonstrates that an instance of many-to-one matching is equivalent to a one-to-one matching in which each school is copied a number of times that corresponds to its student body size. Due to the generalizability of this specific case, we perform both theoretical and simulative analysis of this setting. We first examine the circumstances under which unilateral affirmative action affects the matching and relate the conditions necessary to make all students weakly better or worse off in the new match.

Overall, we hope to understand whether a single school's affirmative action change can affect the matching given by student-proposing deferred acceptance, as well as whether any particular type of change is less or more likely to achieve the goals of enacting affirmative action delineated above. Thus, the rest of the paper is structured as follows: Chapter 3 introduces three types of unilateral affirmative action changes: an arbitrary change in ranking, a single-pair swap between two adjacent

students, and a priority-based affirmative action policy, in which minority students experience a Pareto improvement in ranking.

Chapter 3 also introduces several welfare-based metrics for measuring the utility of unilateral affirmative action to students, as well as metrics that measure how well each type of unilateral preference change is able to achieve the overarching goal of applying affirmative action. This analysis fills an oft-ignored gap in the analysis of affirmative action by considering whether the policy being used actually performs its intended purpose. Additionally, although student-proposing deferred acceptance is lauded for incentivizing truthful behavior, it is rarely examined through the lens of sensitivity of students' welfare – however, this is a necessary tool for measuring the impact of an affirmative action change on students.

Through theoretical analysis, we find in Chapter 4 that the worst cases for several types of welfare are achievable for all three types of unilateral affirmative action. Thus, even a change in a single school change can have severe negative repercussions on all students. However, we also see that unilateral priority-based affirmative action can guarantee weakly better performance for at least one minority student. In order to more definitively give a policy recommendation, we perform simulations for each type of unilateral affirmative action change and compare student welfare and outcomes in each. By supplementing our worst-case analysis and guarantees with comprehensive analysis of the “on-average” effect of unilateral preference changes, we contextualize our theoretical results and determine the expected outcome of such a change. We find that unilateral affirmative action leaves most types of welfare constant in expectation regardless of different values of  $n$  or  $k$  and tends not to benefit many minority students, even as their prevalence in the general student body increases. Even when a positive trend can be noticed, we see high variance in the outcomes of a given student. Combining both types of analysis paints a picture of unilateral affirmative action in which average performance is not entirely consistent with the goals of affirmative action, and worst-case performance can cause alarmingly severe consequences to students. Thus, we caution against its indis-

criminate usage to bridge the socioeconomic and racial gaps in schools and instead encourage school districts to reach greater consensus on an application of affirmative action before implementing such policies.

# 2

## Related Work

### 2.1 DEFERRED ACCEPTANCE IN SCHOOL CHOICE

With numerous possible metrics for evaluating the efficacy of a matching algorithm for school choice, this variety begs the question of which qualities of matching algorithms should be considered most desirable or in-line with producing outcomes for both groups of agents involved in the matching process. Previous work establishes some consensus in this regard, with two desirable

characteristics being stability and strategyproofness [27] [13]. These characteristics are particularly important to generating outcomes in the school choice problem, as they ensure that all mutually beneficial matches are made and reduce the ability of students who know more about the school choice system to try and strategically report. Because it possesses these desirable properties, one common algorithm for finding a solution to the school choice problem, is the student-proposing deferred acceptance (DA) algorithm introduced by Gale and Shapley in [14].

[4] first proposed the use of student-proposing DA to find stable matchings for the school choice problem. DA finds a matching between two parties that have preference rankings over each other, with one party “proposing” matches and the other accepting or rejecting until a complete matching is made. This process can easily be analogized to the school application process – students “propose,” or apply, to schools in order of their preference, and schools accept the most desirable proposals. The DA algorithm possesses the desirable qualities of returning a stable matching and being strategyproof for the proposing side [27] [13]. Since no algorithm for a stable matching can be strategyproof for both groups of agents, student-proposing DA possesses the best guarantee of strategyproofness for the purposes of school choice, as schools prefer that students not be able to manipulate their preferences. It is also computationally efficient, with a runtime of  $O(nm)$ , where  $n$  is the number of students and  $m$  is the number of schools. Thus, student-proposing DA has become a widely accepted solution to the school choice problem. [3] argues that the stability and strategyproofness of student-proposing DA make it a more robust mechanism for finding matchings than predecessors like the Boston mechanism, which could incentivize strategic behavior and thus create advantages for students who have more knowledge about the school application process. Since then, Boston – and many other cities – have switched to using student-proposing DA to match students to schools, and DA has become fairly commonplace for school choice in public high school districts [11]. However, as affirmative action policies become more commonplace in an effort to desegregate and diversify school districts, explorations of the impact of affirmative action when applied to DA

show that the desired effects of such policies may not be guaranteed or even achievable [23] [12].

## 2.2 AFFIRMATIVE ACTION IN SCHOOL CHOICE

Affirmative action has been a long-time source of contention in various school admissions systems, with conflicting empirical results about whether such interventions actually produce the desired increase in welfare for minority students, and whether the potential trade-off between equity and efficiency is justified. [25] demonstrates that affirmative action in Brazil's centralized college system increased minority representation by 73%, while estimating that the benefits outweighed the costs by 1.6 times, as displaced students could seek high quality degrees elsewhere. By contrast, [20] shows that affirmative action based on regional caps in the Japanese hospital matching market can lead to avoidable losses of efficiency. [15] studies the impact of hypothetical affirmative action interventions on Boston high school admissions and finds mixed results, noting that while some more extreme interventions could produce desired outcomes for minority students, most would have relatively small benefits for diversity. Although our work does not involve an empirical study, such studies highlight the lack of clarity surrounding the impacts of affirmative action and thus motivate the need to study affirmative action within the school choice problem.

There are three common models of affirmative action in the school choice problem: priority-based affirmative action, which weakly improves the rankings of minority students relative to majority students but keeps the orderings within the two groups the same, quota-based affirmative action, which places a cap on the number of majority students that can be admitted by a school, and reserve-based affirmative action, which saves a certain number of seats for minority students at each school. Assuming that the desired outcome for an affirmative action policy is to better the matches of minority students, [7] produces a favorable result by showing that student-proposing DA is the unique mechanism that “respects improvements” for one student, meaning that a market

in which a single student has been improved will lead to a weakly better matching for that student. However, [23] demonstrates that applying priority-based and quota-based affirmative action to student-proposing DA are not guaranteed to produce a better outcome for any minority students. Particularly, there are markets for which the application of a stronger quota-based or priority-based affirmative action policy can lead every minority student to have a Pareto inferior outcome. [16] uses simulations to show that the aforementioned instances may occur with nontrivial frequency, though they focus on quota-based affirmative action rather than priority-based affirmative action, which is the focus of this work. The authors then propose reserve-based affirmative action as a new alternative, which they argue is more efficient than quota-based affirmative action because it allows the reserved seats for minority students to be filled with majority students when there are not enough minority students. The authors also demonstrate that minority reserves will make at least one minority student weakly better off.

### 2.3 RESPONSIVE AFFIRMATIVE ACTION

The negative results of [23] and [16] have motivated studies that restrict the school choice problem in an effort to discover what type of intervention for school and student preferences in student-proposing DA is necessary to ensure that outcomes for minority students are truly consistent with affirmative action changes. Specifically, “minimal responsiveness” is satisfied when at least one minority student is strictly better off in the case that any minority student is worse off, a metric that we utilize in our simulations [12]. Other works consider modifying DA itself in order to guarantee such responsiveness. [22] does not consider affirmative action, but rather proposes the efficiency-adjusted DA mechanism (EADAM) to remediate existing inefficient circumstances in DA in which many students receive their first choice. EADAM’s algorithm follows the same steps as deferred acceptance but allows a student to “waive” a higher position in a priority order that would leave

their own final match unchanged but negatively impact another student, a phenomenon that occurs frequently in student-proposing DA. [12] extends this work to the affirmative action space by proposing a minimally responsive but non-strategyproof mechanism for affirmative action based on EADAM. The modified deferred acceptance with minority reserves (MDA) algorithm utilizes the reserve-based affirmative action policy described in [16] but modifies it such that a minority student will not be accepted into the reserve of a school it prefers (e.g. will be treated like a majority student) if this acceptance would worsen other minority students but not change its own outcome. This work also demonstrates that no mechanism is minimally fair, minimally responsive, and strategyproof, demonstrating that there are trade-offs between fairness, efficiency, and strategyproofness. [19] then shows that for priority-based affirmative action, giving “full priority” to the minority by ensuring that each minority student has higher priority than any majority student ensures that *each* minority student is weakly better off in student-proposing DA. [18] builds upon these ideas by showing that a similar variation of EADAM as the algorithm given in [12] is minimally responsive to priority-based affirmative action as well. As such, there is a wide range of work that focuses on adapting and modifying student-proposing DA to meet the minimally responsive (or stronger) criteria. However, as [12] points out as a caveat to their work, ideas like giving “full priority” are potentially inapplicable in practice even if they do create positive guarantees, as it is highly unlikely that all minority students could be ranked above all majority students. Practical implementation of EADAM is also fairly complex and even legally challenging due to the need to perform adjustments in the matching to promote efficiency, as described in [22]. As such, while developing responsive mechanisms for affirmative action presents some interesting solutions to the inefficiency of DA, we find that studying the impact of unilateral affirmative action policies on DA is more reflective of the real-world policies being implemented in the present day [6] [10].

# 3

## Model

### 3.1 SETTING

In accordance with the traditional school choice setting described in [4], we let  $S$  and  $C$  be finite and disjoint sets of students and schools. For the purposes of this analysis, we let  $|S| = |C| = n$ , as mentioned in Chapter 1. Let  $S^m$  and  $S^M$  be nonempty sets that denote minority and majority

students, respectively, where  $S^m \cup S^M = S$  and  $S^m \cap S^M = \emptyset$ . Let  $|S^m| = k$ , where  $k \in [1, n - 1]$ .\*

For each student  $s_i \in S$ ,  $\succ_{s_i}$  represents a strict and complete preference ordering over schools. Each school  $c_j \in C$  also possesses a strict and complete preference ordering over students, denoted by  $\succ_{c_j}$ . A *school choice problem* is denoted by a tuple  $G = \langle S, C, \succ_S, \succ_C \rangle$ . This work examines a specific case of priority-based affirmative action in which one school in  $C$  implements a change in preferences, which we call *unilateral* affirmative action. Since all schools are interchangeable in this setting, WLOG we can say that  $c_1$  is the school that implements affirmative action. A school choice problem *with unilateral affirmative action* can then be denoted by a tuple  $\langle G, G' \rangle$  where  $G = \langle S, C, \succ_S, \succ_C \rangle$  and  $G' = \langle S, C, \succ_S, \succ'_C \rangle$ . We let  $\succ'_{c_1}$  represents school  $c_1$ 's preferences after unilateral affirmative action and  $\succ'_C = \succ'_{c_1} \cup \succ_{c_j \in C \setminus c_1}$  denote the complete preference profile for schools after the affirmative action policy has been implemented.

The exact type of rank change made to  $\succ_{c_1}$  to produce  $\succ'_{c_1}$  has been purposely left ambiguous thus far, as we will examine three different types of changes: *arbitrary*, *single-swap*, and *priority-based* preference changes. For all  $s_i \in S$  and  $c_j \in C$ , we denote the rank of student  $s_i$  in  $\succ_{c_j}$  to be  $r_{c_j}(s_i) \in [1, n]$  and the rank of student  $c_j$  in  $\succ_{s_i}$  to be  $r_{s_i}(c_j) \in [1, n]$ . We also let the rank of student  $s_i$  in  $\succ'_{c_1}$  to be  $r'_{c_1}(s_i) \in [1, n]$ .

**Definition 1 (Arbitrary Preference Change)** A completely randomly reshuffling of  $\succ_{c_1}$  that does not consider minority or majority status of students.

**Definition 2 (Single-Swap Preference Change)** Given two adjacent students  $s_i, s_j \in S$  in  $\succ_{c_1}$  where  $r_{c_1}(s_i) = r_{c_1}(s_j) - 1$ , swap them such that  $s_i \succ_{c_1} s_j$  but  $s_j \succ'_{c_1} s_i$ .

---

\*We allow the size of the minority set  $k$  to be over half of the students because “minority” students may not be a numerical minority but can still be underrepresented within certain schools, particularly elite schools. [15] highlights that Boston’s student population is 75% Black and Hispanic students, but these students only make up approximately 40% of the student body in exam schools.

**Definition 3 (Priority-Based Affirmative Action)** For all students  $s_i, s_j \in S$ , if  $s_i \succ_{c_1} s_j$  and  $s_i \in S^m$ , then  $s_i \succ'_{c_1} s_j$ , and if  $s_i, s_j \in S^M$  and  $s_i \succ_{c_1} s_j$ , then  $s_i \succ'_{c_1} s_j$ .

As such, unilateral priority-based affirmative action promotes the priorities of minority students relative to majority students in  $\succ'_{c_1}$  while keeping the ordering within  $S^m$  and  $S^M$  constant. The single-swap change can be treated as a special case of priority-based affirmative action in which  $k = 1$  and  $S^m = \{s_i\}$ .

Since we assume that the set of schools and students are of equal size, we assume each school has one seat that must be filled up by a single student, and a *matching*  $\mu$  constitutes a one-to-one mapping between students and schools such that if  $s_i$  and  $c_j$  are matched, then  $\mu(s_i) = c_j$  and  $\mu(c_j) = s_i$ . We denote the matching produced by preference profiles  $\succ_S$  and  $\succ_C$  as  $\mu$  and the matching produced by preference profiles  $\succ_S$  and  $\succ'_C$  as  $\mu'$ . We say that a student  $s_i \in S$  is *better off* under  $\mu'$  if  $\mu'(s_i) \succ_{s_i} \mu(s_i)$ , *worse off* under  $\mu'$  if  $\mu(s_i) \succ_{s_i} \mu'(s_i)$ , and *unchanged* under  $\mu'$  if  $\mu(s_i) = \mu'(s_i)$ . Additionally,  $\mu'$  *Pareto dominates*  $\mu$  if  $\mu'(s_i) \succeq_{s_i} \mu(s_i)$  for all  $s_i \in S$  and  $\mu'(s_i) \succ_{s_i} \mu(s_i)$  for at least one  $s_i \in S$ . Since affirmative action policies are meant to improve the minority, we also define an efficiency measure for the minority class in accordance with [16] and [19]:  $\mu'$  *Pareto dominates*  $\mu$  *for the minority* if  $\mu'(s_i) \succeq_{s_i} \mu(s_i)$  for all  $s_i \in S^m$  and  $\mu'(s_i) \succ_{s_i} \mu(s_i)$  for at least one  $s_i \in S^m$ .

We then consider the notion of a stable matching for the school choice problem. For a school choice problem  $G = \langle S, C, \succ_S, \succ_C \rangle$ , a matching is stable if it is individually rational and does not have a blocking pair. Individual rationality requires that  $\mu(s_i) \succ_{s_i} \emptyset$  for all  $s_i \in S$  (e.g. all students prefer to have a match). A blocking pair is defined as a student and school pair that mutually prefer each other to their matches in  $\mu$ . Thus,  $\mu$  has no blocking pair if for all  $s_i \in S$  and  $c_j \in C$  such that  $c_j \succ_{s_i} \mu(s_i)$ ,  $\mu(c_j) \succ_{c_j} s_i$  and for all  $s_i, c_j$  where  $s_i \succ_{c_j} \mu(c_j)$ ,  $\mu(s_i) \succ_{c_i} c_j$ .

### 3.2 DEFERRED ACCEPTANCE ALGORITHM

We next discuss DA and its positive qualities as a mechanism for finding stable matchings. A *mechanism* is a mapping  $\varphi$  that associates a matching  $\varphi(G)$  to any school choice problem  $G$ . The student-proposing DA algorithm is one such matching mechanism that is commonly used for finding matchings in the school choice problem. For a school choice problem with each type of unilateral affirmative action as described above, the algorithm is simply run twice, once with the preferences  $G$  and once with the preferences  $G'$ . As such, we describe the student-proposing DA algorithm as it is outlined by Gale and Shapley in [14]:

*Step 1:* Start with no student or school having been matched. Each student then proposes to their first choice school. Each school that receives proposals gives their single seat to the student they prefer most according to their preference ranking.

*Step k:* In general, at step  $k$ , any student who was rejected in the previous step proposes to their next choice according to the student preference rankings. Each school considers the student it currently holds together with any new proposers and gives its seat to the student they prefer most out of this pool. The remaining students are rejected.

The algorithm terminates once all students are no longer rejected, with the resulting assignments forming the final match  $DA(G)$ . As mentioned above, since unilateral affirmative action only involves a preference change on behalf of  $c_1$ , we consider that  $DA$  is run on two separate school choice problems:  $G = \langle S, C, \succ_S, \succ_C \rangle$  for the baseline matching and  $G' = \langle S, C, \succ_S, \succ'_C \rangle$  for the matching with unilateral affirmative action. We will denote the matchings as  $\mu = DA(G)$  and  $\mu' = DA(G')$ .

Student-proposing DA has the property of being strategy-proof for students, meaning that for all  $s_i \in S$  there is no preference ranking  $\succ'_{s_i}$  such that, holding all other preference profiles constant, the outcome of DA with  $\succ'_{s_i}$  dominates that of  $\succ_{s_i}$ . It is also weakly group strategy-proof for students, so that no group of students  $\hat{S} \in S$  can gain by misreporting their preferences as well [27]. This

property, along with the guarantee of stable matches, has made student-proposing DA a popular solution to the school choice problem. However, we hope to measure other aspects of the matchings  $\mu$  and  $\mu'$  given by the application of unilateral affirmative action when using student-proposing DA. In this way, we determine how one school can impact student welfare, as well as whether welfare is sensitive to changes of different magnitudes. In order to do so, we give several metrics and benchmarks by which student-proposing DA with unilateral affirmative action should be evaluated.

### 3.3 METRICS

We define metrics that can be measured both for  $S^m$  and  $S^M$  as well as ones that capture welfare for all  $S$  regardless of class status, as the arbitrary and single-swap preference changes do not explicitly differentiate between minority and majority students. We first adapt the concepts of utilitarian and egalitarian welfare to this environment. The first metric we consider is *utilitarian count*, which is simply defined as the number of students who are worse off after the unilateral affirmative action change, or the number of students  $s_i \in S$  where  $\mu(s_i) \succ_{s_i} \mu'(s_i)$ . Utilitarian count can also be defined with respect to  $S^m$  or  $S^M$  (e.g. the number of minority/majority students who are worse off under  $\mu'$ ). We also define *egalitarian welfare* as follows:

**Definition 4 (Egalitarian Welfare)** The highest-magnitude rank change for a student  $s_i \in S$  where  $\mu(s_i) \succ_{s_i} \mu'(s_i)$ :

$$\min_{s_i \in S} r_{s_i}(\mu(s_i)) - r_{s_i}(\mu'(s_i))$$

Once again, this can be measured separately for the minority and majority classes in the unilateral priority-based affirmative action case by taking the minimum over  $S^m$  or  $S^M$  rather than  $S$ . Thus far, utilitarian count has quantified how many students are worse off, and egalitarian count takes the

Rawlsian approach to welfare by basing it on the worst-off student. *Utilitarian welfare* then utilizes rank changes  $r_{s_i}$  for all students to characterize the magnitude of change in the entire system:

**Definition 5 (Utilitarian Welfare)** For some set of students  $\hat{S} \in \mathcal{S}$ , let the utilitarian welfare of that set be:

$$U(\hat{S}) = \sum_{s_i \in \hat{S}} r_{s_i}(\mu(s_i)) - r_{s_i}(\mu'(s_i))$$

As such, utilitarian welfare increases if  $\mu'(s_i)$  is ranked higher in  $s_i$  than  $\mu(s_i)$ . In priority-based affirmative action, we can go one step further and measure the *utilitarian welfare difference* between the classes, or  $U(S^M) - U(S^m)$ .

All of the aforementioned metrics provide some insight into the general welfare of the entire system. However, as suggested by the “responsiveness” discourse, another essential aspect of evaluating the efficacy of unilateral preference changes is whether such a change has the desired impact. Even for an arbitrary preference change, in which there is no delineation of group membership, we would hope that for a student  $s_i$ , an improvement in  $r_{c_1}(s_i)$  would imply a weak improvement in  $U(s_i)$  as well. To measure this for arbitrary and single-swap preference changes, we introduce *collateral damage* and *consistency* as follows:

**Definition 6 (Collateral Damage)** The number of  $s_i$  where  $r_{c_1}(s_i) = r'_{c_1}(s_i)$  but  $\mu(s_i) \succ_{s_i} \mu'(s_i)$ .

**Definition 7 (Consistency)**  $\forall s_i \in \mathcal{S}$ , let the rank of  $s_i$  in  $c_1$ 's old preferences be  $r_{c_1}(s_i)$ , and let the rank of  $s_i$  in  $c_1$ 's new preferences be preferences be  $r'_{c_1}(s_i)$ . Then,  $s_i$ 's matching change is consistent with its ranking change if  $r_{c_1}(s_i) > r'_{c_1}(s_i)$  and  $\mu'(s_i) \succeq_{s_i} \mu(s_i)$ ,  $r_{c_1}(s_i) = r'_{c_1}(s_i)$  and  $\mu'(s_i) = \mu(s_i)$ , or  $r_{c_1}(s_i) < r'_{c_1}(s_i)$  and  $\mu(s_i) \succeq_{s_i} \mu'(s_i)$ .

Collateral damage is essentially the number of students unchanged in  $\succ'_{c_1}$  who end up with a less-preferred match in  $\mu'$ , while consistency We seek to find the amount of collateral damage and the proportion of students whose welfare outcomes are consistent with their movement in  $\succ'_{c_1}$ .

For priority-based affirmative action, we stratify based on group membership, so we no longer use collateral damage as a metric. In its place, we observe consistency for the minority and majority classes separately and record *minimal responsiveness*, as defined by [12].

**Definition 8 (Minimal Responsiveness)** If there is  $s_i \in S^m$  such that  $\mu(s_i) \succ_{s_i} \mu'(s_i)$ , then there must be some other  $s_j \in S^m$  such that  $\mu'(s_j) \succ_{s_j} \mu(s_j)$  (e.g.  $\mu$  does not Pareto dominate  $\mu'$  for the minority).

Though [23] demonstrates that student-proposing DA is not guaranteed to be minimally responsive to priority-based affirmative action, measuring how often unilateral affirmative action leads to a matching  $\mu'$  that is in line with minimal responsiveness can be informative towards making policy recommendations. Specifically, simulating the rate at which the algorithm is minimally responsive (e.g. when at least one minority student is worse off, how often is another minority student better off?) can be informative towards determining whether the cases in [23] are commonplace or an anomaly.

However, what minimal responsiveness does not measure is the number of minority and majority students who are strictly better and worse off, respectively. To measure this, we now define two types of consistency to be measured for each class – weak and strict consistency. *Weak consistency* is modeled after the general-form consistency in Definition 7.

**Definition 9 (Minority- and Majority-Class Weak Consistency)**  $\forall s_i \in S^m$ , minority student  $s_i$ 's matching change is consistent with its ranking change if  $\mu'(s_i) \succeq_{s_i} \mu(s_i)$ .  $\forall s_j \in S^M$ , majority student  $s_j$ 's matching change is consistent with its ranking change if  $\mu(s_j) \succeq_{s_j} \mu'(s_j)$ .

In general, we hope that in  $G'$  minority students will be made better off, potentially at the expense of majority students being made worse off. *Strict consistency* measures whether minority and majority students are actually changing their matchings (rather than remaining the same) and is similarly defined as follows:

**Definition 10 (Minority- and Majority-Class Strict Consistency)**  $\forall s_i \in S^m$ , minority student  $s_i$ 's matching change is strictly consistent with its ranking change if  $\mu'(s_i) \succ_{s_i} \mu(s_i)$ .  $\forall s_j \in S^M$ , majority student  $s_j$ 's matching change is strictly consistent with its ranking change if  $\mu(s_j) \succ_{s_j} \mu'(s_j)$ .

We compare these metrics across the three different types of preference changes defined in Section 3.1 to both gain some insight into the contrast between the unilateral affirmative action case and the general affirmative action case, while also trying to determine whether there is any correlation between the type of preference change being made (e.g. the relatively “small” single-swap changes as opposed to the more disruptive arbitrary change) and the welfare outcomes for students.

# 4

## Theoretical Results

We begin by characterizing when and how a unilateral affirmative action policy may actually change the matching  $\mu'$ , before turning to more specific analysis of each type of affirmative action change. First, when we fix the matching algorithm,  $\mu' \neq \mu$  as a result of a unilateral preference change if and only if  $\mu(c_1) \neq \mu'(c_1)$  (Theorem 1 and Corollary 1.1). Assuming that this is true, we establish that  $\mu(c_1)$  being better off under  $\mu'$  implies that  $\mu'$  Pareto dominates  $\mu$  and that  $\mu'(c_1)$  being worse off under  $\mu'$  implies that  $\mu$  Pareto dominates  $\mu'$  (Theorem 2 and 3). We also show that although  $c_1$  is the

school performing the preference change, it is possible that  $\mu(c_1) \succ'_{c_1} \mu'(c_1)$ , and that this outcome for  $c_1$  also implies that  $\mu'$  Pareto dominates  $\mu$  (Corollary 2.2). We then show that for all three types of unilateral preference changes, it is possible to achieve a utilitarian count of  $n - 1$  (Section 4.2.1 and 4.2.2), and that it is possible to achieve the worst-case egalitarian welfare of  $-(n - 1)$  for at least one student (Section 4.2.3). Finally, we see that while student-proposing DA may not be minimally responsive to priority-based affirmative action, unilateral priority-based affirmative action does ensure that the highest-ranked, strictly better off minority student  $s_i$  such that  $c_1 \succ_{s_i} \mu(s_i)$  is weakly better off under  $\mu'$  (Theorem 5). We also show that the same does not hold for a unilateral arbitrary preference change.

#### 4.1 IMPACT OF A UNILATERAL PREFERENCE CHANGE

We first explore the impact of a preference change by a single school  $c_1$  without delving into the specific type of change. WLOG let  $c_1$  be matched to  $s_1$  in the original matching,  $\mu$ . We first show that if  $c_1$  changes its preferences, then either there exists a new match  $\mu'$  for  $G'$  where  $\mu'(c_1) \neq s_1$ , or  $\mu$  is stable for  $G'$ .

**Theorem 1** Let  $\mu$  be a stable matching for  $G$ , and WLOG let  $c_1$  be matched to  $s_1$  in  $\mu$ . Then, only one of the following is true:

1.  $\exists \mu'$  such that  $\mu'(c_1) \neq s_1$
2.  $\mu$  is stable for  $G'$ .

*Proof.* We can first show that if  $c_1$  and  $s_1$  are removed from the matching problem, the truncated matching remains stable for  $G'$ . Let  $S' = S \setminus s_1$  and  $C' = C \setminus c_1$ , and let  $\mu_t$  be the truncated matching, or  $(s_2, \mu(s_2)), (s_3, \mu(s_3)), \dots, (s_n, \mu(s_n))$ .

If  $c_1$  and  $s_1$  are removed, students  $s_2$  through  $s_n$  simply remove  $c_1$  from their rankings, and schools  $c_2$  through  $c_n$  remove  $s_1$  from their rankings. For stability to hold, then for any pair  $(s_i, c_j) \forall i \in [2, n]$  and  $j \in [2, n]$   $(\mu(s_i) \succeq_{s_i} c_j) \vee (\mu(c_j) \succeq_{c_j} s_i)$ , where equality holds if  $s_i$  and  $c_j$  are matched with each other in the original match. Let us assume that stability does not hold in the truncated match  $\mu_t$ . Then, for some  $s_i$  and  $c_j$ ,  $(c_j \succ_{s_i} \mu_t(s_i)) \wedge (s_i \succ_{c_j} \mu_t(c_j))$ . However, if this was the case, then  $s_i$  and  $c_j$  would have formed a blocking pair in the original matching instance, and  $\mu$  would have been unstable, which is a contradiction. Thus,  $\mu_t$  is stable for  $S'$  and  $C'$ .

First, let us consider the case where  $\mu$  is not stable for  $G'$  but it is stable for  $G$ . We know from above that  $\mu_t$  is stable for  $S'$  and  $C'$ . Thus, the stability condition must not hold between some student  $s_i \in S'$  and  $c_1$  or some school  $c_j \in C'$  and  $s_1$ :

**Case 1:** For  $i \in [2, n]$ , check stability for  $(s_1, \mu(s_i) = \mu_t(s_i))$ : Assume that the stability condition does not hold for this pair, so  $(s_1 \succ_{\mu(s_i)} s_i) \wedge (\mu(s_i) \succ_{s_1} c_1)$ , then  $(s_1, \mu(s_i))$  would have been a blocking pair in  $\mu$ . Thus, stability must not hold between some student  $s_i \in S'$  and  $c_1$ :

**Case 2:** For  $i \in [2, n]$ , check stability for  $(s_i, c_1)$ : If the stability condition does not hold for this pair in  $\mu'$ , so  $(c_1 \succ_{s_i} \mu(s_i)) \wedge (s_i \succ'_{c_1} s_1)$  under  $c_1$ 's new preference ranking. This does not imply a blocking pair in  $\mu$ , as  $c_1$  changes its rankings between  $G$  and  $G'$ . However, if this is the case, then  $(s_1, c_1)$  cannot be matched in  $\mu'$ , as  $(s_i, c_1)$  form a blocking pair. Thus, when  $\mu$  is not stable for  $G'$ ,  $\exists \mu'$  where  $\mu'(c_1) \neq s_1$ .

We can use similar reasoning to show that if there is no stable matching  $\mu'$  for  $G'$  where  $\mu'(c_1) \neq s_1$ , then  $\mu$  must remain stable in  $G'$ . As shown above,  $\mu_t$  is stable in  $G'$ , and the stability condition holds between  $s_1$  and any  $c_j$  such that  $j \in [2, n]$ . We then reason about the stability condition between  $s_i$  for  $i \in [2, n]$  and  $c_1$ . As shown above, it is possible that the stability condition does not hold, but only if  $s_1$  and  $c_1$  are no longer matched in  $\mu'$  (because  $(s_i, c_1)$  form a blocking pair). However, if there is no stable matching  $\mu'$  for  $G'$  where  $\mu'(c_1) \neq s_1$ , then  $\mu$  must remain stable for  $G$ . This completes our proof.

□

Now, let us consider what this implies when we fix a matching algorithm. Since we hope to model a college admissions-like setting, we fix the algorithm to be student-proposing deferred acceptance. Then, the algorithm produces one single stable matching for each of  $G$  and  $G'$ . Therefore, we can say the following:

**Corollary 1.1** Fixing the matching algorithm, matching  $\mu' \neq \mu$  if and only if  $\mu'(c_1) \neq \mu(c_1)$ .

This corollary follows as an extension of Theorem 1.

#### 4.1.1 WELFARE OF THE CHANGED SCHOOL'S MATCH

For the following sections, we assume that the matching algorithm has been fixed as student-proposing deferred acceptance. Thus, we take Corollary 1.1 to be true: in this setting, the match changes ( $\mu \neq \mu'$ ) if and only if  $c_1$ 's match changes. This means that  $c_1$ 's old and new partners ( $\mu(c_1)$  and  $\mu'(c_1)$ ) must change their matches.

First, let us establish some constraints for how  $\mu'$  can differ from  $\mu$  using the stability condition.

**Lemma 1** Consider students  $s_i \in S$  where  $s_i \neq \mu(c_1), \mu'(c_1)$  and schools  $c_i \in C$  where  $c_i \neq c_1$ . If such a student or school is better off under  $\mu'$ , then their new match must be worse off under  $\mu'$ . If such a student or school is worse off under  $\mu'$ , then their previous match is better off under  $\mu'$ .

The outline of the proof is as follows: for any such student or school, we consider the possible blocking pairs in both  $\mu$  and  $\mu'$ . Since both are stable matchings, in order to prevent a blocking pair in  $\mu'$  when  $s_i$  or  $c_i$  prefers their match in  $\mu$ , the previous match must be better off, and in order to prevent a blocking pair in  $\mu$  when  $s_i$  or  $c_i$  prefers their match in  $\mu'$ , the new match must be worse off. (see Appendix A.1 for full proof).

Knowing this, we can characterize the relationships between the welfare of  $\mu(c_1)$ ,  $\mu'(c_1)$ , and other students in  $\mu$  and  $\mu'$  for any preference change by  $c_1$ . For the following statements, we assume that  $\mu' \neq \mu$  (e.g. the match has changed), and we let the new match of  $\mu(c_1)$  be  $\mu'(\mu(c_1)) = c_i$  and the previous match of  $\mu'(c_1)$  be  $\mu(\mu'(c_1)) = c_j$ .

**Theorem 2** Given that  $\mu' \neq \mu$ ,  $\mu'$  Pareto dominates  $\mu$  if and only if  $\mu(c_1)$  is better off. Additionally, if  $\mu(c_1)$  is better off,  $\mu'(c_1)$  is also better off.

In the proof, we first see that if all students are weakly better off under  $\mu' \neq \mu$ , then  $\mu(c_1)$  and  $\mu'(c_1)$  must trivially be strictly better off. To prove the other direction, we see that if  $\mu(c_1)$  is better off under  $\mu'$ , then its new match  $c_i$  must be worse off under  $\mu'$ , so that  $(\mu(c_1), c_i)$  is not a blocking pair in  $\mu$ . Lemma 1 states that since  $c_i$  is worse off under  $\mu'$ ,  $\mu(c_i)$  must be better off under  $\mu'$ . We use Lemma 1 to reason that this begins a chain of students necessarily being strictly better off under  $\mu'$ , which only ends with  $\mu'(c_1)$  being matched to  $c_1$ . We then reason about  $\mu'(c_1)$ 's status and find that a contradiction is made if it is worse off under  $\mu'$ . This completes the proof (see Appendix A.1 for full proof).

Although  $c_1$  is the school that changes its preferences, in student-proposing deferred acceptance, it is possible that  $c_1$ 's old match  $\mu(c_1)$  is actually ranked above its new match  $\mu'(c_1)$  in its new preference ranking  $\succ'_{c_1}$ . We define this circumstance as  $c_1$  being “worse off.” We show that it is possible for such an instance to occur and describe the implications for  $\mu(c_1)$  and  $\mu'(c_1)$ 's welfare.

**Corollary 2.1** Let  $c_1$  be “worse off” if  $\mu(c_1) \succ'_{c_1} \mu'(c_1)$  in  $c_1$ 's new ranking. Then, if  $c_1$  is “worse off,”  $\mu'$  Pareto dominates  $\mu$ .

*Proof.* The following example demonstrates that  $c_1$  can be matched to  $\mu'(c_1)$  such that  $\mu(c_1) \succ'_{c_1} \mu'(c_1)$ . Let student preferences in  $G$  and  $G'$  be given by:

$$\succ_{s_1}: c_4 \succ c_3 \succ c_2 \succ c_1$$

$$\succ_{s_2}: c_1 \succ c_4 \succ c_3 \succ c_2$$

$$\succ_{s_3}: c_1 \succ c_4 \succ c_3 \succ c_2$$

$$\succ_{s_4}: c_3 \succ c_4 \succ c_1 \succ c_2$$

Let school preferences in  $G$  be given by:

$$\succ_{c_1}: s_4 \succ s_3 \succ s_1 \succ s_2$$

$$\succ_{c_2}: s_3 \succ s_4 \succ c_1 \succ s_2$$

$$\succ_{c_3}: s_1 \succ s_2 \succ s_4 \succ s_3$$

$$\succ_{c_4}: s_1 \succ s_3 \succ s_4 \succ s_2$$

Let  $c_1$ 's new preferences in  $G'$  be given by:

$$\succ'_{c_1}: s_4 \succ s_2 \succ s_3 \succ s_1$$

Then, the old and new matches are as follows:

$$\mu = \{(s_1, c_4), (s_2, c_3), (s_3, c_2), (s_4, c_1)\}$$

$$\mu' = \{(s_1, c_4), (s_2, c_1), (s_3, c_2), (s_4, c_3)\}$$

In this example,  $c_1$  is matched to  $s_4$  initially and ends up being matched to  $s_2$ , and in  $c_1$ 's new preferences in  $G'$ ,  $s_4 \succ'_{c_1} s_2$ , making  $c_1$  worse off according to the definition above.

Now, we show that  $c_1$  being “worse off” implies that all students are weakly better off. We know from Theorem 2 that all students are at least as well off if and only if  $\mu(c_1)$  is better off. Assume that  $\mu(c_1)$  is worse off while  $c_1$  is “worse off.” Then,  $c_1 \succ_{\mu(c_1)} c_i$ . However, we know that  $c_1$  is “worse off” if  $\mu(c_1) \succ_{c_1} \mu'(c_1)$ . Thus,  $\mu(c_1)$  being worse off creates a contradiction, because  $(\mu(c_1), c_1)$  would form a blocking pair in the stable matching  $\mu'$ .

□

We then consider the complement of Theorem 2 in order to describe the implications for the welfare of all students in  $S$  when  $\mu(c_1)$  is worse off in  $\mu'$ .

**Corollary 2.2** At least one student is made worse off under  $\mu' \neq \mu$  if and only if  $c_1$ 's old partner  $\mu(c_1)$  is worse off in  $\mu'$ .

*Proof.* If  $\mu(c_1)$  is worse off in  $\mu'$ , then at least one person is made worse off in  $\mu'$ , so the if direction holds trivially.

The other direction is essentially the contrapositive of the previous proposition. We showed that if  $\mu(c_1)$  is better off in  $\mu'$ , then all changed students are better off. We know that  $\mu(c_1)$ 's match must change if  $\mu' \neq \mu$  from Corollary 1.1. Thus, the contrapositive of this statement is that if not all changed students are better off (e.g. at least one is worse off), then  $\mu(c_1)$  is worse off. Thus, this statement follows from Theorem 2.

□

We now describe the relationship between the welfare of  $\mu'(c_1)$  and all other students in  $S$  when  $c_1$  performs a preference change.

**Theorem 3** Given that  $\mu' \neq \mu$ ,  $\mu$  Pareto dominates  $\mu'$  if and only if  $\mu'(c_1)$  is worse off. Additionally,  $\mu'(c_1)$  being worse off implies that  $\mu(c_1)$  is also worse off.

The proof follows a similar form as that of Theorem 2. It is trivially true that all students being weakly worse off implies that  $\mu'(c_1)$  and  $\mu(c_1)$  are worse off if the match changed. We then consider the other direction. We reason about the welfare status of  $c_j = \mu(\mu'(c_1))$  and conclude that it must be better off so that there are no blocking pairs in  $\mu'$ .  $\mu'(c_j)$  must then be worse off according to Lemma 1, which begins a chain in which each changed student is necessarily worse off. This chain

only ends with  $\mu(c_1)$  being matched to a school  $c_i$ . We then show that if  $\mu(c_1)$  is better off, this implies that another changed student is worse off, which creates a contradiction (see Appendix A.1 for full proof).

We then use the above statement to describe the relationship between at least one student being better off and  $\mu'(c_1)$ 's welfare.

**Corollary 3.1** At least one student is made better off in  $\mu'$  if and only if  $c_1$ 's new partner  $\mu'(c_1)$  is better off in  $\mu'$ .

*Proof.* If  $\mu'(c_1)$  is better off in  $\mu'$ , then at least one person is made better off in  $\mu'$ , so the if direction holds trivially.

The other direction is essentially the contrapositive of the previous proposition. We showed that if  $\mu'(c_1)$  is worse off in  $\mu'$ , then all students are weakly worse off, meaning that all changed students in  $\mu'$  are worse off. We know that  $\mu'(c_1)$  must change in order for the match to change. Thus, the contrapositive of this statement is that if not all changed students are worse off (e.g. at least one is better off), then  $\mu'(c_1)$  is better off. Thus, Theorem 3 implies the above statement. □

## 4.2 WORST-CASE STUDENT WELFARE FOR SPECIFIC PREFERENCE CHANGES

Now that we have shown that there are relationships between the welfare of all students in  $S$  and the changes experienced by  $\mu(c_1)$  and  $\mu'(c_1)$  in  $G'$ , we turn to characterizing the worst possible impact of the specific types of preference changes  $c_1$  could implement. As described in Chapter 3, we examine arbitrary preference changes, single-swap changes, and priority-based affirmative action.

Previously, we looked at the impact of a change from  $G$  to  $G'$  only from the lens of whether some or all students were weakly better or weakly worse off. We can now capture the magnitude of the

impact using utilitarian count and egalitarian welfare, as described in Section 3.3. Through general-form examples, we can see that that any one of the three types of changes can incur worst cases for both utilitarian count and egalitarian welfare.

#### 4.2.1 WORST-CASE UTILITARIAN COUNT FOR SINGLE-SWAP OR PRIORITY-BASED AFFIRMATIVE ACTION

We present a general-form case where  $n - 1$  students are worse off and 1 student is unchanged through a single swap, with the unchanged student being a student who was made better off in  $c_1$ 's new preference ranking  $\succ'_{c_1}$ .

Let us consider a case where student  $s_i$  for  $i \in [1, n - 2]$  ranks school  $c_i$  as their first preference and school  $c_{i+1}$  as their second preference, while student  $s_{n-1}$  ranks  $c_{n-1}$  as their first preference and  $c_1$  as their second preference, and student  $s_n$  ranks  $c_1$  as their first preference. Thus, student rankings are as such:

$$\begin{aligned} \succ_{s_1}: c_1 \succ c_2 \succ \dots \\ \succ_{s_2}: c_2 \succ c_3 \succ \dots \\ \succ_{s_3}: c_3 \succ c_4 \succ \dots \\ \vdots \\ \succ_{s_{n-1}}: c_{n-1} \succ c_1 \succ \dots \\ \succ_{s_n}: c_1 \succ \dots \end{aligned}$$

Then, let school rankings be as follows: For  $j \in [2, n - 1]$ , school  $c_j$  ranks  $s_{j-1} \succ s_j \succ s_n$ . Let  $c_1$  and  $c_n$  rank  $s_{n-1} \succ s_1 \succ s_n$ , where  $s_1$  and  $s_n$  are at adjacent positions in the ranking. Thus, school

rankings are as below (observe that with the exception of  $s_1$  and  $s_n$  being adjacent to each other in  $c_1$ 's ranking, the exact positions of the students in the rankings can be flexible, as long as the statements below hold):

$$\begin{aligned}
& \succ_{c_1}: \dots s_{n-1} \dots \succ s_1 \succ s_n \dots \\
& \succ_{c_2}: \dots s_1 \dots \succ s_2 \dots \succ s_n \dots \\
& \succ_{c_3}: \dots s_2 \dots \succ s_3 \dots \succ s_n \dots \\
& \quad \quad \quad \vdots \\
& \succ_{c_{n-1}}: \dots s_{n-1} \dots \succ s_1 \dots \succ s_n \dots \\
& \succ_{c_n}: \dots s_{n-1} \dots \succ s_1 \dots \succ s_n \dots
\end{aligned}$$

In  $\mu, \mu(s_i) = c_i$ , as each student  $i \in [1, n - 1]$  has a different first-choice to propose to, while student  $s_n$  is rejected by every school except  $c_n$  because of the preferences above. Now, let us say that in  $\mu'$ ,  $c_1$  changes its preferences by performing the single swap between  $s_1$  and  $s_n$ . Then, in the first round of proposals,  $s_n$  successfully proposes to  $c_1$ , and  $c_1$  rejects  $s_1$ . This causes  $s_1$  to propose to its second choice,  $c_2$ , and be accepted (over  $\mu(c_2) = s_2$ ).  $s_2$  then proposes to its second choice,  $c_3$  and is accepted (over  $\mu(c_3) = s_3$ ). This continues, with  $\mu'(c_j) = s_{j-1} \forall j \in [2, n - 1]$ . Finally,  $s_{n-1}$  proposes to its second choice  $c_1$ . Since  $s_{n-1} \succ'_{c_1} s_n, s_n$  is now rejected by  $c_1$  and  $\mu'(c_1) = s_{n-1}$ . Once again,  $s_n$  can now only propose to and be accepted by  $c_n$ . Thus, the old and new matchings are as follows:

$$\begin{aligned}
\mu &= \{(s_1, c_1), (s_2, c_2) \dots (s_{n-1}, c_{n-1}), (s_n, c_n)\} \\
\mu' &= \{(s_1, c_2), (s_2, c_3) \dots (s_{n-1}, c_1), (s_n, c_n)\}
\end{aligned}$$

In this instance, a single-swap change in  $c_1$ 's preferences (swapping  $s_1$  and  $s_n$ ) causes  $n - 1$  students in the matching to be worse off by switching from their first to their second choices, while the final student ( $s_n$ ) is weakly better off. In short, this example demonstrates that for all  $n > 3$ ,  $n - 1$  students can be made worse off through a single swap or priority-based affirmative action change (as single swaps satisfy the properties of priority-based affirmative action if we let  $S^m = \{s_n\}$  in this case).

#### 4.2.2 WORST-CASE UTILITARIAN COUNT FOR ARBITRARY CHANGE

Next, we show an example of an arbitrary preference change for  $n \geq 4$  in which  $n - 1$  students are worse off and the only student left unchanged in  $\mu'$  was actually moved down in the modified ranking  $\succ'_{c_1}$ . This is, in a sense, “worse” than the example shown in Section 4.2.1 – even if the number of worse-off students in  $\mu'$  is the same, the only weakly better off student in the previous example was still consistent, as its ranking in  $\succ'_{c_1}$  improved. We use similar student and school preferences to those described in the example above:

We once again consider the case where student  $s_i$  for  $i \in [1, n - 2]$  ranks school  $c_i$  as their first preference and school  $c_{i+1}$  as their second preference, student  $s_{n-1}$  ranks  $c_{n-1}$  as their first preference and  $c_1$  as their second preference, and student  $s_n$  ranks  $c_1$  as their first preference. Again, student rankings are as such:

$$\begin{aligned} \succ_{s_1}: c_1 \succ c_2 \succ \dots \\ \succ_{s_2}: c_2 \succ c_3 \succ \dots \\ \vdots \\ \succ_{s_{n-1}}: c_{n-1} \succ c_1 \succ \dots \\ \succ_{s_n}: c_1 \succ \dots c_n \dots \end{aligned}$$

Then, let school rankings be as follows: Once again, for  $j \in [2, n - 1]$ , school  $c_j$  ranks  $s_{j-1} \succ s_j \succ s_n$ . Then, we let  $c_1$ 's initial ranking have  $s_1 \succ s_n \succ s_{n-1}$ , while  $c_n$  can have any set of preferences.

Thus, school rankings are as below:

$$\begin{aligned}
& \succ_{c_1}: s_1 \succ s_n \succ s_{n-1} \dots \\
& \succ_{c_2}: \dots s_1 \dots \succ s_2 \dots \succ s_n \dots \\
& \vdots \\
& \succ_{c_{n-1}}: \dots s_{n-2} \dots \succ s_{n-1} \dots \succ s_n \dots \\
& \succ_{c_n}: \dots
\end{aligned}$$

In  $\mu$ ,  $\mu(s_i) = c_i$  once again, as each student except for  $s_n$  has a distinct first choice, and  $s_n$  is rejected by every school except  $c_n$ . Now, let us say that in  $\mu'$ ,  $c_1$  changes its preferences by reversing them:

$$\succ'_{c_1}: \dots s_{n-1} \succ s_n \succ s_1$$

Then, in the first round of proposals,  $s_n$  proposes to  $c_1$  and is actually successful because  $s_n \succ'_{c_1} s_1$ , thus displacing  $s_1$ .  $s_1$  must then propose to its second choice  $c_2$ , which displaces  $s_2$ , and so on with  $s_i$  for  $i \in [1, n - 2]$  being forced to propose its second choice  $c_{i+1}$ , succeeding, and displacing  $s_{i+1}$ . After being displaced by  $s_{n-2}$ ,  $s_{n-1}$  now proposes to  $c_1$  and is accepted, thus displacing  $s_n$ .  $s_n$  then can re-propose to  $c_n$ , as no student has yet proposed to  $c_n$ . Thus, in  $\mu'$ , every student except  $s_n$  is worse off by 1. Thus,  $\mu$  and  $\mu'$  are as follows:

$$\mu = \{(s_1, c_1), (s_2, c_2) \dots (s_{n-1}, c_{n-1}), (s_n, c_n)\}$$

$$\mu' = \{(s_1, c_2), (s_2, c_3) \dots (s_{n-1}, c_1), (s_n, c_n)\}$$

Although the new matching is the same as that of the example shown for single-swap and priority-based changes in Section 4.2.2, we perform a different rank change in this case to  $s_n$  but achieve the same outcome. Here,  $s_n$  is in a worse rank in  $\succ'_{c_1}$  than it was in  $\succ_{c_1}$ , whereas in the previous example,  $s_n$ 's ranking improved in  $\succ'_{c_1}$ . Thus, this preference change causes  $n - 1$  students to be worse off, while the only unchanged student in  $\mu'$  was moved down in  $\succ'_{c_1}$ . As such, example shows less consistency than the worst case for a single-swap and priority-based change because none of the students who were improved in  $\succ'_{c_1}$  received a better match in  $\mu'$ .

#### 4.2.3 WORST-CASE EGALITARIAN WELFARE FOR ANY PREFERENCE CHANGE

The following example shows that a single-swap change in the ranking of one school can lead a student to jump from their first to last choice. Let student  $s_1$ 's preferences be  $c_1 \succ c_2 \succ \dots c_n$ . Let students  $s_2$  to  $s_n$ 's preferences all be  $c_{n-1} \succ c_{n-2} \dots c_1 \succ c_n$ . Let every school prefer students in the following order:  $s_{n-1} \succ s_{n-2} \dots \succ s_1 \succ s_n$ .

In this original setting,  $s_i$  would be matched to  $c_i$  – students  $s_2$  through  $s_{n-1}$  all propose to the same schools in the same order, and  $s_n$  proposes to  $c_1$  but is rejected because  $s_1$  has already proposed to  $c_1$  and  $s_1 \succ_{c_1} s_n$ .

Now, if  $c_1$  changes its preference ranking by swapping  $s_n$  and  $s_1$  such that its ranking is now  $s_{n-1} \succ \dots s_n \succ s_1$ , then students  $s_2$  through  $s_{n-1}$ 's pairings stay the same, but student  $s_1$  is now rejected by  $c_1$  when student  $s_n$  proposes to  $c_1$ . Thus, the final matching changes such that students  $s_1$  and  $s_n$  switch schools. Thus, student  $s_1$  is now matched to their  $n$ th choice instead of their 1st choice, while student  $s_n$  is now matched to their  $n - 1$ th choice instead of their  $n$ th choice.

The egalitarian welfare in this case is  $-(n - 1)$ . This is the worst possible jump in rank for single student. Since a single-swap preference change is a subset of arbitrary preference changes and a sin-

gle swap can be analogous to a priority-based preference change, this worst-case egalitarian welfare is achievable for all three types of changes.

### 4.3 GUARANTEES FOR SPECIFIC PREFERENCE CHANGES

The examples above show cases for all three types of preference changes that achieve nearly the worst possible utilitarian count welfare ( $n - 1$  students worse off, with the final student unchanged) and the worst possible egalitarian count welfare (a single student worse off by  $-(n - 1)$ ). These discoveries motivate us to explore whether any positive guarantees can be made with regards to student welfare for each type of preference change, as it is potentially troubling that a singular swap, which changes the rankings of only two students out of  $n$ , can achieve the same worst effect on student welfare as an arbitrary change.

In this section, we see that in the single-swap and priority-based affirmative action cases, one student is guaranteed to be weakly better off. For an arbitrary preference change, although the worst utilitarian count we found involved  $n - 1$  students and not  $n$ , we demonstrate through an example that there seems to be no guarantee that any particular student's improvement in  $c_1$ 's ranking implies an improvement from  $\mu$  to  $\mu'$ .

#### 4.3.1 GUARANTEES FOR SINGLE-SWAP AND PRIORITY-BASED AFFIRMATIVE ACTION

First, we use [7] to show that at least one of the students improved under the new ranking  $\succ'_{c_1}$  is guaranteed to be weakly better off in both a single-swap change and priority-based affirmative action.

**Theorem 4** If a single swap is performed, the student  $s_i$  that is made better in  $c_1$ 's ranking must be weakly better off in  $\mu'$ .

This is a specific case of Theorem 5 in [7], which states that student-proposing DA “respects improvements,” meaning that if  $G'$  presents an improvement for a singular student  $s_i$  over  $G$  (e.g.  $s_i$  is weakly better off in  $\succ'_{c_j}$  for every  $c_j$  in  $C$ ), then that student must be weakly better off in  $\mu'$ . [7] shows this by demonstrating that if this were not true, then there would exist a set of false preferences  $\succ'_{s_i}$  that  $s_i$  could independently report in order to improve its outcome in  $G'$ . This presents a contradiction, as student-proposing DA is strategyproof. Since the single-swap preference change strictly improves  $s_i$  in  $c_1$ , it is a special case of this theorem. A.1 also provides a supplementary proof written for this specific case.

A priority-based affirmative action modification consists of performing multiple single swaps between at least 1 minority student and the majority students originally ranked above it in  $c_1$ , while preserving the order of minority students within their group. We now show the guarantee that a single minority student will be weakly better off under priority-based affirmative action.

**Theorem 5** Let  $s_i \in S^m$  be the highest-ranked strictly better off minority student who prefers  $c_1 \succ_{s_i} \mu(s_i)$ .  $s_i$  is guaranteed to be weakly better off under a priority-based affirmative action modification.

The outline for the proof is as follows: We know by definition that if  $s_i$  is matched to  $c_1$  in  $\mu'$ , it must be better off. We also know that if  $s_i \succ'_{c_1} \mu'(c_1)$ , then  $s_i$  must be matched to some  $c_j \succ_{s_i} c_1$  in order to prevent  $(s_i, c_1)$  from being a blocking pair in  $\mu'$ . Thus, we see that the only non-trivial case we must consider is one in which  $\mu'(c_1) \succ'_{c_1} s_i$ , where  $\mu'(c_1)$  did not propose to  $c_1$  in  $G$  but now does in  $G'$ . We show that  $s_i$  cannot be made worse off in  $\mu'$  for this case by comparing the proposals that occur in  $G'$  to those in  $G$ . We reason that  $s_i$  must be, at least temporarily, accepted by  $c_1$  at some point in  $G'$  over another student that was accepted by  $c_1$  in  $G$ . Then, the proof is reduced to showing that if  $\mu'(s_i) \neq c_1$  in this case,  $s_i$  is still at least as well off as before. This can be shown by considering the possible orderings of  $s_i$  and  $\mu(c_1)$  in  $\succ_{c_1}$  and  $\succ'_{c_1}$ . In each case, we demonstrate that  $s_i$  is made weakly better off by following the chain of new proposals induced by  $c_1$  temporarily

accepting  $s_i$  in  $G'$ . The full proof is included in Appendix [A.1](#).

Thus, we have shown that the worst-case utilitarian count welfare described in Section [4.2.1](#) is truly the worst case for single-swap and priority-based affirmative action preference changes, as one (minority) student is guaranteed to be weakly better off under these types of preference changes. However, it is worth noting that only one minority student out of all the students in the minority group in priority-based affirmative action is guaranteed to be weakly better off, and that this can occur at the expense of all other students in the minority group, since the worst-case utilitarian count welfare is achievable.

#### 4.4 LACK OF GUARANTEE FOR AN ARBITRARY CHANGE

We can also show that such a guarantee does not seem to exist for the arbitrary preference change. Although we were unable to show that  $n$  students could be worse off under a preference change by  $c_1$ , we argue that  $n - 1$  students produces a similarly undesirable outcome that points towards potential issues with unilateral affirmative action. Particularly, the following example demonstrates that there can be cases where the highest-ranked strictly better off student  $s_i$  that prefers  $c_1$  to their old match is worse off under an arbitrary preference change.

Let student preferences in  $G$  and  $G'$  be given by:

$$\succ_{s_1}: c_1 \succ c_2 \succ c_3 \succ c_4$$

$$\succ_{s_2}: c_2 \succ c_3 \succ c_4 \succ c_1$$

$$\succ_{s_3}: c_1 \succ c_2 \succ c_4 \succ c_3$$

$$\succ_{s_4}: c_1 \succ c_3 \succ c_2 \succ c_4$$

Let school preferences in  $G$  be given by:

$$\succ_{c_1}: s_4 \succ s_3 \succ s_2 \succ s_1$$

$$\succ_{c_2}: s_2 \succ s_3 \succ s_1 \succ s_4$$

$$\succ_{c_3}: s_4 \succ s_1 \succ s_2 \succ s_3$$

$$\succ_{c_4}: s_1 \succ s_4 \succ s_3 \succ s_2$$

Let  $c_1$ 's new preferences in  $G'$  be given by:

$$\succ'_{c_1}: s_2 \succ s_3 \succ s_1 \succ s_4$$

Then, the old and new matches are as follows:

$$\mu = \{(s_1, c_3), (s_2, c_2), (s_3, c_4), (s_4, c_1)\}$$

$$\mu' = \{(s_1, c_4), (s_2, c_2), (s_3, c_1), (s_4, c_3)\}$$

The highest-ranked strictly better off student  $s_i$  in  $\succ'_{c_1}$  where  $c_1 \succ_{s_i} \mu(s_i)$  in this case is  $s_1$ , as students  $s_2$  and  $s_3$  are made better off in  $\succ'_{c_1}$  as well but do not prefer  $c_1$  to their old matches. However, we see that  $s_1$  receives  $c_4$ , its last-choice school, rather than its previous match  $c_3$ . Thus, it is not guaranteed under an arbitrary preference change that the highest-ranked, strictly better off student who prefer  $c_1$  to their old match receives a better match in  $\mu'$ .

# 5

## Simulation Results

In previous chapters, we determined that unilateral preference changes by a single school in deferred acceptance can incur worst cases for utilitarian count and egalitarian welfare, which demonstrates the extreme sensitivity of student welfare with respect to a single school's decision: a change involving very few students leads to significant degradation in welfare that may even be detrimental the students we intend to uplift. However, we also see that when this preference change adheres to the restrictions of priority-based affirmative action, at least one minority student preserves its welfare.

This positive guarantee could make unilateral affirmative action more viable.

Thus, in order to draw more definite conclusions about whether unilateral affirmative action is potentially effective towards its purpose, we use simulations in Python to characterize the average and general behavior of student-proposing DA under arbitrary preference changes, priority-based affirmative action, and single-swap preference changes. By examining the average and variance of different metrics under each preference change, we give a more accurate representation of the expected impact of a unilateral preference change and can thus more confidently give insights to policymakers regarding school districts that use deferred acceptance.

## 5.1 SIMULATION METHODS

We use Python to implement and simulate the deferred acceptance matching algorithm over 10000 iterations. Since we do not have adequate information regarding how student and school preferences may be correlated, we perform simulations in which students' and schools' preferences are picked randomly and uniformly, with preferences modeled as permutations of the list of all students or schools. We use NumPy's random permutation selector for this purpose, and preference profiles are uniformly and randomly selected for every iteration of the matching in order to ensure that each iteration is performed on a unique  $G$  and  $G'$ .

In the United States, the average school district contains 5.6 schools, although there is a high amount of variation in school district size [1]. As such, we simulate for relatively small  $n \in [3, 4, 5, 6, 7, 8, 9, 10, 12, 16, 20]$ .

For single-swap preference changes, we pick a random index  $i \in [2, n]$  and a student  $s$  such that  $r_{q_1}(s) = i$ . We then swap the student in rank  $i$  with the student in rank  $i - 1$ . For priority-based affirmative action, we also simulate over the number of minority students  $k \in [1, n - 1]$ . For the purposes of these simulations, if there are  $k$  minority students, then all  $k$  students are made strictly

better off. The priority-based rank change is then simulated as follows:

1. First, the minority students  $s_i \in S^m$  are ordered based on their rank in  $\succ_{c_1}$ .
2. Let  $s_i$  be the highest-ranked minority student and have rank  $r_{c_1}(s_i)$ . Then,  $r'_{c_1}(s_i)$  is uniformly and randomly selected from  $[1, r_{c_1}(s_i) - 1]$ .  $s_i$  is then inserted at  $r'_{c_1}(s_i)$  and removed from  $r_{c_1}(s_i)$ . This preserves the relative ordering of all majority students within their group.
3. For any subsequent minority students  $s_k$ , the new rank  $r'_{c_1}(s_k)$  is selected from  $[r'_{c_1}(s_j) + 1, r_{c_1}(s_k) - 1]$ , where  $s_j$  is previous minority student (e.g. the new rank for  $s_k$  must never be higher than that of the new rank for  $s_j$ ). This preserves the ordering of all minority students in their group.
4. This process continues for each minority student.

For each  $n$  or  $(n, k)$  pair (in the case of priority-based affirmative action), we run the selection of preferences and simulate student-proposing DA for 10000 iterations. We then find averages, distributions, and worst cases for the metrics described in Section 3.3: utilitarian count, egalitarian welfare, utilitarian welfare, collateral damage, and consistency.

We first measure the number of iterations out of 10000 where the  $\mu \neq \mu'$  in order to determine how often the matching actually changes due to a single school's preference change. For single-swap and arbitrary preference changes, we and measure the average, standard deviation, and worst case for each metric across all iterations where  $\mu' \neq \mu$  and graph these results for all values of  $n$ . For priority-based affirmative action, we also do the same over different values of  $k$  for each  $n$ .

We also describe how to measure minimal responsiveness and strict and weak consistency for both classes. We calculate the rate of minimal responsiveness by taking the proportion of runs with at least one worse-off minority student  $s_i \in S^m$  where the criteria for minimal responsiveness (defined in Definition 8) is satisfied. The rate of minority- and majority-class weak or strict consistency

is the proportion of students out of each class whose outcomes were consistent, as defined in Definitions 9 and 10.

Finally, for each  $n$  or  $(n, k)$  pair, we categorize students  $s_i \in S$  by the value of their rank change from  $\succ_{c_1}$  to  $\succ'_{c_1}$ , or  $r_{c_1}(s_i) - r'_{c_1}(s_i)$ . Then, for each set of students with the same rank change, we take the average and standard deviation of their utilitarian welfare  $U(s_i) = r_{s_i}(\mu(s_i)) - r_{s_i}(\mu'(s_i))$  over all students in the set. This comparison shows what the expected welfare change after unilateral affirmative action is for a rank change of magnitude  $r$ . We hope to see a positive correlation between utilitarian welfare change and rank change, as this would indicate that being improved in  $c_1$ 's ranking as an individual student would correlate with a better match in  $\mu'$ .

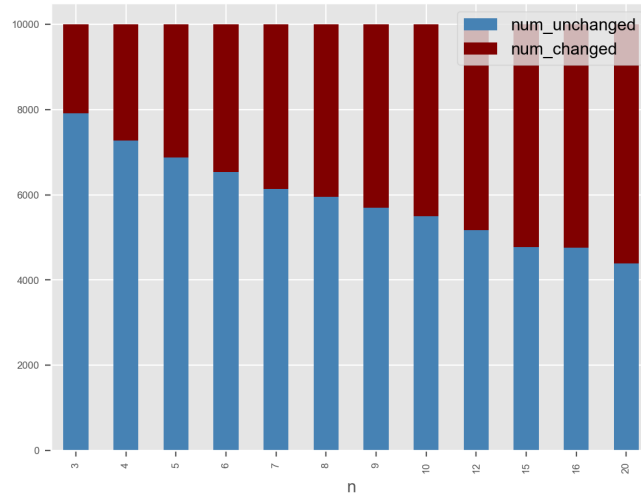
These metrics measure the sensitivity of student welfare to an affirmative action change. They show whether the ranking change performed by  $c_1$  is actually reflected in the changed matching  $\mu'$ , how consistently the changes in  $\mu'$  follow the changes made in  $\succ'_{c_1}$ , and how much students who were unchanged in  $\succ'_{c_1}$  can be affected by changes made to others' rankings.

## 5.2 ARBITRARY CHANGE

In Corollary 1.1, we established the relationship between the match  $\mu'$  changing and  $c_1$ 's match changing. However, we must also examine how often such a change actually occurs.

Figure 5.1 shows that the proportion of iterations in which  $\mu' \neq \mu$  is increasing with  $n$ . The proportion begins around 20% for  $n = 3$  and grows to over 50% for  $n = 20$ . This demonstrates that for small values of  $n$ , even an arbitrary change will not be highly likely to change the match at all, so there is a limit to how much change can actually be incurred by changing the preferences of just one school out of  $n$ .

We then observe the average, standard deviation, and worst/best values of the metrics described above across each value of  $n$ .

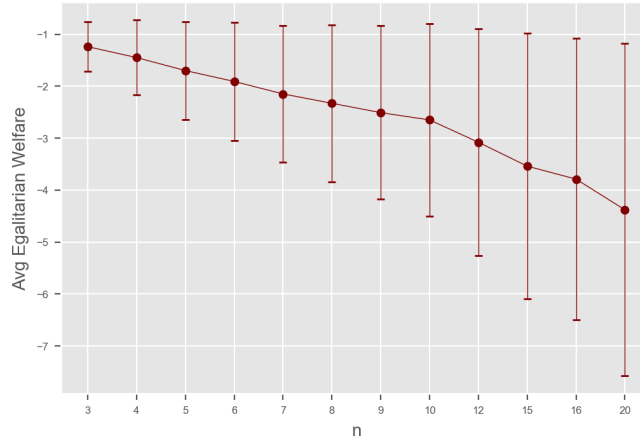


**Figure 5.1:** Number of iterations out of 10000 where the matching changed due to an arbitrary preference shuffling  $\succ'_{c_1}$ . Frequency of a match change increases with  $n$ .

n	Worst Util Count	Worst Egal Welfare	Worst Util Welfare	Best Util Welfare	Worst Collateral Damage
3	2	-2	-2	2	1
4	3	-3	-6	5	2
5	4	-4	-9	10	3
6	5	-5	-13	10	4
7	6	-6	-17	14	3
8	7	-7	-23	19	3
9	7	-8	-28	23	3
10	8	-9	-30	23	5
12	9	-11	-34	32	3
16	12	-15	-45	74	3
20	14	-19	-77	69	5

**Table 5.1:** Summary of worst and best performance for several metrics when an arbitrary preference change is made.

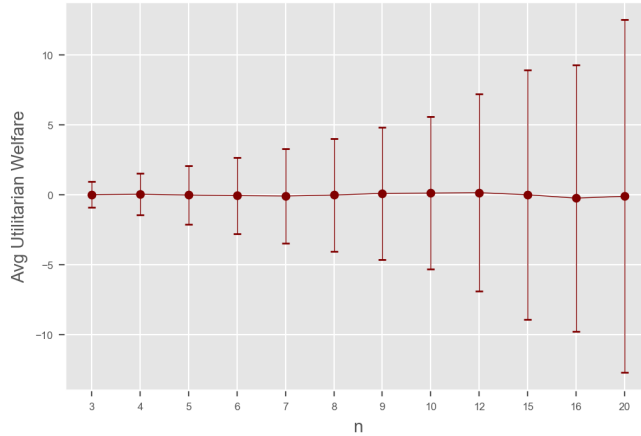
First, we can examine utilitarian count and egalitarian welfare. In Section 4.2.2, we showed that a utilitarian count of  $n - 1$  and the worst egalitarian welfare  $-(n - 1)$  can be achieved by an arbitrary preference change. In Table 5.1, we see that as  $n$  increases past 8, the worst case may not be observable in 10000 iterations, thus showing that these worst cases are achievable but rare.



**Figure 5.2:** Egalitarian welfare averaged across 10000 iterations, which show a decreasing trend over  $n$ . Error bars represent standard deviation.

Figure 5.2 shows that average egalitarian welfare, or how much worse off the worst off student was on average, decreases with  $n$ , while its standard deviation increases with  $n$ . On average, the egalitarian welfare is not worse than  $-5$  for any  $n$ , but within one standard deviation,  $n = 20$  can achieve egalitarian welfare of  $-7$ . The increasing standard deviation with  $n$  shows that the worst-case welfare for a single student can be unpredictable. In fact, the worst-case egalitarian welfare described in Section 4.2.3 is actually observable for all  $n$  within 10000 iterations, as seen in Table 5.1.

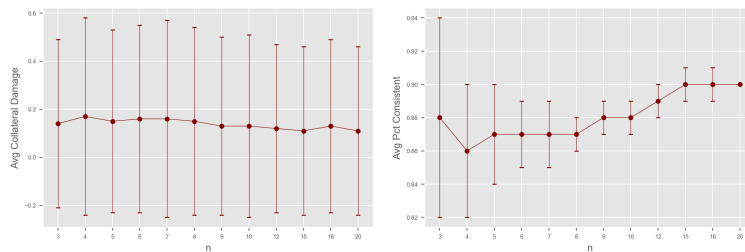
Let us now consider students' utilitarian welfare over values of  $n$  in Figure 5.3. Figure 5.3 shows that average utilitarian welfare for an arbitrary change is around 0, but that variance seems to increase exponentially with  $n$ , as seen by the range covered in one standard deviation. For each value



**Figure 5.3:** Utilitarian welfare averaged across 10000 iterations. Error bars show standard deviation, which increases over  $n$  while average utilitarian welfare remains relatively constant around 0.

of  $n$ , utilitarian welfare is also distributed approximately normally, though a slight left skew can be observed for larger  $n$  (see Figure A.1). Comparing the specific values in Table 5.1 also shows that in general, over 10000 iterations, the magnitude of the worst  $U(S)$  is larger than that of the best  $U(S)$ , and that the worst  $U(S)$  per student is generally increasing with  $n$ , but increases more slowly as  $n$  increases.

We now turn to the metrics of collateral damage and consistency, as well as the comparison of average match rank change for each possible value for preference rank change.

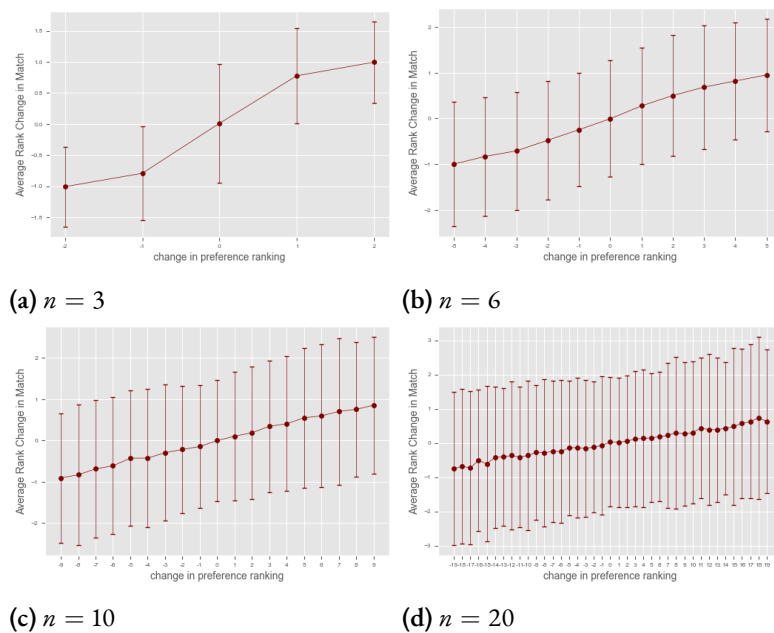


(a) Average collateral damage.

(b) Average % of consistent students.

**Figure 5.4:** Collateral damage (Def. 6) and consistency (Def. 7) averaged over 10000 iterations. Error bars show standard deviation.

Figure 5.4 shows that the amount of collateral damage remains relatively constant in the range of  $(0.1, 0.2)$  with standard deviation  $\approx 0.40$ , meaning that there are close to 0 students  $s_i$  who were unchanged by  $c_1$  that ended up worse off, and that the number of such students does not scale with  $n$ . The worst collateral damage recorded in Table 5.1 shows that the number of unchanged students who are negatively affected is not necessarily monotonically increasing with  $n$ . The consistency graph shows that as  $n$  increases, the proportion of students whose match outcomes are consistent with their rank changes in  $c_1$  is over 80% and generally increasing with  $n$ . The standard deviation also decreases with  $n$ , showing that small values of  $n$  are actually more volatile with regards to consistency.



**Figure 5.5:** Average change in student welfare for each possible change in rank number, graphed for several  $n$ . Averages show a positive correlation, but the high standard deviation represented by error bars makes it difficult to infer a true positive trend.

Figure 5.5 shows that there is a positive correlation between the the amount of change a student experiences in  $c_1$ 's preference ranking and  $U(s_i)$ . While this positive trend seems to indicate that a

unilateral change by  $c_1$  leads to outcomes that are relatively consistent (e.g. being worsened in the ranking leads to receiving a worse match), the growing variance with  $n$  demonstrates that the average case is not a reliable indicator of student welfare for larger  $n$ . Additionally, the range of the average change in match welfare stays constant within  $[-1, 1]$  and does not scale with  $n$ , which is somewhat troubling – it would seem intuitive that worsening a student by 10 ranks should incur much worse performance than worsening a student by 3 ranks. However, this is not the case. Moreover, it can be stated that the ideal trend for this graph would be if the error bars for preference rank changes  $> 0$  remained above 0, and vice versa. However, the  $n = 10$  and  $n = 20$  cases shows that within one standard deviation, students who have been moved far down in the ranking can experience positive changes in their match and vice versa. This stands in contrast to the trend shown in Figure 5.4. However, the consistency being measured is based on weak orderings – a student who was improved in  $\succ'_{c_1}$  is considered “consistent” only if it is not *strictly* worse off. Thus, we can conclude that the large standard deviation is likely because many students are remaining unchanged in spite of their movement in  $\succ'_{c_1}$ . This allows us to reconcile the high consistency rate and high standard deviation in welfare change for a particular rank change in  $\succ'_{c_1}$ .

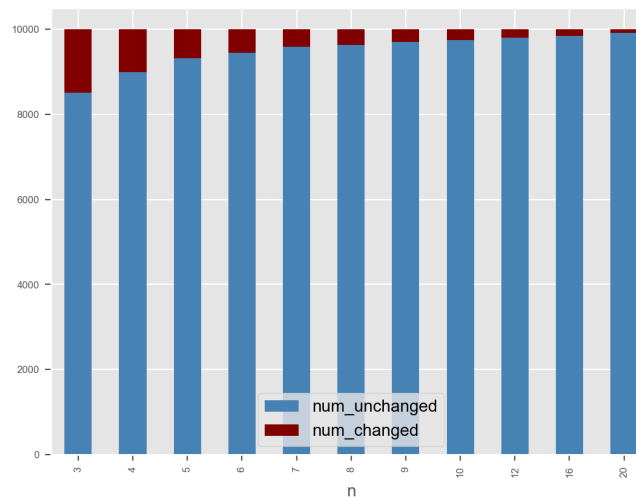
Our simulative analysis of  $c_1$ 's arbitrary preference changes shows that  $c_1$  has more “power” to change the match as  $n$  increases. Utilitarian welfare is generally distributed normally around 0, a result that is corroborated by our runs producing similar numbers of worse and better off students in  $\mu'$  (see Figure A.2). We do see some positive trends, as the worst cases for utilitarian count found in Section 4.2.2 are not even observed in 10000 iterations for larger  $n$ , the collateral damage is low for an arbitrary preference change, consistency is high, and there is positive correlation between average match welfare change  $U(s_i)$  and preference ranking change in  $\succ'_{c_1}$ . However, the high variance for utilitarian and egalitarian welfare as well as for match welfare show that the results in expectation cannot entirely be interpreted in favor of unilateral preference changes. Rather, the high variance and worst cases show that student welfare that is both consistent and actually sensitive to the

changes made in  $\succ'_{c_1}$  is difficult to come by for an arbitrary preference change.

### 5.3 SINGLE-SWAP CHANGE

In Section 4.2.1, we found that single-swap preference changes in  $c_1$  could produce worst-case results for utilitarian count and egalitarian welfare similar to those of arbitrary preference changes. However, Theorem 4 states that the student swapped to a higher ranking is guaranteed to be at least as well off as before. We now look to visualizations and metrics to determine if single-swap preference changes display more consistency than arbitrary changes, or if the “magnitude” of the change to the preference ranking does not seem to be of consequence to student welfare.

As before, we examine how often the match actually changes over 10000 iterations with a single swap:



**Figure 5.6:** Number of iterations out of 10000 where the matching changed due to a single swap of two students in  $\mu'_{c_1}$ . Frequency of a match change decreases with  $n$ .

Figure 5.6 shows that the proportion of iterations  $n$  where  $\mu' \neq \mu$  is much lower than when an arbitrary change is made and decreases with  $n$ . This makes intuitive sense, as we would expect that

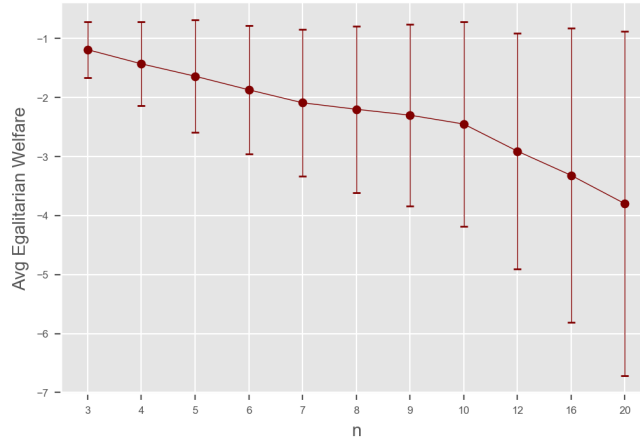
n	Worst Util Count	Worst Egal Welfare	Worst Util Welfare	Best Util Welfare	Worst Collateral Damage
3	2	-2	-2	2	1
4	3	-3	-4	4	2
5	4	-4	-8	7	3
6	5	-5	-9	10	4
7	5	-6	-12	11	4
8	6	-7	-14	15	5
9	6	-8	-14	12	5
10	6	-8	-17	19	5
12	8	-10	-22	19	7
16	8	-12	-28	41	7
20	9	-14	-27	37	8

**Table 5.2:** Summary of worst and best performance for several metrics when a single-swap preference change is made.

changing less students would be less likely to create changes in the matching, and that the impact of a single swap would be lessened with more students and schools.

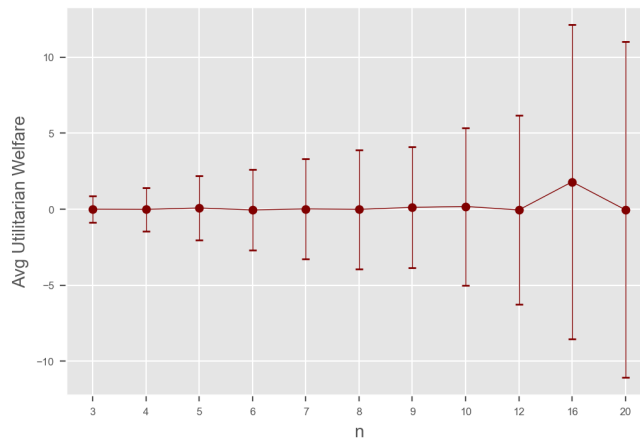
We then observe the average, standard deviation, and worst/best values of the metrics described above across each value of  $n$ . First, we can examine utilitarian count and egalitarian welfare. In Section 4.2.1, we show that the worst possible utilitarian count for a single-swap preference change is  $n - 1$ , and the worst egalitarian welfare is  $-(n - 1)$ . In Table 5.2, we see that as  $n$  increases past 9, the worst case may not be observable in 10000 iterations, thus showing that these worst cases are achievable but rare. Particularly, the worst case for egalitarian welfare is not observable for  $n > 9$ . Overall, if we compare these results to those for the arbitrary preference change in Table 5.1, we can see that both the worst utilitarian count and egalitarian welfare for the single-swap case are never “worse” than those for the arbitrary case.

Figure 5.7 shows that average egalitarian welfare shows a similar trend to that of an arbitrary change. On average, the egalitarian welfare is not worse than  $-4$ , but within one standard deviation,  $n = 20$  can achieve egalitarian welfare of over  $-6$ . This average and standard deviation is only slightly less than that of the arbitrary preference change (as shown in Figure 5.2), which shows that



**Figure 5.7:** Egalitarian welfare averaged across 10000 iterations, which show a decreasing trend over  $n$ . Error bars represent standard deviation.

the worst-off student can receive a similar welfare through a much smaller change in the preference ranking. We now consider the utilitarian welfare over  $n$  in Figure 5.8.



**Figure 5.8:** Utilitarian welfare averaged across 10000 iterations. Error bars show standard deviation, which increases over  $n$ . Averages remain relatively constant at 0, with the exception of an outlier at  $n = 16$ .

Figure 5.8 shows that average utilitarian welfare for a single-swap change is around 0 with the ex-

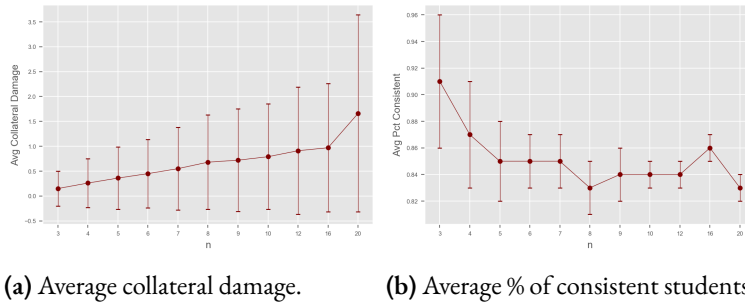
ception of an outlier at  $n = 16$ , and that standard deviation seems to increase exponentially with  $n$ , as seen by the range covered in one standard deviation. Though we may expect that the range of utilitarian welfare within one standard deviation would be lower for a single-swap preference change, comparing this to the average utility for arbitrary preference changes shown in Figure 5.3 shows that there is little significant difference between the average and variance of utilitarian welfare for the two types of changes.

We observe a similar trend in the best and worst cases for utilitarian welfare and compare them to the observations for arbitrary preference changes.

Comparing the specific values in Table 5.2 shows that in general, over 10000 iterations, the magnitude of the best  $U(S)$  is larger than that of the worst  $U(S)$ . This is supported by the fact that the distributions of utilitarian welfare for each  $n$  are slightly skewed right (see Figure A.3), which stands in contrast to the slight left skew and clearer normal distribution observed for arbitrary preference changes. We can compare the worst and best  $U(S)$  to those shown in Table 5.1 and see that the range of  $U(S)$  is indeed smaller for a single swap than for an arbitrary preference change. Thus, although the distribution of utilitarian welfare within one standard deviation seems similar for these two types of preference changes, the worst and best cases are less extreme for a single swap.

So far, we have seen that match change is less probable under a single swap than under an arbitrary preference change, and that when it does happen, the total impact on student welfare is less drastic, but the impact on the worst-off student can be similar. We now turn to the metrics of collateral damage and consistency, as well as the comparison of average match rank change  $U(s_i) = r_{s_i}(\mu(s_i)) - r_{s_i}(\mu'(s_i))$  for each preference rank change amount  $r_{c_i}(s_i) - r_{c'_i}(s_i)$ .

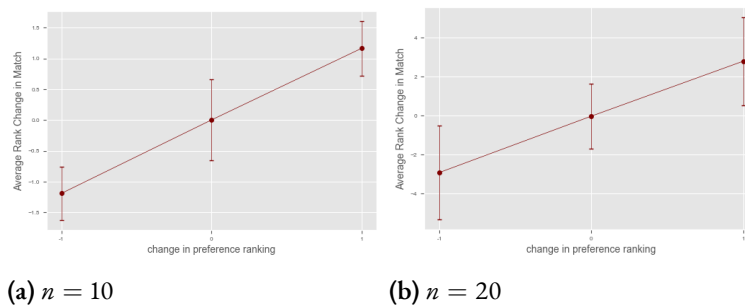
Figure 5.9 shows that the average and standard deviation of collateral damage both increase with  $n$ . This seems to stand in contrast to the trend for collateral damage with an arbitrary preference change (shown in Figure 5.4), which remains relatively constant over  $n$ . The consistency graph shows that as  $n$  increases, the proportion of students whose match outcomes are consistent with



**Figure 5.9:** Collateral damage and consistency averaged over 10000 iterations for all  $n$ . Error bars show standard deviation.

their rank changes in  $c_1$  is also  $> 80\%$  but generally declines as  $n$  increases (with two outliers at  $n = 8$  and  $n = 16$ ). We conjecture that the declining consistency rate is due to unchanged students being made better or worse off, as students who are unchanged from  $G$  to  $G'$  are more common in a single-swap change than for an arbitrary preference change.

We now corroborate our findings about the consistency rate through graphing average welfare change  $U(s_i)$  for a student  $s_i$  with a given change in preference ranking in  $\succ'_{c_1}$ . Note that since we only perform single swaps, the only possible preference ranking changes are  $[-1, 0, 1]$ .



**Figure 5.10:** Average change in student welfare for each possible change in rank number, graphed for two different  $n$ . Averages show a positive correlation. These graphs are representative of the trends found for all  $n$ .

Figure 5.10 shows that there is a positive correlation between the the amount of change occurring in the preference ranking and  $U(s_i)$ . We can compare these results to those for the arbitrary preference change shown in Figure 5.5. For the single swap, the range of the average  $U(s_i)$  seems

to increase with  $n$ , whereas it stayed relatively constant within  $[-1, 1]$  for an arbitrary preference change. Additionally, each follows the ideal trend, with the error bars for preference rank changes  $> 0$  remaining above 0, and vice versa. The standard deviation does also increase with  $n$ , but overall, it seems that single-swap preference changes produce outcomes for student welfare that are more consistent than those of arbitrary preference changes.

Our simulative analysis of single-swap preference changes for  $c_1$  shows that  $c_1$  has less “power” to change the match than it does through an arbitrary preference change and as  $n$  increases. This makes intuitive sense, as we would expect the impact of a single swap of two adjacent students to be less than that of an arbitrary change and to be diluted as more schools and students are added. The worst-case egalitarian welfare is also less severe than that of an arbitrary preference change. Although there is more collateral damage than in the arbitrary preference change case, we can attribute at least some of this to the simple fact that more students remain unchanged under a single-swap preference change. Finally, single swaps show a positive correlation between average  $U(s_i)$  and rank change in  $\succ'_{c_1}$ , with the worsened student ending with a worse match and the improved student ending with a better match even within one standard deviation of its average  $U(s_i)$ . This aligns with our findings in Theorem 4 and points to the idea that making single-swap and arbitrary changes in  $\succ'_{c_1}$  do differ in their impact on student welfare, and that singular changes can produce the desired outcomes for student welfare.

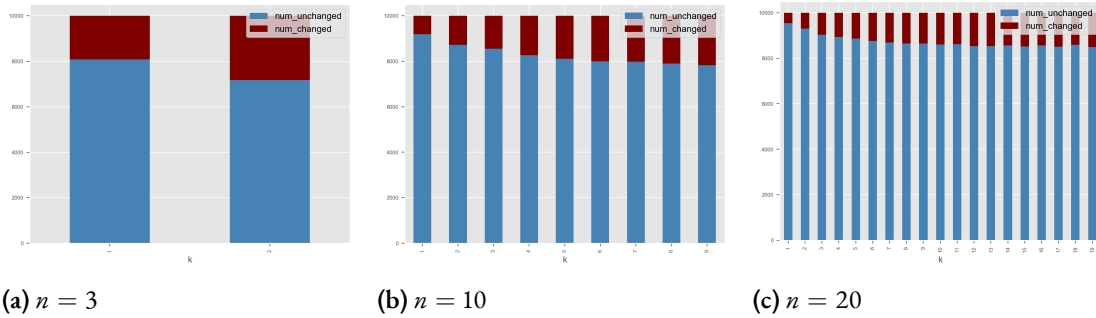
#### 5.4 PRIORITY-BASED AFFIRMATIVE ACTION

Thus far, the outcomes for single-swap preference changes point towards potentially consistent and positive outcomes, but the results for arbitrary preference changes show that there can be wide variation in the outcomes when multiple students are changed at once in  $\succ'_{c_1}$ . As a subset of arbitrary preference changes and a superset of single-swap changes, priority-based affirmative action involves

both of these driving forces, with multiple minority students being changed through multiple single swaps. Since priority-based affirmative action is closest to the type of unilateral affirmative action policy that might be implemented by a school, we simulate the outcomes for the entire system, as well as the outcomes of the minority and majority student groups across  $n$  and  $k$ .

Priority-based affirmative action changes involve selecting a number of students and schools  $n$  and a number of minority students  $k$ . Thus, our analysis will primarily consist of comparing metrics across the two groups for various values of  $n$ , as well as across different values of  $k$  for each  $n$ .

We begin by examining the number of times that  $\mu' \neq \mu$  for various values of  $n$  across each possible value of  $k$ :

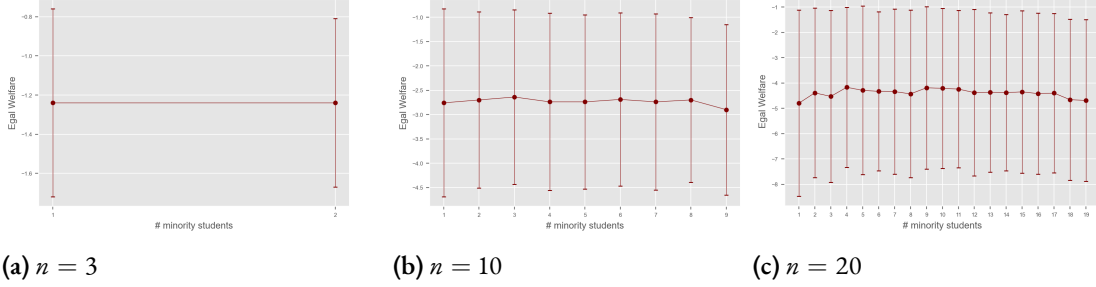


**Figure 5.1:** Number of iterations out of 10000 where the matching changed due to priority-based affirmative action. The change over possible values of  $k$  is graphed for three values of  $n$ . Frequency of a match change increases when  $k$  is small but seems to plateau as  $k$  increases.

We see from Figure 5.1 that the number of iterations out of 10000 where  $\mu' \neq \mu$  generally increases across  $k$ , although the increase seems to become indiscernibly small after  $k = 5$  for  $n = 10$  and  $k = 8$  for  $n = 20$ . Regardless of the value of  $k$ , each  $n$  seems to experience less matching changes than for an arbitrary preference change (see Figure 5.1).

We then observe the average, standard deviation, and worst/best values of the metrics described in Section 3.3 across each value of  $k$  for some informative values of  $n$  to show trends. As described in Section 5.1, we measure utilitarian count, egalitarian welfare, and utilitarian welfare for each  $(n, k)$

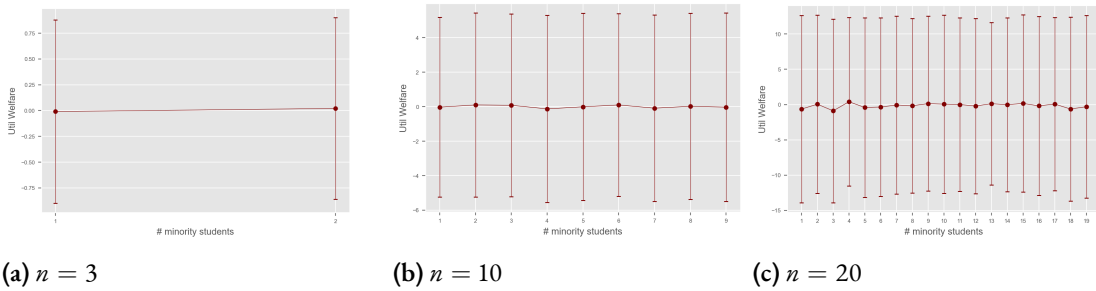




**Figure 5.13:** Egalitarian welfare averaged across 10000 iterations. Egalitarian welfare is constant across different values of  $k$ , but decreases as  $n$  increases. This mirrors the trend observed for egalitarian welfare when an arbitrary or single-swap change is made.

Figure A.5). Just as with the other two types of unilateral preference changes, this shows that minority students' match welfare in  $\mu'$  can still experience a large drop, though these cases are relatively rare.

Next, we observe average utilitarian welfare  $U(S)$ , worst and best utilitarian welfare for each group, and the utilitarian welfare difference between  $U(S^M)$  and  $U(S^m)$ .

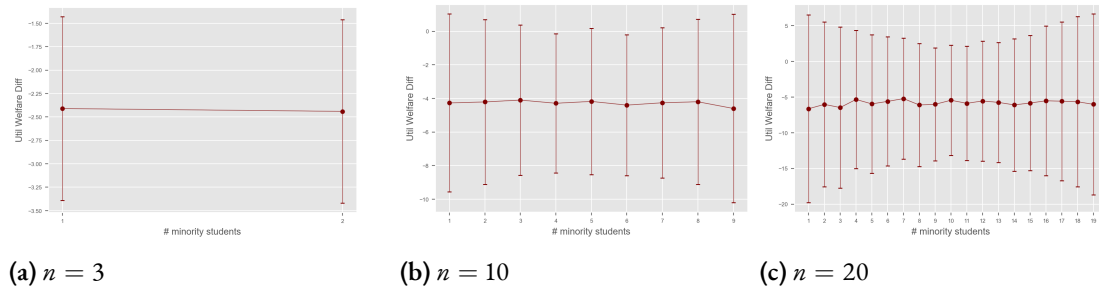


**Figure 5.14:** Utilitarian welfare averaged across 10000 iterations. Error bars show standard deviation, which is constant over different values of  $k$  but increases with  $n$ . Averages remain relatively constant at 0 across both  $k$  and  $n$ .

We see from Figure 5.14 that the average  $U(S)$  is similar to the trends seen for arbitrary preference changes (Figure 5.3) and single swaps (Figure 5.8). The average utilitarian welfare is around 0 for all  $n$  and  $k$ , and the standard deviation remains constant across  $k$  but increases across  $n$ . We can also observe that the trend in worst and best case utilitarian welfare across  $n$  and  $k$  for  $S^m$  and  $S^M$  is also

similar to that of worst utilitarian count by class. Over increasing values of  $k$ , worst utilitarian welfare worsens for minority students and increases for majority students, and best utilitarian welfare improves for minority students and worsens for majority students. It is also worth noting that for all  $n$ ,  $\min(U(S^m)) > \min(U(S^M))$  and  $\max(U(S^M)) < \max(U(S^m))$ , so the minority class achieves a higher best  $U$  and a better worst  $U$  than the majority class (see Figure A.6).

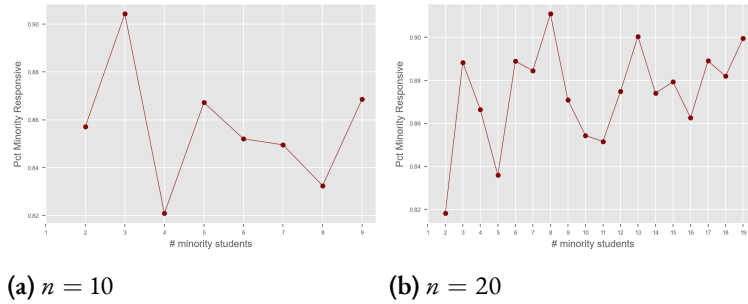
Next, we graph the average of  $U(S^M) - U(S^m)$ , or the difference in utilitarian welfare for the two classes. What we hope to see is that this value is generally negative, so the majority class is, on average, worse off than the minority class.



**Figure 5.15:** Average difference in utilitarian welfare  $U(S^M) - U(S^m)$  between the minority and majority classes across 10000 iterations. The average difference remains constant over all  $k$ , is decreasing over  $n$ . The difference is negative for all  $n$  on average, but is positive within one standard deviation, showing that it is possible for the majority class to do better than the minority class in utilitarian welfare.

The averages shown in figure 5.15 do generally follow the desired trend, with the average  $U(S^M) - U(S^m)$  being negative and becoming larger in magnitude with  $n$ , though it seems to stay about the same with  $k$ . The standard deviation is generally larger for more extreme values of  $k$  (closer to 1 or  $n - 1$ ). However, as  $n$  increases, we can see that within one standard deviation, it is possible for this difference to be positive (e.g. the majority class does better than the minority class). This is somewhat troubling for priority-based affirmative action as a policy, as it shows that regardless of the actual values of  $U(S^m)$  and  $U(S^M)$  it is common and likely for larger values of  $n$  that the majority class would do better than the minority class.

We then observe the rate of minimal responsiveness over all runs, as well as the minority- and majority-class strict and weak consistency over several values of  $n$  and  $k$ . As stated in Definition 8 and Section 5.1, the rate of minimal responsiveness is defined as the proportion of runs in which at least one minority student was worse off in  $\mu'$  where at least one other minority student was better off in  $\mu'$ .

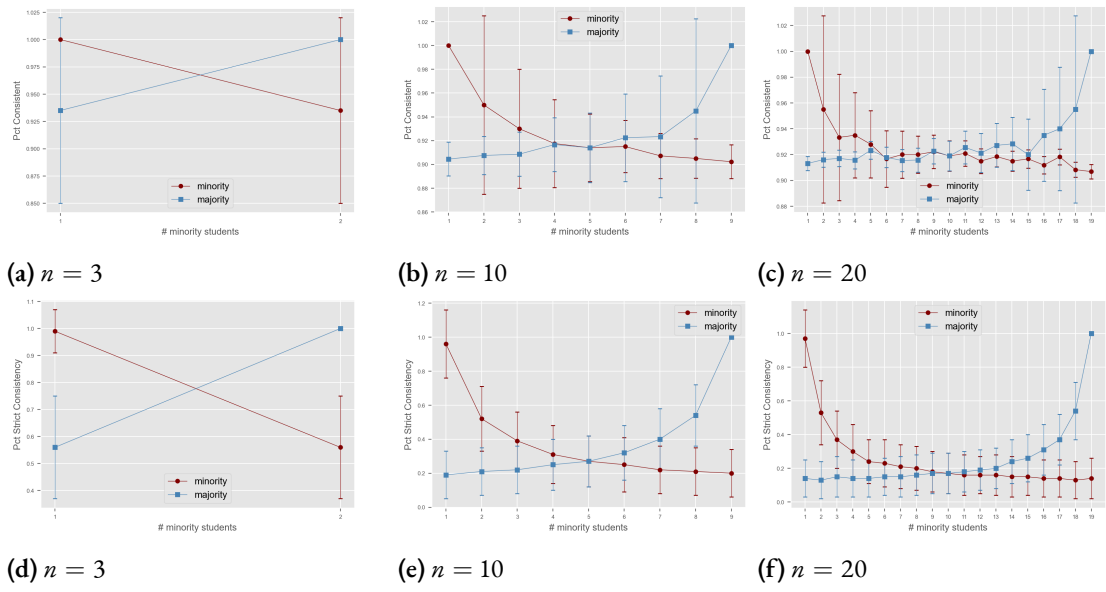


**Figure 5.16:** Percent of iterations out of 10000 in which the new match  $\mu'$  displays minimal responsiveness, as defined in Definition 8.

Figure 5.16 shows that the rate of minimal responsiveness for  $n = 10$  and  $n = 20$  is over 80%, meaning that in under 20% of runs where at least one  $s_i \in S^m$  is worse off,  $\mu$  Pareto dominates  $\mu'$ . There does not seem to be a clear trend in the relationship between minimal responsiveness and  $n$  or  $k$ , however.

Minimal responsiveness only measures whether *at least one* minority student is better off when a majority student is worse off. However, this does not capture the trend in *how many* minority students out of  $k$  total are generally made better off. Thus, we observe the weak and strict consistency rate for both classes.

Figure 5.17 shows that the proportion of minority students who are weakly better off (weakly consistent) is decreasing with  $k$ . For majority students, the proportion of weakly worse off (weakly consistent) students increases with  $k$ . However, we can see that the average proportion of students in both classes who are consistent is above 90% for all  $n$ . This demonstrates that for the most part,



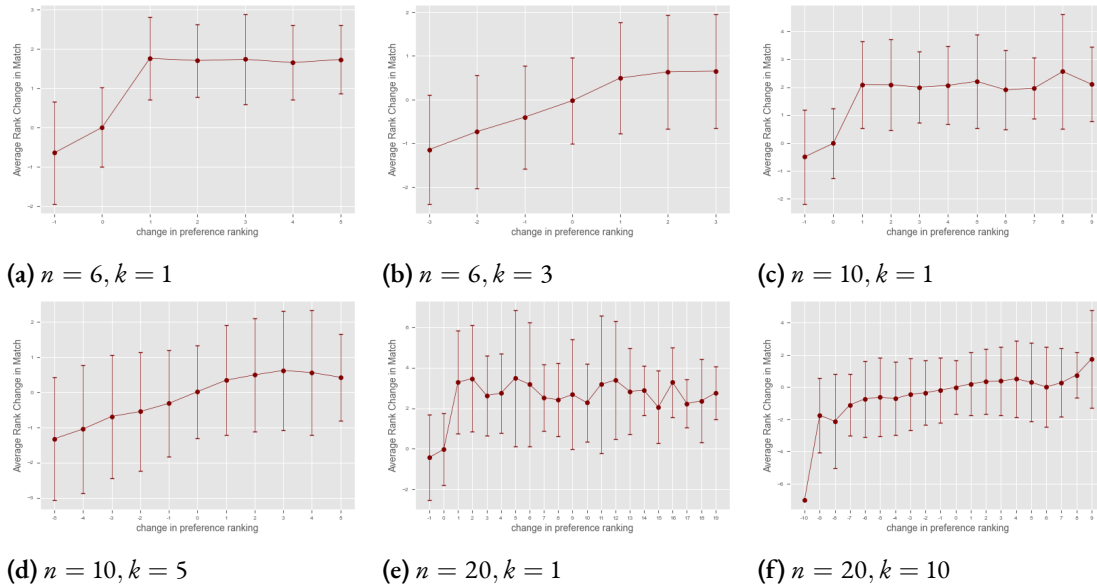
**Figure 5.17:** Average weak consistency (top) and strict consistency (bottom) rates across all possible values of  $k$  and several  $n$ . These rates are defined according to Definitions 9 and 10, respectively. We see generally high rates of weak consistency and low rates of strict consistency, showing that most students remain unchanged unless  $k$  nears its extremes ( $1$  and  $n - 1$ ).

minority students are weakly better off and majority students are weakly worse off.

However, the graphs of strict consistency show that strict consistency exponentially decreases for the minority class. When there is just 1 minority student, this student will almost always improve. This strengthens the result in Theorem 5, as not only is the single minority definitely weakly better off, but it is actually highly likely to be strictly better off. These results do not necessarily extend to other values of  $k$  – the strict drop-off in average strict consistency rates indicates that for most  $k$ , between 0 and 2 minority students will be improved on average. Combined with the high weak consistency rate, this shows that most minority students remain unchanged in their matches.

The trends for majority students are symmetric and opposite to those for minority students – the weak consistency rate is over 90%, so few majority students are made strictly better off. Strict consistency increases exponentially with  $k$  – most majority students remain unchanged, though for larger  $k$ , a higher proportion of minority students are made worse off. This means that when there is a high proportion of minority students in the overall student population, on average, only a couple students will be made better off, while most of the majority students will be made worse off. Intuitively, it makes sense that if there are many minority students, most majority students will be made worse off. However, the fact that only a few minority students are able to benefit at the cost of a larger proportion of majority students may be cause for concern.

Finally, we turn to graphs of average change in match welfare over change in preference ranking for several  $n$  over values of  $k = 1, \frac{n}{2}$ . In Figure 5.18, we can first focus on the change in welfare  $U(s_i)$  for the minority students  $s_i \in S^m$  (e.g. students moved up in  $\succ_{c_1}$ ). We see that for each value of  $n$ , when  $k = 1$ , no matter how much the minority student is improved, they have fairly consistently positive match welfare, while the worsened majority students do less well. The standard deviations show that majority students may improve in match welfare as well. This differs from the findings for arbitrary changes and single swaps (shown in Figures 5.5 and 5.10), in which the average improvement in match welfare seems to scale linearly with improvement in  $c_1$ .



**Figure 5.18:** Average change in student welfare for each possible change rank number, graphed for several different combinations of  $n$  and  $k$ . Averages generally show a less clear positive correlation than in the arbitrary and single-swap preference change cases.

When  $k = \frac{n}{2}$ , the graphs become smoother and thus more similar to those for the arbitrary and single-swap preference changes. Standard deviation actually decreases from the case where  $k = 1$ , but so does the average welfare change for minority students, such that within one standard deviation of the average, it is possible that a minority student whose rank was increased by any amount can be worse off in  $\mu'$ . Much like the arbitrary preference ranking change case, we can argue that there is a positive trend, but the high variance shows unilateral priority-based affirmative action is unpredictable.

What these overall trends demonstrate is that the value of  $k$  is relatively inconsequential towards the number of minority students improved: even if  $k$  is large, only a couple minority students seem to be strictly better off under  $\mu'$  on average, possibly at the expense of a potentially large proportion of majority students. We do see positive trends in utilitarian welfare difference, as well as high weak consistency rates for both classes. However, we also see trends similar to those for the arbitrary pref-

erence change, in which the average match welfare  $U(s_i)$  is positively correlated with rank change, but the standard deviation is high, likely because minority students remain unchanged with such high frequency. Additionally, with most minority students remaining unchanged on average, it seems that Theorem 5 may be the strongest guarantee of better performance that we can make for the minority group. Thus, it is difficult to recommend unilateral priority-based affirmative action as a policy when there is little assurance that most of the minority students will actually change, and it is possible that  $n - 1$  students can be worse off.

Overall, the only type of unilateral affirmative action change that seems to have a consistently positive impact on the students being improved in  $\succ'_{c_1}$  is the single-swap preference change or the priority-based affirmative action case where  $k = 1$ . However, this is a highly manufactured case that would rarely occur in real life, and our results regarding arbitrary preference changes and priority-based affirmative action show that unilateral affirmative action is volatile and not necessarily consistent with the desired effects of affirmative action.

# 6

## Conclusion and Future Work

In recent years, many school districts using student-proposing DA as a school assignment mechanism have faced scrutiny surrounding the lack of representation in their student bodies. Particularly, magnet and elite exam schools have begun reconsidering their admissions criteria by removing their exams or boosting the rankings of minority students [6] [10]. These policies have caused controversy both over whether they are ethical and whether they really do produce the desired effect of improving minority students' outcomes. Though previous work has shown impossibilities in student-

proposing DA’s responsiveness to different types of affirmative action policies, this work reflects the changes made by individual elite schools in public school districts by examining the setting in which one school attempts to pioneer a unilateral affirmative action policy in a school choice problem  $G$  that uses student-proposing DA. In our analysis, we examine a special case of the school choice problem in which there are  $n$  students and  $n$  schools, assuming WLOG that school  $c_1$  changes its preferences.

We introduce three types of preference changes that can constitute a unilateral affirmative action change (arbitrary, single-swap, and priority-based affirmative action) in order to determine whether students’ outcomes are sensitive to the magnitude of the change. We also contribute a set of metrics for student welfare sensitivity as well as metrics for “consistency,” which extends the “minimal responsiveness” described in [23] and [12] to be more descriptive of the amount of change induced by unilateral affirmative action. Our theoretical results show that even with a single school, single-swap preference change, it is possible to achieve worst cases in utilitarian count and egalitarian welfare, which points to the potential volatility of unilateral affirmative action in student-proposing DA. We also find that when adhering to priority-based affirmative action, we can guarantee weak consistency for the highest-ranked, strictly better off minority student who prefers  $c_1$  to their old match in  $\mu$ .

Another contribution of this paper is to simulate the average, best, and worst behavior for various metrics under unilateral affirmative action. Though theoretical analysis can determine the type of behavior that is or is not guaranteed, the average and variance in welfare and consistency can be more informative for policy recommendations, as worst cases may be rare. Our simulations show that the type of preference change does impact the ability of  $\succ'_{c_1}$  to change the matching  $\mu'$ , as arbitrary preference changes result in the highest proportion of runs where  $\mu' \neq \mu$ , and single-swap preference changes result in the lowest. However, we also see that single swaps are the only type of change in which there is a strong enough positive correlation between rank change and average match welfare  $U(s_i)$  – both arbitrary and priority-based preference changes show a general upwards

trend but contain too much variance to conclude that the trend will hold in general. Additionally, though priority-based affirmative action tends to produce weakly consistent results for most students, only a couple minority students seem to improve on average. Thus, in a case like Boston's where there is a high proportion of minority students among all applicants, this implies that a low percentage of these students would actually receive better outcomes after priority-based affirmative action even if all students are actively uplifted in  $c_1$ 's new ranking [15]. As such, we see that the positive impact of unilateral priority-based affirmative action may not scale with increased numbers of improved minority students.

Overall, we see that average performance in every case except the single-swap preference change (which is a highly manufactured case that is likely improbable in real school choice problems) does not necessarily achieve the desired effect on average. Since both overall and minority-class performance can also achieve severe worst cases, unilateral affirmative action is difficult to recommend as a policy. This should not, however, be taken as a total indictment of applications of affirmative action to student-proposing DA or in the general school choice problem for several main reasons. First, additional analysis should be performed in the future to examine settings where there are many more students than schools (e.g.  $n \gg m$ ). Additionally, this work uses uniformly and randomly distributed preferences for both students and schools, but it may also be informative to perform the same simulations using correlated preferences for students, schools, or both to see what happens when preferences are more consistent (e.g. there is a group of universally desirable schools in a district). Since such correlations will only be representative if extracted from empirical studies of student reports under deferred acceptance, simulating this setting with a basis of real student and school preference data is a vital next step towards definitively concluding whether unilateral affirmative action (or any affirmative action policy at all) should or should not be implemented.

Beyond the potential limitations in the realism of our simulations, our work also only investigates the specific case of unilateral affirmative action. What policy-makers can take away from this

analysis is not necessarily that affirmative action for student-proposing DA is unadvisable in general, but rather that unilateral affirmative action may be too volatile to produce the desired effect. As such, although the introduction of affirmative action by a single school is often a well-intentioned response to a lack of diversity, this work demonstrates that better outcomes may be more probable if all schools achieve consensus on both the decision to implement an affirmative action policy and the exact criteria of such a policy. As shown in [12] and [19], strong affirmative action policies that give “full priority” to the minority or variations of deferred acceptance can lead to stronger guarantees of minority welfare in  $G'$ . Additionally, [12] and [18] propose alternatives to student-proposing DA that may temper some of the efficiency loss that can occur when implementing affirmative action. Though these methods may be more costly and difficult to implement, they show improved responsiveness to priority-based affirmative action. These policies can potentially serve as a more effective alternative to a unilateral preference change. Thus, our work demonstrates that individual schools hoping to increase minority enrollment in districts like Boston and Fairfax County should prioritize establishing robust and consistent criteria for affirmative action across the entire district over immediate implementation of an affirmative action policy. Although it has become customary for schools to look favorably upon the idea of boosting a minority student’s ranking, we conclude that hasty implementation of such a policy could lead to lukewarm results on average and unpredictably severe worst cases that should be avoided if possible.

## References

- [1] Overview of the 100 largest public elementary and secondary school districts in the united states.
- [2] ABDULKADIROĞLU, A., PATHAK, P. A., ROTH, A. E., AND SÖNMEZ, T. The boston public school match. *American Economic Review* 95, 2 (2005), 368–371.
- [3] ABDULKADIROĞLU, A., PATHAK, P. A., ROTH, A. E., AND SÖNMEZ, T. Changing the boston school choice mechanism, 2006.
- [4] ABDULKADIROĞLU, A., AND SÖNMEZ, T. School choice: A mechanism design approach. *American economic review* 93, 3 (2003), 729–747.
- [5] ANDREYEVA, E., AND PATRICK, C. Paying for priority in school choice: Capitalization effects of charter school admission zones. *Journal of Urban Economics* 100 (2017), 19–32.
- [6] BAILEY, M. Opening the doors to elite public schools, Mar 2021.
- [7] BALINSKI, M., AND SÖNMEZ, T. A tale of two mechanisms: student placement. *Journal of Economic theory* 84, 1 (1999), 73–94.

- [8] BARRY, E. Boston overhauls admissions to exclusive exam schools, Jul 2021.
- [9] CAHN, S. M. *The affirmative action debate*. Routledge, 2013.
- [10] CAMERA, L. Race divides elite new york city high schools - us news & world report, Mar 2019.
- [11] CHEN, Y., AND SÖNMEZ, T. School choice: an experimental study. *Journal of Economic theory* 127, 1 (2006), 202–231.
- [12] DOĞAN, B. Responsive affirmative action in school choice. *Journal of Economic Theory* 165 (2016), 69–105.
- [13] DUBINS, L. E., AND FREEDMAN, D. A. Machiavelli and the gale-shapley algorithm. *The American Mathematical Monthly* 88, 7 (1981), 485–494.
- [14] GALE, D., AND SHAPLEY, L. S. College admissions and the stability of marriage. *The American Mathematical Monthly* 69, 1 (1962), 9–15.
- [15] GOODMAN, J., AND RUCINSKI, M. Increasing diversity in boston’s exam schools. *Rappaport Institute for Greater Boston Policy Brief* (2018).
- [16] HAFALIR, I. E., YENMEZ, M. B., AND YILDIRIM, M. A. Effective affirmative action in school choice. *Theoretical Economics* 8, 2 (2013), 325–363.
- [17] HU, L., AND CHEN, Y. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference* (2018), pp. 1389–1398.
- [18] JIAO, Z., AND SHEN, Z. School choice with priority-based affirmative action: A responsive solution. *Journal of Mathematical Economics* 92 (2021), 1–9.

- [19] JIAO, Z., AND TIAN, G. Responsive affirmative action in school choice: A comparison study. *Economics Letters* 181 (2019), 140–145.
- [20] KAMADA, Y., AND KOJIMA, F. Stability and strategy-proofness for matching with constraints: A problem in the Japanese medical match and its solution. *American Economic Review* 102, 3 (2012), 366–370.
- [21] KANNAN, S., ROTH, A., AND ZIANI, J. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), pp. 240–248.
- [22] KESTEN, O. School choice with consent. *The Quarterly Journal of Economics* 125, 3 (2010), 1297–1348.
- [23] KOJIMA, F. School choice: Impossibilities for affirmative action. *Games and Economic Behavior* 75, 2 (2012), 685–693.
- [24] NATANSON, H. Fairfax county school system faces second lawsuit over changes to thomas jefferson admissions, Mar 2021.
- [25] OTERO, S., BARAHONA, N., AND DOBBIN, C. Affirmative action in centralized college admission systems: Evidence from Brazil. *Unpublished manuscript* (2021).
- [26] PROBOLUS-CEDRONI, K. Bright flight: Desegregating Boston’s elite public schools, 1960–2000. *Journal of Urban History* 48, 3 (2022), 657–677.
- [27] ROTH, A. E. The economics of matching: Stability and incentives. *Mathematics of operations research* 7, 4 (1982), 617–628.
- [28] SETHURAMAN, J., TEO, C.-P., AND QIAN, L. Many-to-one stable matching: geometry and fairness. *Mathematics of Operations Research* 31, 3 (2006), 581–596.

- [29] SONI, V. Morality vs. mandate: Affirmative action in employment. *Public Personnel Management* 28, 4 (1999), 577–594.
- [30] SOWELL, T. *Affirmative action around the world: An empirical study*. Yale University Press, 2004.

# A

## Appendix

### A.1 PROOFS

PROOF OF LEMMA 1: Observe that only the preferences of  $c_1$  change, so preferences stay the same between  $G$  and  $G'$  for all students and all other schools. Let  $\mu' \neq \mu$ . Then, consider all students  $s_i$  such that  $s_i \neq \mu(c_1), \mu'(c_1)$  and  $\mu(s_i) \neq \mu'(s_i)$ , and all schools  $c_i$  such that  $c_i \neq c_1$  and  $\mu(c_i) \neq \mu'(c_i)$ .

If  $s_i$  is better off under  $\mu'$ , then  $\mu'(s_i) \succ_{s_i} \mu(s_i)$ . Assume that  $\mu'(s_i)$  is better off. Then,  $s_i \succ_{\mu'(s_i)} \mu(\mu'(s_i))$ , and  $(s_i, \mu'(s_i))$  forms a blocking pair in  $\mu$ . Thus, if  $s_i$  is better off, then  $\mu'(s_i)$  is worse off. If  $c_i$  is better off under  $\mu'$ , then  $\mu'(c_i) \succ_{c_i} \mu(c_i)$ . Assume that  $\mu'(c_i)$  is better off. Then,  $c_i \succ_{\mu'(c_i)} \mu(\mu'(c_i))$ , and  $(\mu'(c_i), c_i)$  forms a blocking pair in  $\mu$ . Thus, if  $c_i$  is better off, then  $\mu'(c_i)$  is worse off.

If  $s_i$  is worse off under  $\mu'$ , then  $\mu(s_i) \succ_{s_i} \mu'(s_i)$ . Assume that  $\mu(s_i)$  is worse off. Then,  $s_i \succ_{\mu(s_i)} \mu'(\mu(s_i))$ , and  $(s_i, \mu(s_i))$  forms a blocking pair in  $\mu'$ . Thus, if  $s_i$  is worse off, then  $\mu(s_i)$  is better off. If  $c_i$  is worse off under  $\mu'$ , then  $\mu(c_i) \succ_{c_i} \mu'(c_i)$ . Assume that  $\mu(c_i)$  is worse off. Then,  $c_i \succ_{\mu(c_i)} \mu'(\mu(c_i))$ , and  $(\mu(c_i), c_i)$  forms a blocking pair in  $\mu'$ . Thus, if  $c_i$  is worse off, then  $\mu(c_i)$  is better off.

□

**PROOF OF THEOREM 2:** First, if all students are weakly better off under  $\mu' \neq \mu$ , then  $\mu(c_1)$  and  $\mu'(c_1)$  must both be better off because both of their matches must change under  $\mu'$ .

Now, we must show that if  $\mu(c_1)$  is better off, then all other students are as least as well off. We know that for  $\mu(c_1)$  to be better off and for  $(\mu(c_1), c_i)$  to not form a blocking pair in  $\mu$ ,  $c_i$  must be worse off in  $\mu'$ :

$$c_i \succ_{\mu(c_1)} c_1$$

$$\mu(c_i) \succ_{c_i} \mu(c_1)$$

Since  $\mu(c_1)$ 's match changes under  $\mu'$ , then  $\mu(c_i)$  is now matched to a new school,  $c_a$  in  $\mu'$ . By Lemma 1, we know that because  $c_i$  is worse off under  $\mu'$ ,  $\mu(c_i)$  is better off. Lemma 1 also states that because  $\mu(c_i)$  is better off in  $\mu'$ ,  $c_a$  must then be worse off. Then,  $\mu(c_a)$  must be matched to a new school  $c_b$ . Because  $c_a$  is worse off in  $\mu'$ ,  $\mu(c_a)$  must be better off under  $\mu'$ . This chain of students being necessarily better off under  $\mu'$  continues until  $\mu'(c_1)$  is matched to  $c_1$ .

Since  $c_1$ 's preference ranking changes between  $\mu$  and  $\mu'$ , we cannot use the blocking pair of  $(\mu'(c_1), c_1)$  in  $\mu$  to reason about  $\mu'(c_1)$ 's status – it is possible in the new matching instance  $G'$  for  $\mu'(c_1)$  and  $c_1$  to prefer each other to their matches in  $G$ . Let us assume that  $\mu'(c_1)$  is worse off. Then, its old match  $c_j$  must be better off under  $\mu'$  so that  $(\mu'(c_1), c_j)$  does not form a blocking pair in  $\mu'$ . This implies that  $\mu'(c_j)$  is worse off. However, it was established above that any student besides  $\mu'(c_1)$  that was changed must be better off, so we reach a contradiction. Thus,  $\mu'(c_1)$  must be better off, meaning that all changed students must be better off.

This demonstrates that  $\mu(c_1)$  being better off in  $\mu'$  implies that  $\mu'(c_1)$  is also better off – and in fact, that every changed student is better off.

□

**PROOF OF THEOREM 3:** First, if all students are weakly worse off and  $\mu' \neq \mu$ , then since  $\mu'(c_1)$  and  $\mu(c_1)$ 's match changed in  $\mu'$ , they must both be worse off.

Once again, we let  $\mu(\mu(c_1)) = c_i$ , and let  $\mu(\mu'(c_1)) = c_j$ . We then show that if  $\mu'(c_1)$  is worse off, all students are weakly worse off (e.g. no student is better off).

We know that for  $\mu'(c_1)$  to be worse off and for  $(\mu'(c_1), c_j)$  to not form a blocking pair in  $\mu'$ ,  $c_j$  must be better off in  $\mu'$ :

$$c_j \succ_{\mu'(c_1)} c_1$$

$$\mu'(c_j) \succ_{c_j} \mu'(c_1)$$

By Lemma 1, we know that because  $c_j$  is better off,  $\mu'(c_j)$  is worse off under  $\mu'$ . Let  $\mu(\mu'(c_j)) = c_a$ . Lemma 1 also states that if  $\mu'(c_j)$  is worse off, then  $c_a$  is better off. Thus,  $c_a$  is matched to a new student  $\mu'(c_a)$ , who must again be worse off, according to Lemma 1. This leads  $c_b = \mu(\mu'(c_a))$  to be better off, which implies that  $\mu'(c_b)$  is worse off. This chain of students being worse off under  $\mu'$

continues until  $\mu(c_1)$  is matched to  $c_i$ . This closes the loop because  $\mu(c_1)$ 's former match,  $c_1$ , has been matched to  $\mu'(c_1)$ .

We now reason about  $\mu(c_1)$ . Assume that  $\mu(c_1)$  is better off, so  $c_i \succ_{\mu(c_1)} c_1$ .  $c_i$  must then be worse off in order to prevent  $(\mu(c_1), c_i)$  from being a blocking pair in  $\mu$ . If  $c_i$  is worse off, then according to Lemma 1,  $\mu(c_i)$  must be better off. However, we have established that any changed student other than  $\mu(c_1)$  must be worse off. Thus, we have a contradiction, and  $\mu(c_1)$  must be worse off.

This demonstrates that  $\mu'(c_1)$  being worse off in  $\mu'$  implies that  $\mu(c_1)$  is also worse off – and in fact, that every changed student is worse off.

□

**PROOF OF THEOREM 4:**\* Let  $s_j$  be the student that  $s_i$  was swapped with (e.g. in  $c_1 s_j \succ_{c_1} s_i$  and in  $\succ'_{c_1}, s_i \succ'_{c_1} s_j$ ). First, we consider that the match only changes if the set of proposals being made changes. Since only  $c_1$  changes its preferences, the proposals change only if a student who was previously rejected by  $c_1$  is now accepted under  $c_1$ '.

In this case, since only one student is changed in  $c_1$ 's ranking, this implies that  $s_j$  is, at least temporarily, now accepted by  $c_1$  while  $s_i$  is rejected. This also means that it must be true that  $c_1 \succ_{s_i} \mu(s_i)$ . If this were not the case, then the proposals would not change, as  $s_i$  would propose to all the exact same schools as in  $G$  (as would all other students), which would lead to the same set of acceptances.

**Case 1:**  $\mu(c_1) = s_j$ . First, we consider the case where  $s_j$  was actually  $c_1$ 's previous match. Then,  $s_j$  must propose to schools it has not yet proposed to and is worse off. Let  $\mu'(s_j) = c_a$ . In order for this proposal to be accepted by  $c_a$ , it must be true that  $c_a$  is made better off or that  $c_a$ 's previous match was  $s_i$ . If the latter is true, then  $s_i$  and  $s_j$  simply switch. No other proposals change as a result of this. If the former is true, then  $\mu(c_a)$  must now propose to schools that it also has not yet proposed to,

---

\*This is a supplementary proof that shows minimal responsiveness for the specific case we examine (a unilateral, single-swap preference change). The result is already implied by [7].

making it worse off than before. This means that the new school it proposes to,  $c_b$ , must be made better off, or that  $c_b$ 's previous match was  $s_i$ .

This sequence of students being made worse off by being "displaced" from their previous matches by new proposals continues until one of two things occur: either a displaced student  $s_m$  proposes to and is accepted by  $\mu(s_i)$ , thus ending this sequence with  $s_i$  better off with  $s_1$ , or a displaced student  $s_m$  (where  $s_m \succ_{c_1} s_i$ ) proposing to  $c_1$  and being accepted over  $s_i$ . In this case,  $s_i$  becomes the displaced student and can once again propose to  $\mu(s_i)$  and be accepted, as  $\mu(s_i)$  cannot have been taken by any other displaced student because, as mentioned above, the sequence would have ended if another displaced student had taken  $\mu(s_i)$ . Thus,  $s_i$  is at least as well off as before or better off through being accepted by  $c_1$ .

**Case 2:**  $\mu(c_1) \neq s_j$ . If  $s_j$  was not  $c_1$ 's previous match, then there exists some student  $s_k = \mu(c_1)$ . We consider the possible orderings of  $s_k, s_i, s_j$  in the original matching problem  $G$ .

First, we consider  $s_j \succ_{c_1} s_i \succ_{c_1} s_k$ . In this case, it is not possible for the matching to change due to a switch between  $s_i$  and  $s_j$ . We know this is true because if  $s_k = \mu(c_1)$ , then  $s_i$  and  $s_j$  must not have proposed to  $c_1$  in  $G$ , or  $\mu(s_i) \succ_{s_i} c_1$  and  $\mu(s_j) \succ_{s_j} c_1$ . However, we stated above that the match only changes if  $c_1 \succ_{s_i} \mu(s_i)$ . Thus, the matching does not change in this case due to the swap of  $s_i$  and  $s_j$ .

We then consider the case when  $s_k \succ_{c_1} s_j \succ_{c_1} s_i$ . If the matching does change in this case, then again, it must occur because  $c_1$  accepts  $s_i$  over  $s_j$  at some point. In order for  $s_k$  to no longer be matched to  $c_1$  in  $\mu'$  and for  $(s_k, c_1)$  to not form a blocking pair in  $\mu'$ , it must be that  $s_i$ 's acceptance by  $c_1$  allows  $s_k$  to propose to and be accepted by some school  $c_j \succ_{s_k} c_1$  where  $s_i \succ_{c_j} s_k$ . This means that  $c_1$  is necessarily "worse off" (as we defined in Corollary 2.1) because  $s_k$  no longer proposes to it. If  $c_1$  is "worse off," then Corollary 2.1 states that no student is worse off and at least one student is better off. Thus,  $s_i$  must also be weakly better off in this case as well.

□

PROOF OF THEOREM 5: First, we consider possible constraints on  $\mu'(c_1)$  under this change. We know that if  $\mu'(c_1) = s_k$  where  $s_i \succ'_{c_1} s_k$ , then it must be that  $s_i$  is now matched to  $\mu'(s_i)$  where  $\mu'(s_i) \succ_{s_i} c_1$  in order for  $(s_i, c_1)$  to not form a blocking pair in the new matching. Thus, in this case,  $s_i$  is better off, as  $\mu'(s_i) \succ_{s_i} c_1 \succ_{s_i} \mu(s_i)$ .

We also know that if  $\mu'(c_1) = s_i$ , then  $s_i$  is better off because  $c_1 \succ_{s_i} \mu(s_i)$ .

Then, we consider the case where  $\mu'(c_1)$  is  $s_k$ , where  $s_k \succ'_{c_1} s_i$ . Here,  $s_k$  may either be a majority student or a minority student that prefers  $\mu(s_k) \succ_{s_k} c_1$ , and  $s_k \succ_{c_1} s_i$  as well – if  $s_k$  is a majority student, then it must have already been ranked above  $s_i$ , and if  $s_k$  is a minority student, since priority-based affirmative action preserves the ordering of students within their respective groups, then  $s_k$  was also already ranked above  $s_i$ . This means that  $s_k$  must not have proposed to the  $c_1$  in  $G$  but now does.

We also observe that the proposals being made in  $G'$  only differ from those in  $G$  because  $s_i$  is, at some point, accepted by  $c_1$  where it was not before. Even if some other minority student  $s_m$  where  $s_i \succ_{c_1} s_m$  is newly accepted by  $c_1$  at some point, since  $s_i$  definitely proposes to  $c_1$  in both  $G$  and  $G'$ ,  $s_m$  ends up being rejected in both  $G$  and  $G'$ .

Now, we examine the relative ordering of  $\mu(c_1)$  and  $s_i$ . We know that it cannot be true that  $s_i \succ_{c_1} \mu(c_1)$  in  $G$ , as  $(s_i, c_1)$  would have then formed a blocking pair for  $\mu$ . Thus, we assume that  $\mu(c_1) \succ_{c_1} s_i$ .

If  $\mu(c_1) \succ_{c_1} s_i$  and in the new ranking,  $s_i \succ'_{c_1} \mu(c_1)$ , then we can say that  $s_i$  is at least temporarily accepted by  $c_1$ , displacing  $\mu(c_1)$ .  $\mu(c_1)$  must then propose to new schools that it did not propose to previously.

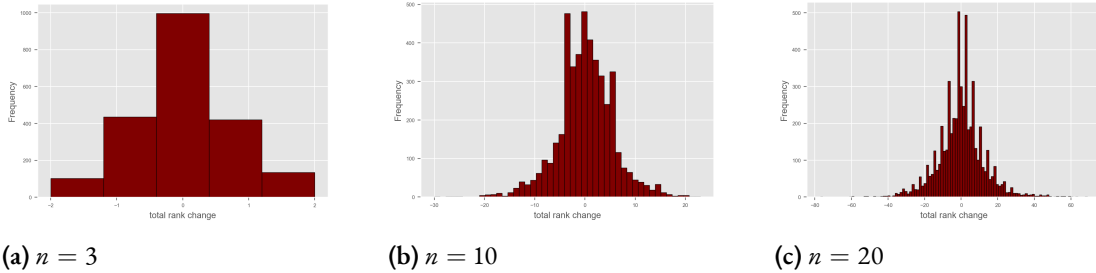
Let  $\mu(c_1)$  be accepted by some school  $\mu(s_a)$ , which was previously matched to student  $s_a$ . We know that this school cannot be  $\mu(s_i)$ , as doing so would end the sequence of new proposals with  $\mu'(s_i) = c_1$ , and we know that  $s_k \succ'_{c_1} s_i$  is  $c_1$ 's new match in this case. This leads  $s_a$  to propose to schools it did not propose to in  $G$ . Let the school it is now accepted by be  $\mu(s_b)$ , where  $\mu(s_b) \neq \mu(s_i)$

for the same reasons mentioned above. Then,  $s_b$  is displaced and must also now propose to schools it did not propose to in  $G$ . At some point in this sequence of displaced students who propose to new schools,  $s_k$  must be displaced and propose to  $c_1$  successfully. This means that  $s_i$  must now propose to schools it has ranked below  $c_1$ . As mentioned above, since no displaced students made new proposals to  $\mu(s_i)$ ,  $\mu(s_i)$  must still be available for  $s_i$  to propose to. Thus, even if  $\mu(c_1) = s_k$  where  $s_k \succ'_{c_1} s_i$ ,  $s_i$  can still be at least as well off as before by proposing to its old match  $\mu(s_i)$ .

We can also consider the case where  $\mu(c_1) \succ_{c_1} s_i$  in the old ranking and  $\mu(c_1) \succ'_{c_1} s_i$  as well in the new ranking. The only way that a change in proposals can occur is if  $s_i$ , through being ranked above some other student  $s_a$  in  $\succ'_{c_1}$ , now no longer need propose to one of the schools  $c_j$  that it ranked below  $c_1$  and above  $\mu(s_i)$ . If one such  $c_j$  is preferred by  $\mu(c_1)$  to  $c_1$ , then  $\mu(c_1)$  can now propose successfully to this school. As stated in Theorem 2,  $\mu(c_1)$  being better off ensures that every student is weakly better off, so  $s_i$  must remain better off.

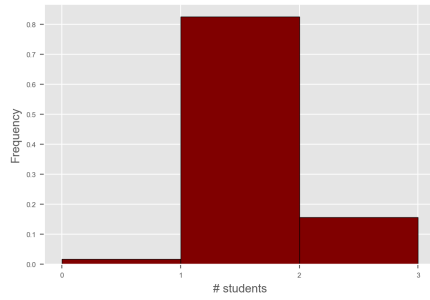
□

## A.2 ADDITIONAL RESULTS FOR ARBITRARY PREFERENCE CHANGE

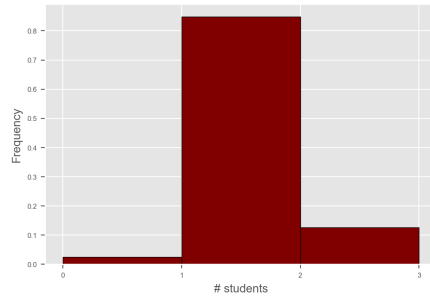


**Figure A.1:** The distribution of utilitarian welfare for several representative values of  $n$ . We see a largely Gaussian trend in the distributions, with a slight left skew as  $n$  increases.

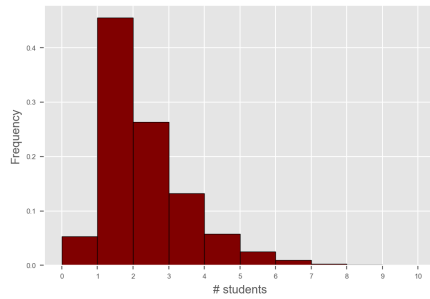
Figure A.2 shows that the proportion of runs with  $k = [0, n - 1]$  improved or worsened students, grouped by the value  $k$ . We see that the proportion of runs with  $k$  improved students is ap-



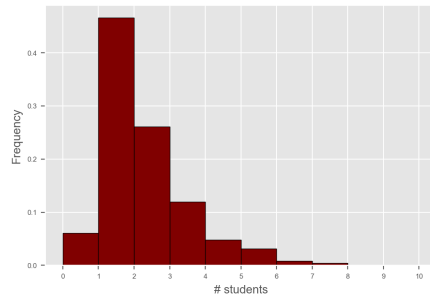
(a) Improved Students,  $n = 3$



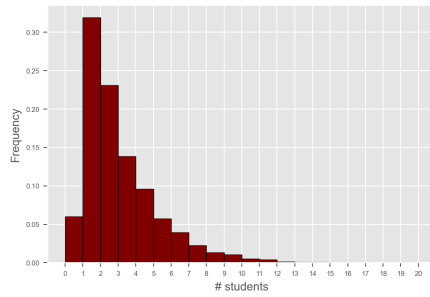
(b) Worsened Students,  $n = 3$



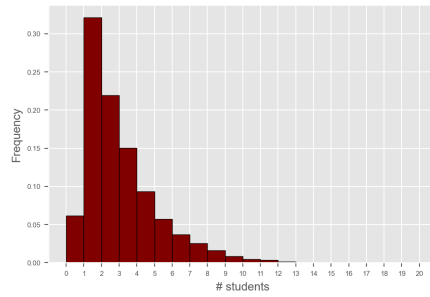
(c) Improved Students,  $n = 10$



(d) Worsened Students,  $n = 10$



(e) Improved Students,  $n = 20$

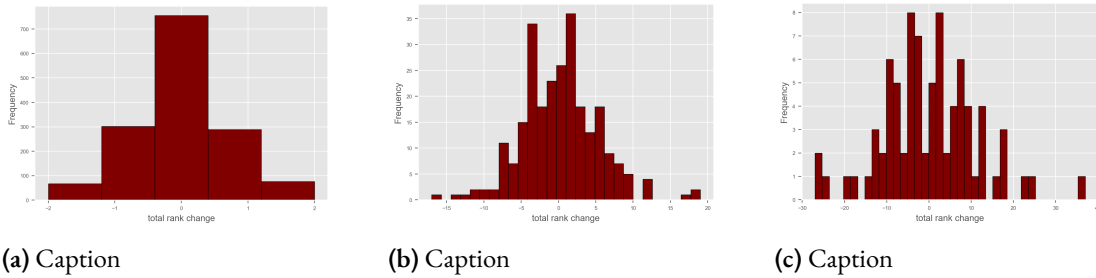


(f) Worsened Students,  $n = 20$

**Figure A.2:** Distributions of improved and worsened students across 10000 iterations for an arbitrary preference change. Note that the distributions seem largely symmetric between improved and worsened students.

proximately the same as the proportion of runs with  $k$  worsened students, and that the proportions across  $k$  have the same right-skewed distribution, with a peak at  $k = 1$ . This supports the symmetry in the graphs of utilitarian welfare in Figure A.1.

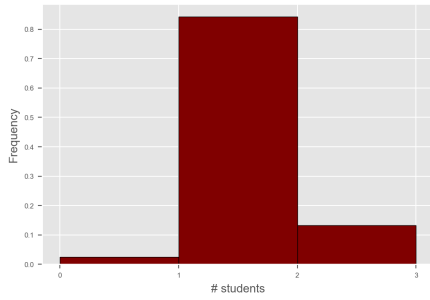
### A.3 ADDITIONAL RESULTS FOR SINGLE-SWAP PREFERENCE CHANGE



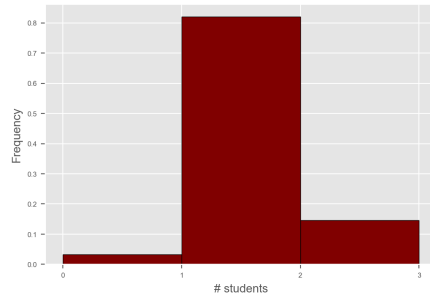
**Figure A.3:** The distribution of utilitarian welfare for several representative values of  $n$ . We see a somewhat Gaussian trend in the distributions, with a slight right skew as  $n$  increases.

Figure A.4 shows the proportion of runs with  $k = [0, n - 1]$  improved or worsened students, grouped by the value  $k$ . We see that for  $n = 3, 10$  the proportion of runs with  $k$  improved students is approximately the same as the proportion of runs with  $k$  worsened students, and that the proportions across  $k$  have the same right-skewed distribution, with a peak at  $k = 1$ . At  $n = 20$ , we see a different trend – the percentage of runs with  $k$  improved students peaks at  $k = 2$  rather than 1, while the peak remains at  $k = 1$  for worsened students. This supports the right skew in the graphs of utilitarian welfare in Figure A.3, which become more apparent as  $n$  increases.

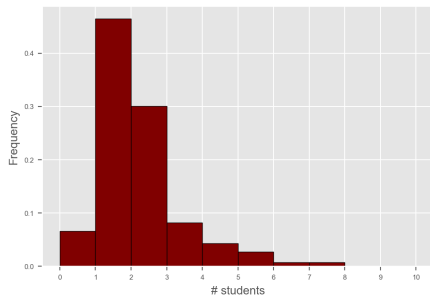
### A.4 ADDITIONAL RESULTS FOR PRIORITY-BASED AFFIRMATIVE ACTION



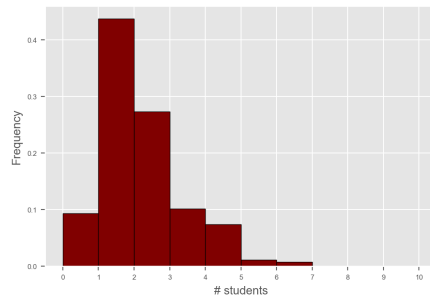
(a) Worsened



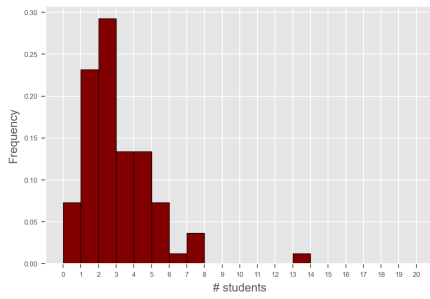
(b) Worsened



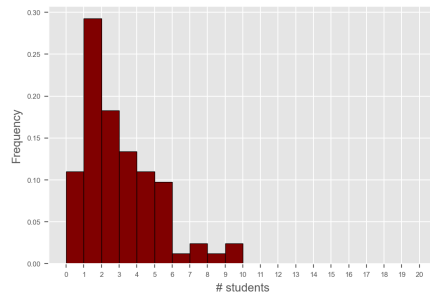
(c) Improved



(d) Worsened

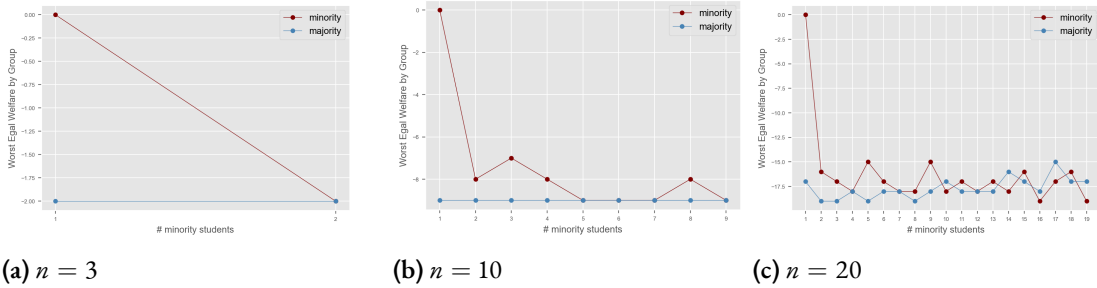


(e) Improved

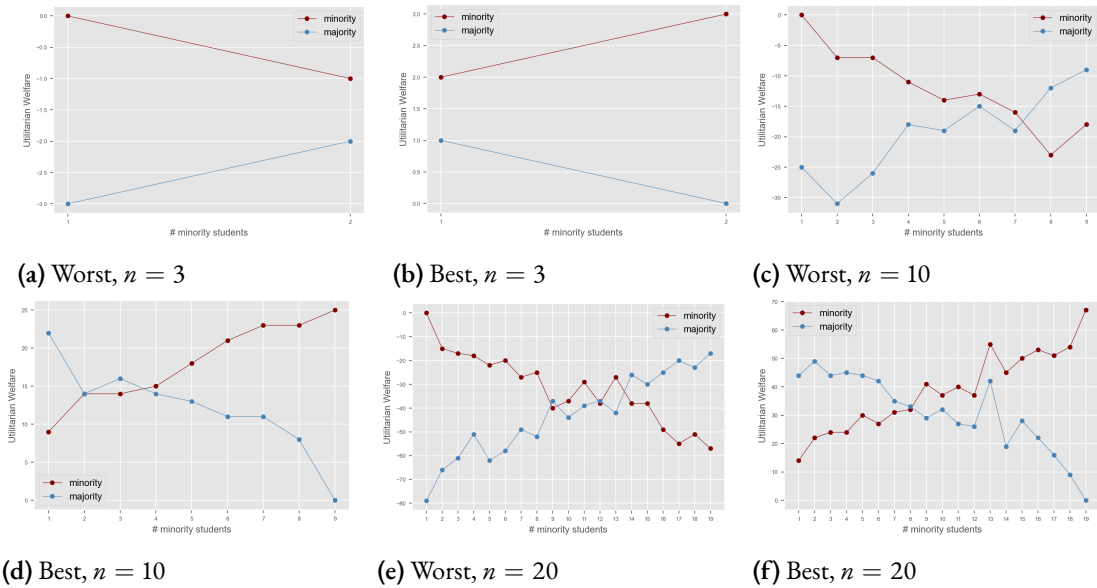


(f) Worsened

**Figure A.4:** Distributions of improved and worsened students across 10000 iterations for a single-swap preference change. Note that the distributions seem largely symmetric between improved and worsened students, though the trend is not as clear.



**Figure A.5:** Worst egalitarian welfare for minority and majority groups from 10000 iterations and graphed for several values of  $n$  and all possible  $k$ .



**Figure A.6:** Worst and best utilitarian welfare by class. We see that the welfare of the two classes have opposite trends.