Designing Affirmative Action Policies under Uncertainty

Corinna Hertweck

Helsinki May 8, 2020

UNIVERSITY OF HELSINKI
Master's Programme in Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Studieprogram — Study Programme
Faculty of Science	Master’s Programme in Computer Science

Tekijä — Författare — Author

Corinna Hertweck

Työn nimi — Arbetets titel — Title

Designing Affirmative Action Policies under Uncertainty

Ohjaajat — Handledare — Supervisors

Michael Mathioudakis and Antti Ukkonen

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master’s thesis	May 8, 2020	58 pages

Tiivistelmä — Referat — Abstract

In this work, we seek robust methods for designing affirmative action policies for university admissions. Specifically, we study university admissions under a real centralized system that uses grades and standardized test scores to match applicants to university programs. For the purposes of affirmative action, we consider policies that assign bonus points to applicants from underrepresented groups with the goal of preventing large gaps in admission rates across groups, while ensuring that the admitted students are for the most part those with the highest scores. Since such policies have to be announced before the start of the application period, there is uncertainty about which students will apply to which programs. This poses a difficult challenge for policy-makers. Hence, we introduce a strategy to design policies for the upcoming round of applications that can either address a single or multiple demographic groups. Our strategy is based on application data from previous years and a predictive model trained on this data. By comparing this predictive strategy to simpler strategies based only on application data from, e.g., the previous year, we show that the predictive strategy is generally more conservative in its policy suggestions. As a result, policies suggested by the predictive strategy lead to more robust effects and fewer cases where the gap in admission rates is inadvertently increased through the suggested policy intervention. Our findings imply that universities can employ predictive methods to increase the reliability of the effects expected from the implementation of an affirmative action policy.

ACM Computing Classification System (CCS):
Social and professional topics → Governmental regulations
Computing methodologies → Machine learning
Applied computing → Law

Avainsanat — Nyckelord — Keywords

algorithmic fairness, affirmative action, university admissions

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Thesis for the Algorithms study track

Acknowledgements

This research would not have been possible without my supervisor, Michael Mathioudakis. Thank you for your ongoing support and guidance and for allowing me to work on a topic that I am deeply passionate about.

I owe many thanks to Carlos Castillo. Thank you for providing the university admission dataset and for your collaboration on this project - your expertise has helped define the path of this work.

My sincere thanks also go to my second supervisor, Antti Ukkonen. Thank you for your valuable feedback in the research and writing process.

I am deeply grateful to my family, especially to my parents, for their support not only in the past two years, but throughout my entire life. Thank you for always having my back and for reminding me that there is a life outside of university and work.

My heartfelt thanks go to my friends Adelina, Aidan, Christina, Eoin, Puru and Sara who have accompanied me in my studies for the last two years. Thank you for all your advice and encouragement, for the shared food and all the fun.

Contents

1 Introduction

In an analysis of university admissions data, we find that one of the most prestigious engineering programs in Chile frequently admits female applicants at a lower rate than male applicants. This raises the question of what the reason for this disparity is. If you are not familiar with the Chilean university admission system, you might be inclined to think that university administrators are less likely to admit female applicants because they hold explicit or implicit biases [42] against women in engineering. Such biases have indeed been shown to play a role in hiring (e.g., [15, 81]) and university admission processes where each application is reviewed individually (e.g., [24]). An easy solution to this problem then appears to be the automation of the admission process based solely on grades or standardized test scores – after all, this appears to leave no room for human biases.

However, this is exactly how the Chilean admission system is already working. In their *central admission system*, students apply to a central institution that matches applicants with university programs. Such centralized admission systems are used in several countries worldwide, such as Turkey (see, e.g., [5, 87]), Germany (see, e.g., [86]), China (see, e.g., [27]), and Ghana (see, e.g., [4]). The matching of students to programs is typically based on the students' grades and standardized test scores [44]. This raises the question of how this admission process can still result in such unequal admission rates. A large body of research (e.g., [6, 65, 73, 76]) points to differences in grades and standardized test scores as the source of inequality. These scores are – for various reasons discussed in Section 2 – correlated with demographic markers, such as gender, race and income, which can then lead to unequal admission rates. As shown in [64], affirmative action policies can be applied to reduce inequalities in such settings. Their objective is the increase of the representation of groups that have historically been underrepresented.

Achieving this goal is challenging for a number of reasons. First, we assume that universities still want to reward merit and admit the best applicants – an objective that may contradict the one of increasing representation of specific groups. A trade-off between the two objectives has to be carefully considered. Second, the implementation of affirmative action policies should be transparent. Policies are therefore typically announced before the start of the application period, and so they must be designed under uncertainty about which students will apply to each program. And third, the search for optimal and robust policies quickly becomes infeasible for human policy-makers as their limited computational capacities pose a constraint on their ability to make rational choices [79].

The question we want to answer with this thesis is how computational methods can be leveraged to evaluate a wider range of alternatives and therefore to identify more robust policies. We develop an approach for designing robust and effective affirmative action policies in the application scenario of university admissions. In particular, we design *bonus policies*, i.e., policies that assign a number of bonus points to applicants from disadvantaged backgrounds. These policies do not alter the admission priority of applicants within each group and have equivalent effects to setting admission quotas [64]. The technical problem we face is to choose the right number of bonus points, so that the policy will have a robust beneficial effect on the admission rate of the given group.

Our approach is based on the simulation of the application process as this will allow us to compare the effects of different policies. Since we are unable to measure the effect of a given bonus policy on next year's application set (consisting of the applicants with their applications), we could instead estimate its effect by evaluating the policy on a large number of historical application sets. As historical application data are, however, limited, we generate our own application sets. To this end, we first sample sets of students based on application data from previous years. Even though these students' original applications are known, they cannot be used for our purposes as some of the programs that students have applied to in the past no longer exist. A predictive model trained on the most recent data is thus employed to generate applications for the sampled students. We assume that a policy that is on average beneficial to the resulting application sets will also reduce the gap in admission rates in the upcoming application round. We develop this predictive approach as an applied data analysis case over a large real dataset of university admissions in Chile and demonstrate its performance.

The data analysis for this study is run on the University of Helsinki's Ukko2 cluster (urn:nbn:fi:research-infras-2016072533).

An earlier version of this work has been submitted to the Applied Data Science track of ECML PKDD 2020 and is under review at the time of writing this thesis. The thesis expands on the submitted paper in that it not only examines the design of policies for a single demographic group but for multiple ones. Additionally, it includes details that had to be omitted from the paper due to space limitations.

1.1 Research Gap

This thesis belongs to the field of algorithmic fairness. In this field, evaluating the fairness of a given algorithm is common [71]. If biases are detected, the algorithm is oftentimes adapted to not exhibit these biases anymore as seen in, e.g., [22, 55, 88]. The algorithm that could be adapted in this work is the algorithm that assigns students to university programs. One possible adaption would be changing its input variables. As previously mentioned, admission decisions are only based on grades and standardized test scores. While these features are correlated with demographic markers, we assume that they should remain the input variables as they are the standard features for admission decisions. Another possible adaption would be the change of the algorithm that assigns students to university programs. As there are many possibilities to match students with university programs, even if the admission priority is given by students' scores, the algorithm could simply produce another matching (see, e.g., [2, 78]). There is already a large body of research (see Section 2) into how the algorithm can incorporate affirmative action policies, such as the reservation of spots for female students, in order to increase the representation of disadvantaged groups. However, there is little research into how the parameters of these policies, such as the number of spots reserved for women, should be chosen. This is what our research is aimed at.

Hence, instead of addressing inequalities through changes in the computational aspects of the admissions process directly, this thesis explores how computational methods can be applied to the policy-making process. The policies then address the inequalities in the admissions process.

1.2 Thesis Contributions

To our knowledge, this is the first study that discusses how historical data can be leveraged in the design of affirmative action policies for university admissions when the applicants are unknown. An advantage of our approach is that it incorporates the knowledge of domain experts, i.e., university administrators. This is achieved by allowing them to make conscious decisions about the trade-off between merit and social inclusion, i.e., the potential decrease in the scores of admitted students they are willing to accept for increased social inclusion. Central university admission systems can adopt the presented approach to design robust policies based on existing data.

1.3 Structure

We begin in Section 2 by situating this work in the broader fields of policy-making and algorithmic fairness. In particular, we highlight a gap in the literature with respect to how computational methods are oftentimes already utilized in policy-making, but not in the design of affirmative action policies. Section 3 explores the Chilean university admissions data with a focus on differences between demographic groups and discusses the (un)predictability of admission rates. Section 4 discusses explores ways to model students' application behavior. The model decided on based on experimental results is employed to generate application sets on which the effects of different bonus policies are evaluated. Section 5 describes in detail how policy suggestions are made based on the generated application sets and experimentally compares our predictive approach to simpler strategies. Section 6 presents a method for efficiently finding close-to-ideal combinations of bonus policies when policies for multiple demographic groups are to be implemented. To achieve this goal, it extends the strategies introduced in Section 5. Finally, Section 7 discusses possibilities for future work as well as the implications of our findings for the design of affirmative action policies.

2 Background and Related Work

This section examines literature at the intersection of the policy-making process and algorithmic fairness. We begin with an introduction to affirmative action policies and their current usage in Chile. We then highlight how computational methods can be an aid in the policy-making process, but find a lack of research into how this can be applied to the context of affirmative action policies. Since "fairness" has no single definition, we conclude with a discussion of "fair" university admissions and justify how we define fairness in the context of this work.

2.1 Affirmative Action Policies in Chile

Affirmative action policies aim at increasing the representation of disadvantaged groups by treating them favorably, for example, in university admissions or hiring. According to [60], Hobart Taylor, Jr., first introduced the term in Execute Order 10925 [72] in 1961. Since then affirmative action policies have been implemented and debated in many countries, such as India (see, e.g., [16, 59]), Brazil (see, e.g., [50]) and South Africa (see, e.g., [63]).

Measures like this are necessary in Chile as the inequality in the education system has persisted for years [21]. While higher education is no longer regarded as a privilege of the elites, multiple issues prevent equal access to higher education for all students [30]. In 2009, the Chilean Ministry of Education found that a majority of the students that choose vocational training for their upper secondary education are from a disadvantaged socio-economic background [68]. Consequently, an OECD report [69] found that students from low-income households are underrepresented at universities.

Chile's university admission system is already implementing affirmative action policies in order to reduce inequalities. In an effort to make grades more comparable across schools, for example, a transformed version of the high school grade score average has been introduced. The transformation compares each student's grades to students who have studied at similar schools in the past years. This measure has been shown to help better judge students' academic abilities [67]. Moreover, it can be viewed as an affirmative action policy as it tends to increase the grade averages of students from disadvantaged backgrounds [31].

Additionally, one of the most prestigious universities in the country implements its own affirmative action policies. Through these policies the first women on the waiting list of certain science and engineering programs, such as industrial civil engineering, and the first men on the waiting list of the social work program are automatically admitted [84]. Bastarrica et al. [10] evaluated the effect of this policy and found not only a positive effect on the number of admitted women but also on the number of applications received from women.

Similar studies have been conducted internationally (see, e.g., [25, 33, 48, 82] for studies on affirmative action policies in the United States). Typically, they study the effect of affirmative action policies on minority students by observing the outcomes of the implementation or ban of such policies. Other studies, such as [7], examined the effects of affirmative action policies in a laboratory setting. Specifically, Balafoutas and Sutter compared four different types of affirmative action policies, e.g., giving bonus points to the disadvantaged group or preferring the member of the disadvantaged group if two competitors are equal in terms of merit. The mentioned studies have found that affirmative action policies are beneficial for the representation of disadvantaged groups.

While the existing literature reveals differences between the various types of affirmative action policies, it lacks a discussion of how the numerical parameters in these policies affect the outcomes. Instead, numerical parameters are oftentimes seen as a given.

2.2 Matching Algorithms

In centralized admission systems, algorithms are typically deployed to handle the admissions process. Such algorithms are well-known in game theory where they are referred to as *matching algorithms*. The case of matching a set of students with a set of programs is a many-to-one, two-sided matching problem [1].

This matching problem was first described in [38] in 1962 as the *college admissions problem*. To match students with programs, Gale and Shapley proposed the Deferred Acceptance (DA) algorithm which they proved to have several desirable qualities. One of them, called *stability*, is particularly important in the case of Chile. In a stable assignment, there is no student-program-pair (s, p) where s prefers p over their current assignment and p prefers s over any of the students it admitted. As noted in [75], the admission results in Chile are publicly visible. Unstable matchings could therefore lead to lawsuits if students realize that someone with a lower score has been admitted to a program that they would have preferred over their own. Although the matching algorithm employed in Chile is not publicly available, it can thus be assumed to be a version of the DA algorithm. This claim is supported by the experimental results in [75].

Matching algorithms, such as DA, can be implemented to incorporate affirmative action policies. One possible affirmative action policy is that of a minority-reserve that represents a lower bound on the number of minority students. [56] evaluates this strategy for DA. [1, 2, 43, 58] demonstrate further affirmative action extensions of matching algorithms. However, as previously observed for the field of affirmative action policies, how the parameters of such policies should be set is rarely considered.

2.3 Evidence-Based Policy Design

With the ongoing digitalization of many aspects of human life, there is an increasing demand for data being used in the design of public policies. Computational approaches to policy-making can be seen as part of the field of *evidence-based policy design* that grounds policy-making in data. Desouza and Lin [32] argue that computational modeling methods enable the exploration of more scenarios which helps to reduce uncertainty about the potential effects of a policy. This way, computational methods create more robust policies. Such methods have, for example, been applied to the design of policies in the fields of public health [9, 29] or counter-terrorism [57, 66]. In the area of university admissions, recent work [64] addressed the problem of policy design for affirmative action. Contrary to our work, it designed policies for a given set of applications and not under uncertainty.

2.4 Notions of Fairness for Affirmative Action

As Abebe et al. [3] note, computing can bring attention to historical inequalities by formalizing and quantifying them. The field of algorithmic fairness has thus brought forth a variety of metrics to measure fairness between socio-demographic groups which are distinguished by a *sensitive attribute*, e.g., gender. The literature commonly focuses on the binary case where each sensitive attribute is associated with two *subgroups*, e.g., female and male.

We are particularly interested in measures for the fairness of university admissions. While several fairness measures exist, we will see that the goal of this work – designing affirmative action policies – and its underlying assumptions limit the number of statistical measures to choose from.

Verma and Rubin [85] categorize fairness measures based on whether they measure differences in predictions, outcomes or both. This makes sense in cases where the fairness of a predictive algorithm is to be evaluated. In the case of our matching algorithms, however, we do not produce predictions, but only outcomes, i.e., admission decisions. For this reason, we are restricted to measures based on outcomes, which Verma and Rubin refer to as *statistical parity* measures. Within this category, two groups of measures are defined. Measures in the first group calculate the disparity in the probability of being assigned a positive outcome. The second group of measures conditions this disparity on "legitimate" attributes. Differences in the distribution of these features are perceived to be legitimate reasons for unequal outcomes. Such attributes might, for example, be the grades and test scores of students.

In order to decide which group of statistical parity measures we should select, we need to understand the assumptions underlying each of these groups. Friedler et al. [36] introduced two worldviews that differ in how they reason about differences in feature distributions between demographic groups: *what you see is what you get* (WYSIWYG) and *we're all equal* (WAE). WYSIWYG considers differences as innate to the subgroups. On the other hand, WAE sees them as the result of structural biases and aims for *equality of outcome*, i.e., equal outcomes across subgroups. This leads to disparate treatment of similar individuals, e.g., individuals who have similar grades but belong to different demographic groups. Green [41] refers to this as the conflict between *formal* and *substantive equality* where formal equality refers to *equal treatment* and substantive equality refers to *equal outcomes.* He states that in an unequal society, equal treatment of individuals is guaranteed to lead to unequal outcomes across subgroups. Thus, a seemingly "fair" algorithm that is free from direct *human bias* might still indirectly reinforce patterns of structural discrimination due to the underlying *population inequity.* An example of an algorithm free from human bias is the previously mentioned DA matching algorithm. In the case of Chilean university admissions, students are solely ranked on seemingly "legitimate" criteria: their grades and standardized test score. Yet, such algorithms might still create unequal outcomes due to the unequal nature of society.

In the education system, structural bias can manifest itself in an achievement gap across various groups, such as in a racial, indigenous, gender or income achievement gap. The reasons for differences in the distribution of grades and standardized test scores are complex and range from a lack of resources (see, e.g., [34]) through parental education (see, e.g., [65]) to the fear of confirming stereotypes (see, e.g., [80]). This, in turn, means that higher test scores do not necessarily imply more talent or more diligence. The realization that merit is not sufficiently represented by test scores is an argument for the WAE worldview and thus for affirmative action policies.

As grades and test scores are imperfect proxies for merit and correlated to demographic variables, avoiding the reinforcement of existing inequalities demands the usage of statistical parity measures that are not conditioned on scores. Instead the probabilities of being assigned a positive outcome, i.e., admission, are compared across subgroups. Two measures are typically used in this context. We follow the notation used in [11] of referring to them as statistical parity difference (SPD) and disparate impact (DI). SPD calculates the difference in the probabilities while DI refers to their ratio. DI is common in the United States where it is also referred to as the *4/5ths rule*. We choose SPD as our fairness measure as it has already been used in previous work on the same dataset [64].

3 Dataset of Chilean University Admissions

This section begins by describing the dataset of Chilean university admissions used in this study (§3.1). Based on the available data, sensitive attributes, i.e., demographic groups to which we apply affirmative action policies, are chosen (§3.2). Then, we explore current disparities in admission rates and possible reasons for them (§3.3-§3.5). We assume that if affirmative action policies are to be implemented for the next year, the current year highly influences this decision. As Sections 5 and 6 attempt to find bonus policies for the years 2016 and 2017, this section mainly analyzes the data for the previous years, i.e., 2015 and 2016.

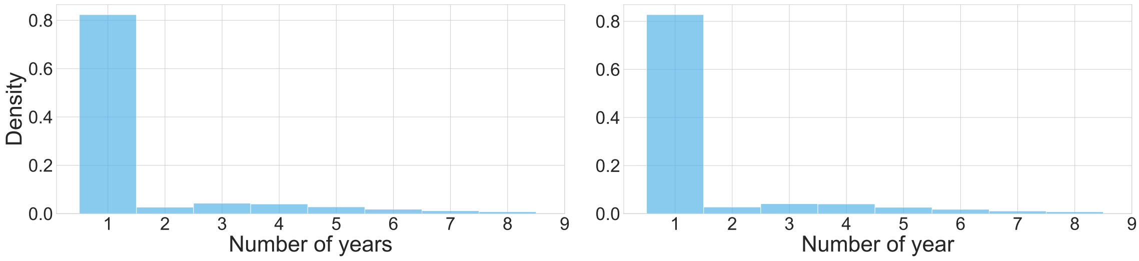
3.1 Data Description

We analyze anonymized data from the central admission system of Chile. This dataset contains information of all students who applied for university programs between 2004 and 2017, as well as the available programs. As seen in Table 1, the number of students and programs has generally increased slowly across these years, with a sharper increase in the number of programs in 2012, as multiple universities joined the central admissions process in that year.

Table 1: Number of programs and applicants per year.

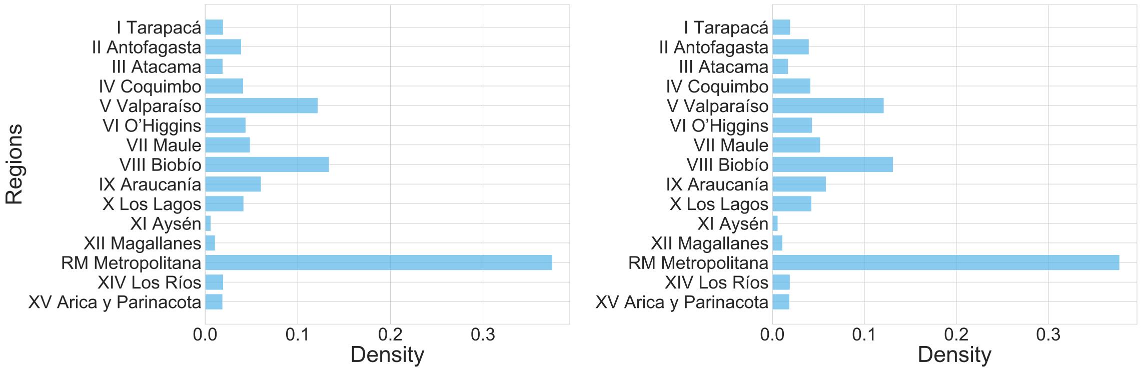
Year	Applicants	Programs
2004	40268	824
2005	44924	868
2006	46207	911
2007	48282	950
2008	48172	952
2009	48746	942
2010	50501	962
2011	49709	980
2012	51730	1335
2013	55488	1395
2014	57073	1419
2015	58773	1423
2016	59289	1436
2017	59743	1481

A variety of features is available for each student. Figure 3.1 shows, for example, how many years lie between students' graduation and the start of university. One year indicates that students start university right after their graduation, which is the case for the majority of students (81% in both 2015 and 2016). More than a third of applicants in the dataset come from the metropolitan region, which includes the capital Santiago. This is evident from Figure 3.2, which shows how many students lived in each of Chile's regions at the time of application. All regions – except for the metropolitan region, which is referred to as *RM* – are listed with their official Roman numeral. Other features in the dataset include the high school students attended and information about their families, such as their parents' education and job.



(a) 2015 (b) 2016

Figure 3.1: Distribution of number of years between high school graduation and university start



(a) 2015 (b) 2016

Figure 3.2: Geographic distribution of applicants by region

After taking standardized tests and obtaining their results, students submit a ranked list of the university programs that they are willing to enter. We will refer to this ranking of programs as the student's *preferences*. Students can list up to ten preferences, but typically only list a few programs: The average number of preferences between 2004 and 2017 ranged between 1.6 and 2.1. For each preference, i.e., each student-program-pair, a weighted score of the student's grade average in high school and standardized test scores is calculated. The weight of the individual components is determined by each program individually. The central institution knows the available spots of each program and matches students to programs based on these average weighted scores. We note that between 2012 and 2017 the majority of programs received fewer applicants than the number of spots they offered.

3.2 Choice of Sensitive Attributes

For the further analysis of the effects of affirmative action policies, we first need to decide what the demographic groups are to which we want to apply affirmative action policies. Commonly used sensitive attributes for affirmative action policies are race, gender and income [19, 28]. We have no data on the race or ethnicity of students, but the data contain a binary gender variable and students' household income. The latter is given on a scale whose range varies between years. As mentioned in Section 2, we want to measure the equality of admission rates between two groups. We therefore transform the household income variable into a binary variable. While we could simply split students into two groups based on household income, this would not reflect per capita income. For each student in the year's applicant pool, the household income per household member is thus calculated from the household income and household size features. Students are categorized as *low-income* if their household income per household member is below the median of the year's applicant pool, and *high-income* otherwise.

In what follows, we therefore compare the demographic subgroups *women* and *men* for the sensitive attribute *gender* and *low-income* and *high-income* students for the sensitive attribute *income*. While the share of female and male students is fairly stable between years, the share of low- and high-income students varies more (see Table 2). This can be attributed to the way in which students are split into income groups: If a relatively high number of students fall into the median household-income-per-household-member group in one year, the share of high-income students is higher than average in that year.

3.3 Differences in Application Behavior across Subgroups

There are two main factors that determine which students are admitted to which programs: the students' scores and the students' list of preferences. This section explores potential differences in these factors between subgroups.

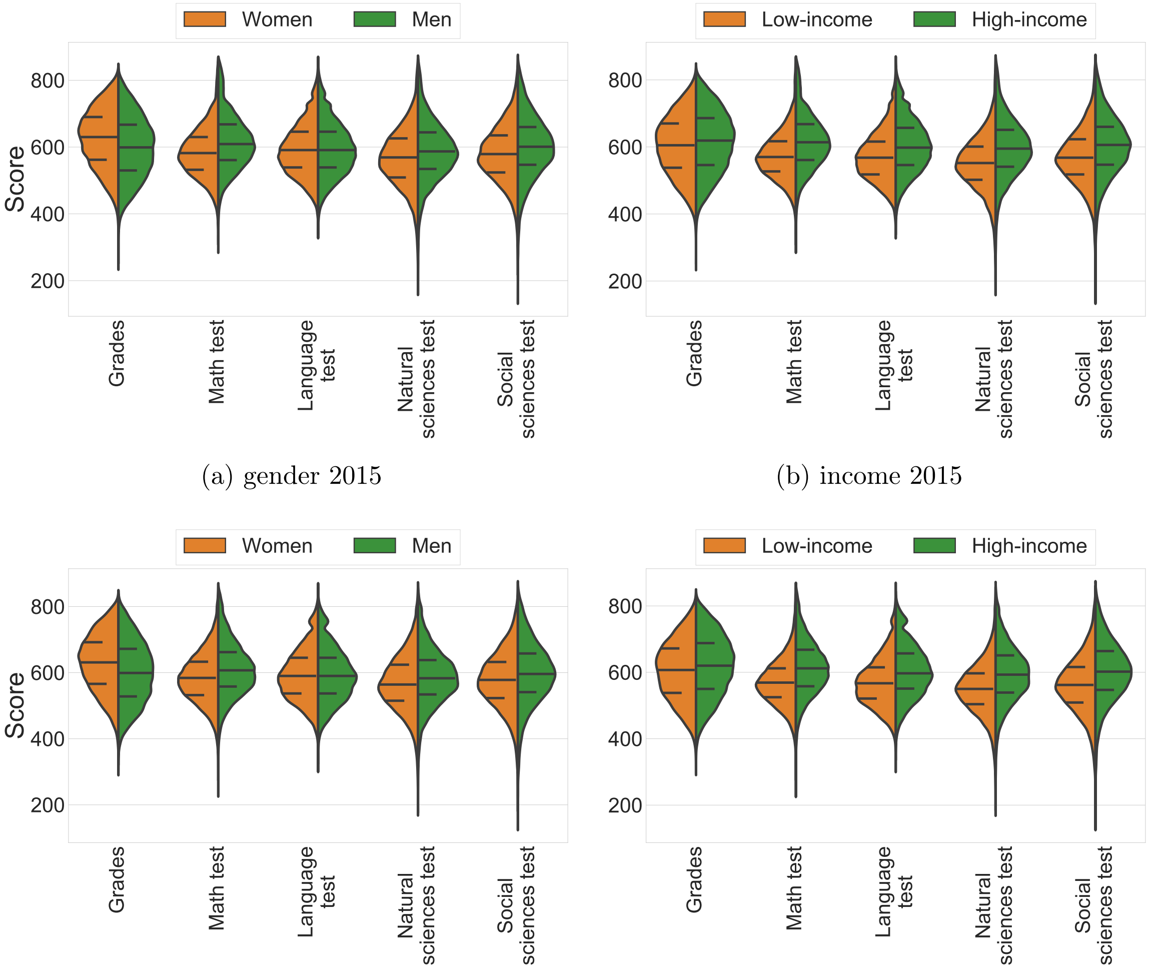
Table 2: Share of subgroups per year as percentages.

Year	Women	Men	Low-income	High-income
2004	49	51	38	62
2005	48	52	40	60
2006	48	52	38	62
2007	49	51	37	63
2008	49	51	46	54
2009	48	52	46	54
2010	47	53	44	56
2011	47	53	41	59
2012	48	52	46	54
2013	48	52	45	55
2014	48	52	41	59
2015	48	52	39	61
2016	49	51	36	64
2017	49	51	47	53

3.3.1 Differences in Score Distributions

In order to apply to university programs that are part of the Chilean centralized admissions system students have to take standardized tests. The two essential tests are in mathematics and the Spanish language. In addition, students choose a test in natural or social sciences (or both) depending on what is required by the programs that they want to apply to. Each of these standardized tests is scored on a scale from 150 to 850 with 850 being the highest achievable score.

We observe differences in the distributions of grades and test scores between gender and income groups (see Figure 3.3) in line with previous work (e.g., [6, 64, 73]). These differences are persistent across years. Note that "grades" in Figure 3.3 refers to the original high school grades and not their transformed version discussed in Section 2. In summary, women outperform men in high school, but do worse on standardized tests, which is consistent with previous work that finds women at a disadvantage when taking “high-stakes” tests (e.g., [8, 77]). The differences between income groups are more marked than the differences by gender. High-income students have slightly higher high school grades than low-income students and much higher scores across all standardized tests. As programs generally put more weight on standardized test scores than on high school grades, these differences might lead to a disadvantage for both women and low-income students.



(a) gender 2015

(b) income 2015

(c) gender 2016

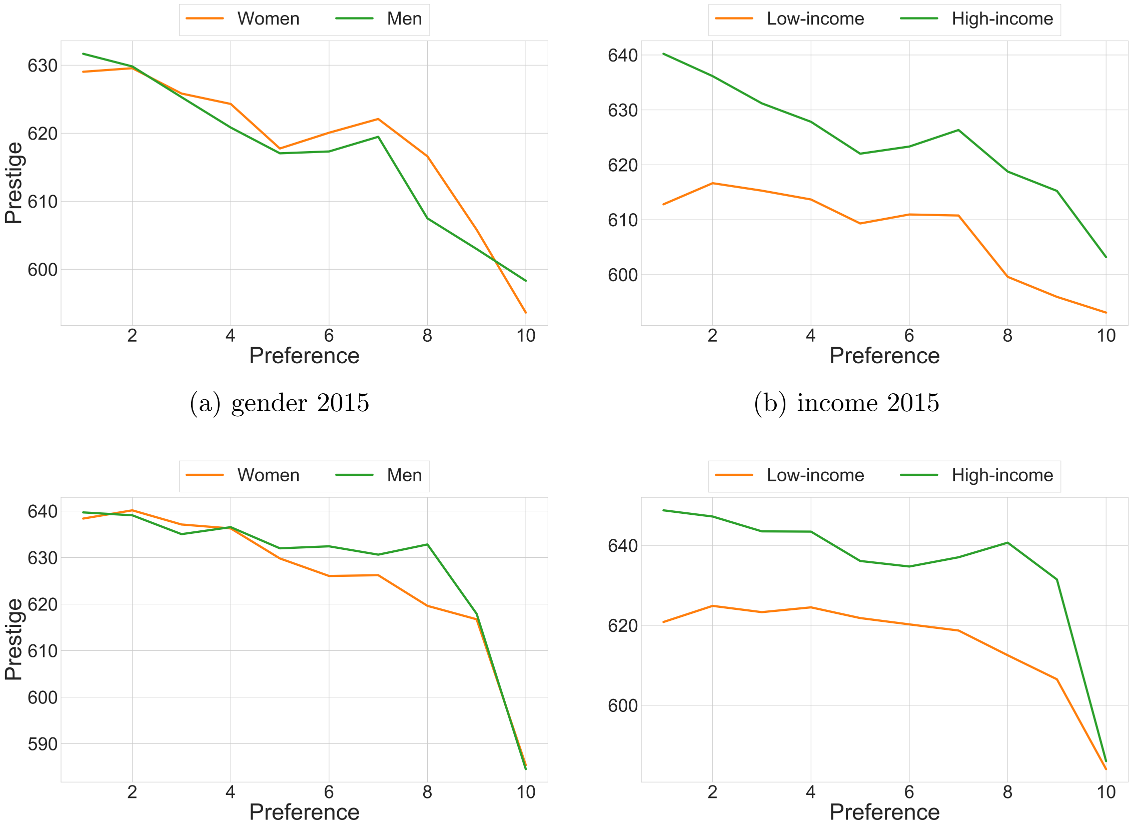
(d) income 2016

Figure 3.3: Distributions of grades and standardized test scores.

3.3.2 Differences in Preferences

We also observe differences in the prestige of the programs to which students apply. To describe these differences, we note that some programs are more sought-after than others by high-scoring students. We define the *prestige* of a program as the (weighted) average score of its admitted students in the previous three years. In Figure 3.4, we show that there is no noticeable difference in the prestige of programs women and men list as their preferences. In contrast, high-income students prefer programs of higher average prestige compared to low-income students, particularly in their first preference.

This difference of behavior could be intrinsic to the group, e.g., if high-income students were more ambitious or optimistic about their chances than low-income stu-



(a) gender 2015

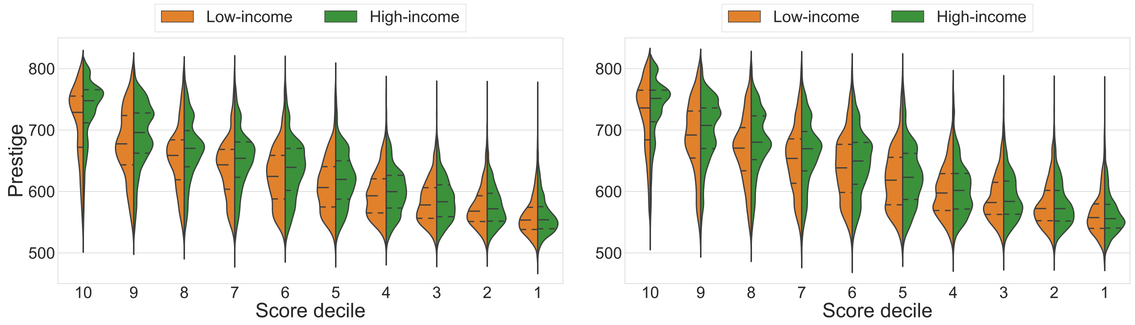
(b) income 2015

(c) gender 2016

(d) income 2016

Figure 3.4: Average prestige of students' preferences across subgroups.

dents. However, we note that high-income students tend to have higher scores than low-income students, and hence, we should compare the prestige of preferences between groups controlling for test scores. Indeed, we find that this greatly reduces differences, in particular for students with scores below the median. As Figure 3.5 shows, differences between low-income students and high-income students whose scores are above the median are notably lower, too. This suggests that application behavior is more driven by the scores of the applicants than by their demographics. We note, however, that high-income students in the top decile of scores still apply to programs of higher prestige than low-income students in the same score decile. This is in line with research conducted by Hoxby and Avery [49] that found that high-achieving high-income students in the United States apply to more selective colleges than high-achieving low-income students. Reasons they mention for this disparity are that low-income students are less often encouraged to apply to selective colleges and that they are less likely to know a person who has attended such a college.



(a) 2015

(b) 2016

Figure 3.5: Average prestige of first choice by income, controlled for score.

3.4 Differences in Admission Rates across Subgroups

We consider subgroups that are defined by a sensitive attribute A and as justified in Section 2 measure the inequality between them through their SPD:

$$P(Y = 1|A = a) - P(Y = 1|A \neq a), \tag{1}$$

where $Y = 1$ indicates a positive outcome, i.e., being admitted into the program. In what follows, a always marks female students if the sensitive attribute A is gender and low-income students if the sensitive attribute is income. Low absolute values of SPD are desired – and perfect equality is achieved for a value of 0. As this might be almost impossible to achieve in practice, we follow the thresholds given by [11] and consider values inside [-0.1, 0.1] as acceptable, while values outside this range are considered *strongly unequal*.

Before setting out to design policies that would equalize admission rates, we first explore their current disparities. For this, we measure the SPD between female and male applicants as well as between low- and high-income applicants. Note that the calculation of the admission rates for a program disregards students that had listed this program as one of their preferences, but who were accepted into a program that they had ranked higher, as such students were neither admitted nor rejected by the program.

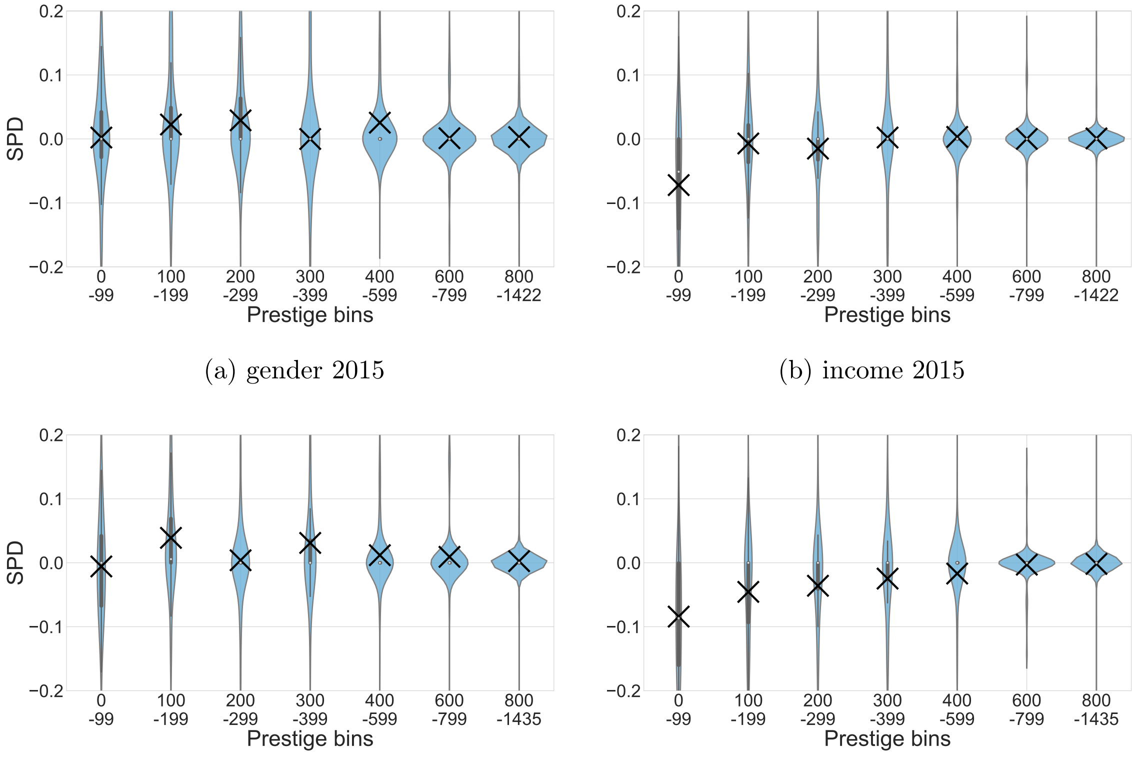
Figure 3.6 shows the distribution of SPD values for gender and income for all programs, ordered by their prestige. According to Eq. 1, values below 0 indicate a lower admission rate for women and low-income students. While the SPD is around 0 for most programs and hence well within the thresholds, 9% of programs in 2015 and 8% in 2016 have strongly unequal gender admission rates. 9% of programs in

2015 and 10% of programs in 2016 have strongly unequal income admission rates. However, not all programs with strongly unequal admission rates require affirmative action policies: Among the programs with strongly unequal admission rates are also programs that accept the underrepresented group at a higher rate than the overrepresented group. If we only consider programs where the group with the lower admission rate is underrepresented, this leaves 4-5% of programs for gender and 6-7% for income in 2015 and 2016.

We note that these disparities occur despite two previously described observations: First, high-achieving low-income students tend to "undersell" themselves by applying to less prestigious programs than high-income students with similar scores. Secondly, the high school grades between low- and high-income students have already been equalized to some extent (see Section 2.1). The share of programs with strongly unequal admission rates between low- and high-income students would likely be more pronounced without the difference in application behavior and the described form of affirmative action.

In order to identify in which types of programs particular subgroups are disadvantaged, we analyze the programs with strongly unequal admission rates to the detriment of the underrepresented group. For 2015 and 2016, we find that 58-65% of the programs that disadvantaged women are in the field of technology, such as engineering. 44-48% of the programs that disadvantaged men in 2015 and 2016 are in the field of health, in particular nursing and obstetrics. Low-income students are at a disadvantage across a larger array of programs. In both 2015 and 2016, they are primarily (37-39%) disadvantaged in the field of health. Different from the disadvantage of men in this field, low-income students tend to have lower acceptance rates for medicine programs. We also note that the programs in which low-income students are disadvantaged tend to be in the capital Santiago (41-45%). As this is where the more prestigious programs tend to be found, this finding corroborates what can be seen in Figure 3.6: Programs that are notably less likely to accept low-income students than high-income students tend to be the most prestigious ones. This is in line with previous observations that the most selective institutions tend to be the ones in which the representation of minority students is lower (see, e.g., [46]).

We note that while the differences in score distribution (shown in Section 3.3) might seem small, their consequences are large, and men and high-income students could “crowd out” women and low-income students. To demonstrate this, we consider a hypothetical scenario in which all students apply to the same program. Figure 3.7



(a) gender 2015

(b) income 2015

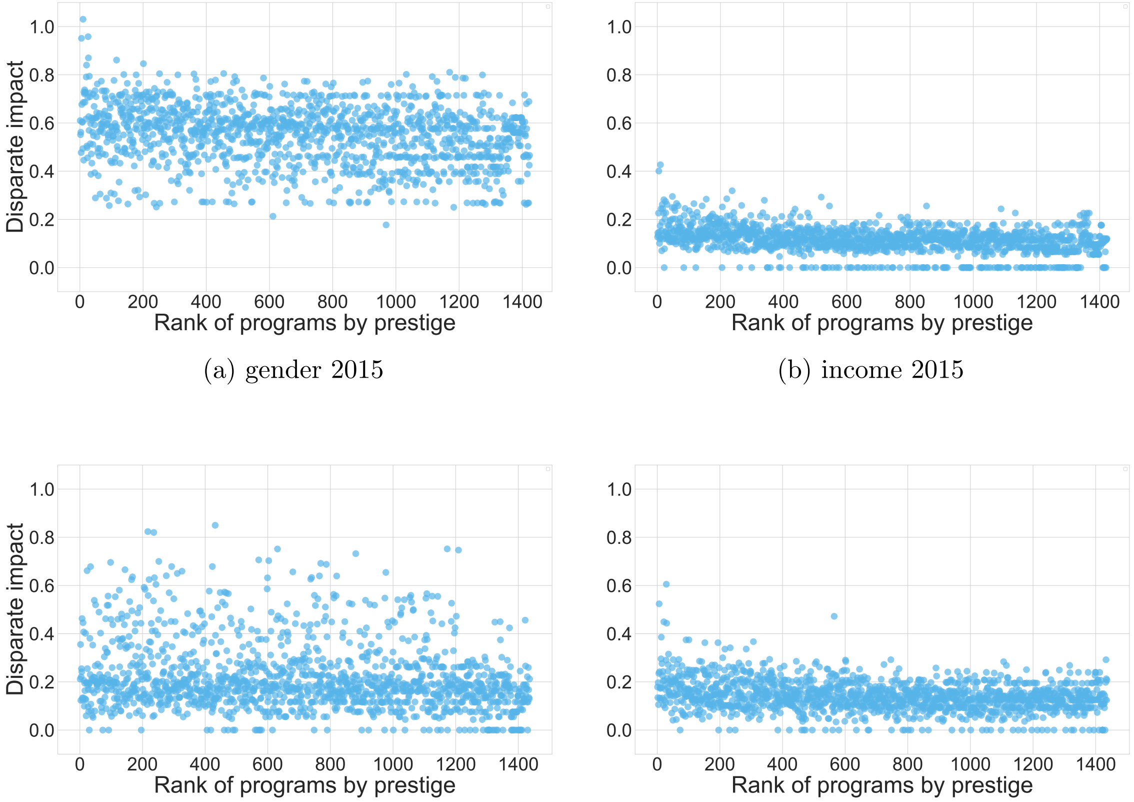
(c) gender 2016

(d) income 2016

Figure 3.6: Distribution of SPD with crosses marking the mean. Programs have been ranked by prestige (higher to lower). Labels in the x-axis indicate program ranks in each bin.

shows the resulting difference in admission rates we would observe if all students applied to the first, second, third, etc., most prestigious program. As the admission rates are bound to be very low if all students compete for a spot at the same program, differences between admission rates would be extremely small and thus uninformative. The plot therefore does not show the SPD, but the previously mentioned DI measure, i.e., the ratio of the resulting admission rates. Values below 1 indicate a lower admission rate for women and low-income students compared to men and high-income students, respectively. Programs have again been ranked by their prestige (higher to lower). We see that under this hypothetical scenario, the vast majority of programs would exhibit high disparities in admission rates.

In light of this, why do not all programs actually show such high disparities in their admission rates?



(a) gender 2015

(b) income 2015

(c) gender 2016

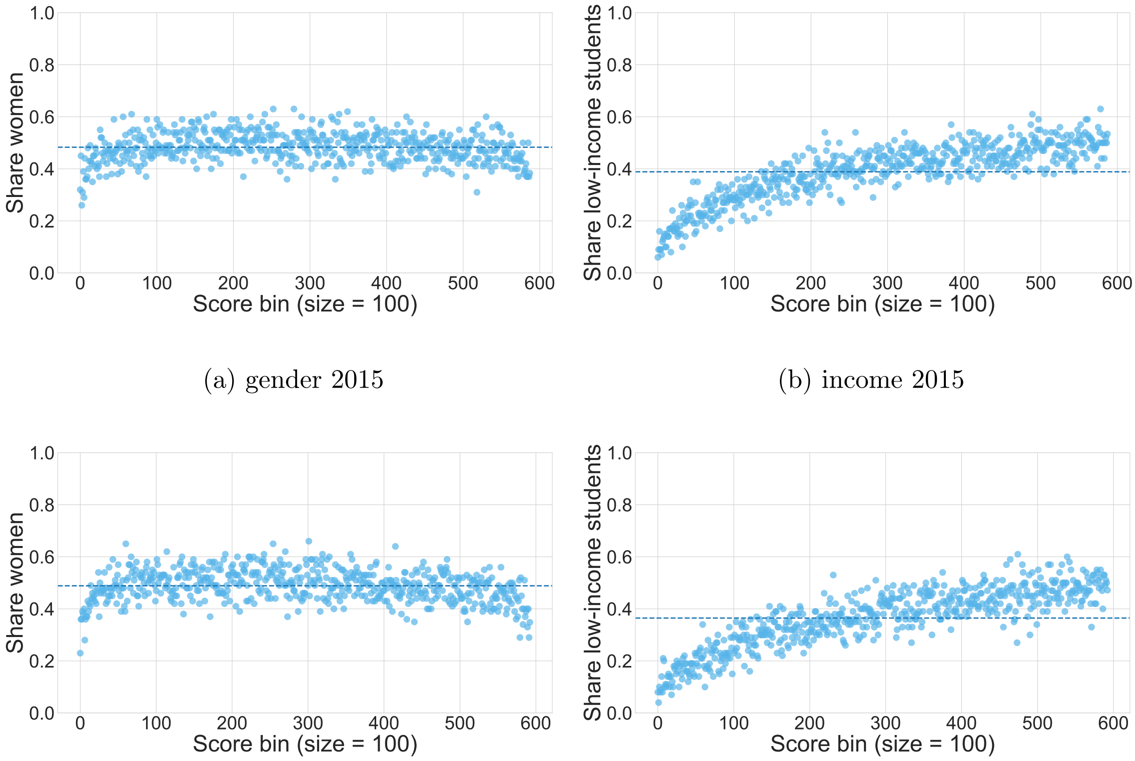
(d) income 2016

Figure 3.7: Distribution of DI assuming every student applied to the same program.

First, note that, while Figure 3.6 showed that differences in admission rates are more pronounced for prestigious programs, this is not observed in the hypothetical scenario of Figure 3.7. This is because in the actual data low-prestige programs tend to get fewer applications (as mentioned in Section 3.1, a majority of programs do not receive enough applications to fill all their seats) and often accept all their applicants – which inevitably leads to perfect equality between admission rates.

The fact that the disparities of competitive programs are more pronounced in Figure 3.7 than in Figure 3.6 can be explained by what we saw in Section 3.3: The application behavior of students depends heavily on their scores. Programs are thus likely to receive applications from students with similar grades which makes the differences less pronounced.

As all students apply to the same program in Figure 3.7, differences in admission rates depend on the best students. The group of the best students differs between programs as each program weighs high school grades and standardized test scores differently. For a generalization, we calculate a weighted average of each student's scores with the same weights. These weights are set to the median of the weights programs used in the respective year. Figure 3.8 ranks students by the resulting average scores and shows the share of women and low-income students for groups of 100 students. Women are underrepresented among the students with the highest and lowest average scores. The share of low-income students is particularly low among the best students. This underrepresentation of women and low-income students in the group of the best students explains why men and high-income students are admitted at higher rates in the hypothetical scenario shown in Figure 3.7.



(a) gender 2015

(b) income 2015

(c) gender 2016

(d) income 2016

Figure 3.8: Demographics of students grouped by (weighted) average score (higher to lower). Labels in the x-axis indicate the bin number. The dotted line marks the share of women and low-income students in the entire applicant pool for comparison.

3.5 Variance in Admission Rates across Years

The previous section has shown the difference in admission rates for one year at a time. While the share of programs with strongly unequal admission rates is fairly constant across years, this section demonstrates that this is not the case if we analyze the SPD of each program individually across years.

To demonstrate to what extent the SPD of a program in a given year t is predictable based on its SPD in previous years, we train a regression model. The features on which the model is trained are *lags* and *differences* [45]. Lags, in our case, are the SPDs of previous years. The first lag, for example, is the SPD of the previous year, denoted as SPD_{t-1}. Differences are measured in the SPDs of two consecutive years: $\text{SPD}_{t-i} - \text{SPD}_{t-i-1}$. To evaluate this regression approach, we build two time series datasets on which we train separate regression models. One dataset includes the lags (SPD_{t-i}) and differences ($\text{SPD}_{t-i} - \text{SPD}_{t-i-1}$) of the previous three years ($i = \{1, 2, 3\}$) and one of the previous five years ($i = \{1, 2, 3, 4, 5\}$). We do not evaluate higher values for i as the construction of lags and differences is limited by the number of years for which admission data is available and we want to ensure that the model is trained and tested on a sufficient number of programs. As a first inspection of the admission rates of individual programs has exhibited instances where a program favored different subgroups between years, we do not expect the relationship between the SPDs to be linear. Therefore, we train a random forest regressor [18, 39], i.e., a non-linear model, on the constructed time series data.

We compare this approach to three baselines: (i) always predicting 0 (i.e., equal admission rates), (ii) assuming that the SPD is the same as in the previous year, and (iii) taking the average of the SPD of the last five years as the predicted SPD. Specifically, we compare how well these approaches predict the difference in admission rates for gender and income groups in the years 2016 and 2017 – the years for which we suggest bonus policies in Sections 5 and 6.

The first interesting result is that predicting perfectly equal admission rates leads to the lowest MAE. It thus performs better than predicting last year's SPD or the average of the last five years' SPDs. The random forest approach showed the second-best performance. Its results depend on the number of features included in the time series data. In our experiments, we have not been able to find MAEs lower than baseline (i) by varying the number of years.

Table 3: Mean absolute error (MAE) relative to true SPD. Smaller values are better.

		All programs		Strong disparity predicted	
	Strategy	Gender	Income	Gender	Income
2016	Equal admission rates	**0.0260**	**0.0276**	-	-
	Last year's SPD	0.0383	0.0372	0.1791	0.1693
	Last 5 years' average SPD	0.0395	0.0360	0.1281	**0.0927**
	Random forest - 3 years	0.0278	0.0294	0.1059	0.1818
	Random forest - 5 years	0.0314	0.0293	**0.0861**	0.0948
2017	Equal admission rates	**0.0205**	**0.0211**	-	-
	Last year's SPD	0.0330	0.0310	0.1932	0.1737
	Last 5 years' average SPD	0.0319	0.0276	0.1656	**0.0887**
	Random forest - 3 years	0.0247	0.0251	0.1121	0.1458
	Random forest - 5 years	0.0229	0.0234	**0.0832**	**0.0887**

This means that there are large variations in admission rates of the same program over the years. This observation is important when it comes to applying affirmative action policies, as we have to ensure that each suggested policy has the intended effect. If the admission rates favored women in the last year, but favor men in the next year, we risk advantaging a group who without intervention would have the higher admission rate. One of the reasons for why this might occur is the underrepresentation of a subgroup in a program's pool of applicants. In 2016, for example, two engineering programs had similar statistics in terms of application numbers: Program A received 2 applications by women and 53 by men while program B received 5 applications by women and 61 by men. Both programs accepted 2 women which in the case of program A means an admission rate of 100% for women. Program B, on the other hand, admitted 40% of the female applicants. The admission rate of the smaller subgroup therefore heavily depends on the scores of the few members that applied in that year. As these scores might vary between years, differences in admission rates can easily vary, too.

If the goal of the policy is equalizing admission rates, we thus only want to apply our policy to programs where this is not the case. Hence, we are particularly interested in the programs that are predicted to have a strongly unequal SPD. First, we note that the number of programs that the random forest predicts to have strong inequalities is much lower than the number of programs that actually show strong inequalities. The reason for this is that predicted SPDs tend to be close to 0 as we saw that baseline (i) produces the lowest MAE. As we can see in Table 3, the smallest MAE is in the range of 0.08 to 0.09. Recall from Section 3.4 that 90% of programs are within the thresholds of -0.1 and 0.1. The variance in the predicted SPD is thus quite high compared to most observed SPDs, which means that often a subgroup that is predicted to be accepted at a lower rate is actually accepted at a higher rate than the other subgroup in the following year. The MAEs are therefore too high for our context and we conclude that the difference in admission rates is difficult to predict. This again underlines why the programs to which admission-rates-equalizing affirmative action policies are applied have to be selected carefully.

4 Predictive Models of Application Behavior

To ensure that the admissions process is transparent, potential affirmative action policies have to be announced before the start of the application period when only data about the programs are available. Consequently, the application behavior of students (i.e., who will apply and for what program) is largely uncertain when the policies are determined. To deal with this uncertainty, we generate samples of possible future application sets based on historical data and, as we describe in Section 5 and 6, use them to estimate the effect of different bonus policies.

We generate application sets in the following two steps: First, we sample a pool of applicants by randomly choosing students from previous years in the data. The size of each pool is drawn from a Poisson distribution with rate s, where s is the size of the student cohort for the most recent year. s is chosen as the rate because Section 3 has shown that two consecutive years tend to receive similar numbers of applicants. Second, we generate applications for the sampled applicants. Ideally, we could simply sample students with their actual applications. This way, the result of the sampling step would be an application set on which we can test different bonus polices. While this is technically possible as the necessary data is available, it is not useful in our context as programs change from year to year. The applications of students who applied to university in previous years include applications to programs that no longer exist. At the same time, older application data do not include applications to programs that have only been created in recent years. However, in order to ensure that the application sets on which we test bonus policies are sufficiently similar to next year's application set, the programs in the dataset have to be similar. We therefore train a model to predict which programs a given student will apply to. This model is trained on the most recent year of data as we assume that the programs in that year largely overlap with the programs in the upcoming application round. The resulting model is then used to generate random applications for the aforementioned sampled applicant pools. In what follows, Section 4.1 describes the three different types of models we consider for predicting the application behavior of students, while Section 4.2 compares them experimentally.

4.1 Description of Predictive Models

We describe three different approaches to model students' application behavior. First, we compare two types of predictive models: a classifier and a regression. Both predict a (probabilistic or continuous) value for a given student and a given program. Instead of trying to correctly predict the value of each program individually, we could also attempt to rank all programs by their relevance to a student – without regard to exact probabilistic or continuous values. The third approach that we examine is therefore a ranking model that orders all programs for a given student. Each of these three approaches can be used to select a set of $n \in \{1, 2, \ldots, 10\}$ programs as potential preferences of a given student.

4.1.1 Multi-Label Probabilistic Classifier

We first consider a classifier that takes a student's features as input. These features are the preprocessed data described in Section 3 and thus include, for example, the student's standardized high school grades and household size. The output could be the programs to which a student is predicted to apply. This corresponds to a multi-label classification task [83] where each student can be assigned multiple labels, i.e., programs. However, we are specifically interested in the probabilities with which students are predicted to apply to each program. The Multi-Label Probabilistic Classifier (MLPC) approach thus outputs the probability with which the given student applies to the given program. To select the student's predicted applications for the generated application set, we sample n programs without replacement based on the application probabilities predicted for all programs. The sampled programs are then ranked by decreasing application probability.

4.1.2 Multi-Output Regression

As MLPC, Multi-Output Regression (MOR) [74] also takes a student's features as its input. However, instead of outputting a binary variable (will the student apply to this program or not?) or probability for a given program, MOR outputs a continuous value [17]. This continuous value represents how relevant the program is to the given student. We can then select the n programs with the highest predicted relevance as the student's applications.

To this end, each combination of student and program is assigned a relevance score which is used as the target variable. As each student can list up to ten programs as their preferences, the score ranges between 10 (for the top preference of the student) and 0 (for programs not applied for).

4.1.3 Learning-to-Rank

While the previously described models consider each program individually, Learning-to-Rank (LTR) models aim to solve ranking problems [61]. Their task is thus the ordering of a list of items. In our case, the list of all programs should be ranked by the relevance of each program to a given student. For this learning task, the programs' features are part of the training. These features include the location of the university and the weights the programs assign to, e.g., the standardized test in natural sciences.

Liu et al. [61] categorize LTR models into three groups: the pointwise approach, the pairwise approach and the listwise approach. The pointwise approach tries to predict each item's relevance individually to then rank the items based on the predicted scores [61]. This is thus similar to MOR – it simply additionally takes each program's features into account. Both the pairwise and the listwise approach, however, consider the relevance of more than one item [61]. While the pairwise approach would in our case learn to rank pairs of programs by relevance, the listwise approach would directly order all programs. As the listwise approach has the advantage that it can compare the rankings of all items (i.e., programs), it frequently outperforms the other two (see, e.g., [23, 53]). We thus only consider a listwise ranking model for our experimental evaluation.

In this case, the LTR approach's input consists of the features of one student combined with the features of all programs. It is trained to rank the programs by their relevance, so that the output is a ranking of all university programs for the given student. Out of these ranked programs the n highest-ranked ones are selected as the student's applications for the generated application set.

Having the programs' features be part of the input has the advantage that this approach could even be applied to previously unseen programs. Imagine, for example, a program that is newly offered in the upcoming application round. As the model was trained on a combination of student and program features, all we need to know to make predictions for a new program are the program's features, which are available in advance.

4.2 Experimental Comparison of Predictive Models

Using data from the year 2016, we train the aforementioned predictive models and evaluate their performance in predicting the application behavior of students. We use the well-known normalized discounted cumulative gain (nDCG) [54] as the performance measure. nDCG is a metric designed to evaluate rankings of results from search engines. Its advantage is that it does not treat all search results equally, but emphasizes the importance of giving highly relevant search results a high ranking. This is important in the context of university applications where in our case students apply to at most ten programs out of a pool of about 1 500 programs. What is most relevant to students is their first choice. Our metric therefore has to put a large weight on the predicted first choice whereas it is less important whether the predictor can correctly distinguish between a program that was a student's 10th choice and a program they did not apply to. This is achieved by the nDCG score by discounting the influence of lower-ranked search results on a logarithmic scale.

Note that for all of the models, the number n of preferences listed by each student is drawn from a (truncated) Poisson distribution with rate λ, as $n \sim P_\lambda$, where λ is the average number of preferences students listed in the year on which we train the model. The distribution is truncated to sample only values in $\{1, 2, \ldots, 10\}$ as applying to more programs is not permissible in the Chilean admissions process.

4.2.1 Experimental Results of Multi-Label Probabilistic Classifier

Initially, we evaluate two baselines: random and unigram. For the random baseline, we predict each program with the same probability; and for unigram, we predict every program p with $\frac{\text{\# applications to } p}{\text{\# total applications}}$. The random baseline has an nDCG score of 0.18 on the testing data and the unigram baseline achieves an nDCG score of 0.24. We train a random forest classifier [18] and tune its hyperparameters with Bayesian optimization [70] over 100 iterations to reach an nDCG score of 0.43 on the testing data (see Section 6 for technical details on Bayesian optimization).

4.2.2 Experimental Results of Multi-Output Regression

A simple linear regression model serves as MOR's baseline. It achieves an nDCG score of 0.27. To improve on this metric, we run a random forest regression model [18, 39] and tune its parameters with Bayesian optimization over 100 iterations. This leads to an nDCG score of 0.44.

4.2.3 Experimental Results of Learning-to-Rank

If we were to train an LTR model on an entire year, we would have to form all possible combinations of students and programs in that year. For 2016, this would result in a dataset of approximately 90 million rows. Such a large dataset considerably slows down the training process, so we decide on working with a smaller dataset. For a first evaluation, we sample 1 000 students and combine them with all programs.

As for MOR, a linear regression serves as the baseline for this approach. It reaches an average nDCG of 0.29 on the testing set. In our attempt to find a better model than the linear regression, we train a LambdaMART model [20], which is a tree boosting model [37] adapted to the ranking problem. `XGBoost`'s [26] implementation of LambdaMART allows for the performance of both pairwise and listwise ranking. As described above, we choose the listwise approach that aims to optimize the nDCG score. We train this model with `XGBoost`'s default parameters and reach an nDCG score of 0.31. By manually testing small changes in the hyperparameters (e.g., increasing the maximum depth of the boosted trees from 6 to 10) we reach an nDCG score of 0.35.

This score is notably lower than the nDCG benchmarks achieved by the previous two approaches. However, we have trained and tested this model on less than 2% of the actual dataset. While the performance could likely be improved by using the full dataset and tuning the hyperparameters further, this first evaluation has already shown that the LTR approach is inefficient for generating applications. The reason for this is that for each sampled applicant pool, this approach would require building the Cartesian product of the sampled applicants and all programs which would lead to a dataset that is in size comparable to the previously described dataset with 90 million rows. We thus do not try to optimize the model's performance further.

4.2.4 Choice of Model

We opted to use MLPC for the rest of the analysis, as it has practical advantages over the other methods.

Both MLPC and MOR proved to be more efficient to train and deploy than LTR. Not using LTR as the prediction model means forgoing one of its advantageous characteristics: Its ability to make predictions for new programs. However, such a feature is not required in this context since affirmative action policies should only be implemented sparingly and when necessary. New programs are thus unlikely to warrant such an intervention.

Contrary to MOR, MLPC outputs categorical probability distributions. This allows for randomness in the students' applications as programs can be sampled from this probability distribution. MOR, on the other hand, always picks the n programs with the highest predicted ranking, which often leads to unpopular programs never appearing in the sampled preference lists because their values are always dominated by more popular programs. As the performance of MLPC is slightly lower than that of MOR, we thus trade this loss in performance for increased practicality. For comparison, we also trained MLPC on the applications from 2015 to predict students' applications in 2016. Within 100 iterations of Bayesian optimization, we found a model with an nDCG score of 0.42 on a separate testing set.

5 Policy Design Strategies for Single Sensitive Attribute

As Section 3.4 has shown, some programs have strongly unequal admission rates due to differences in their applicants' test scores. To reduce these gaps, we consider the use of bonus policies. Essentially, we set out to award bonus points to students from disadvantaged groups in order to make their admission rates similar to that of students outside the group. Recall that, in the setting we study, such policies have to be announced in advance, so that students can consider those bonuses when applying. This section outlines and compares different strategies for designing a robust bonus policy.

5.1 Policy Design

We formalize the goal of policy design in terms of an objective function – and then describe the policy design strategies through which we aim to optimize it.

A *design strategy* is an approach for finding a bonus policy for a given program and a given sensitive attribute. The strategies we compare suggest a policy based on the ideal policies of multiple application sets. The ideal bonus policy for each application set is found by evaluating the effects of bonus values in the interval of $\{0, 1, ..., 120\}$. This interval is chosen as students' weighted average scores range between 150 and 850. Since we want to balance utility with the equality of admission rates, we do not expect ideal bonus point suggestions to exceed 120.

Bonuses in this interval are assigned to both subgroups of the given sensitive attribute. In the case of the search for a gender bonus policy, bonuses for both women and men are thus evaluated. Bonuses for low- and high-income students are compared when searching for an income bonus policy. For practical reasons that will be further described in Section 6, this work always applies bonuses to the same subgroup: In the case of gender policies, bonuses are always given to women and in the case of income policies, bonuses are always given to high-income students. It can easily be seen that increasing the scores of students with sensitive attribute $A \neq a$ by b points is equivalent to decreasing the scores of students with sensitive attribute $A = a$ by b points. Hence, bonus points to men and high-income students are in practice implemented as negative bonus points for women and low-income students, respectively.

For each tested bonus policy, the DA algorithm (see Section 2) is employed to simulate the application process. The objective function is then evaluated on the resulting matching. For each application set, the bonus that minimizes the objective function is considered ideal for it. Each design strategy simply suggests the average of the ideal bonuses over all its application sets.

5.1.1 Problem Definition

We wish our admission policy to lead to the admission of the students with the highest scores, as well as to a reduction in admission rate disparities. This section thus defines the objective function of the bonus policy as a linear combination of both the equality of admission rates and utility, i.e., the scores of accepted students. Specifically, we measure utility μ_b as the average score of the students who are admitted when b bonus points are given. As utility might vary strongly between application sets, we calculate the loss of μ_b compared to μ_0, i.e., the utility when no bonus policy is implemented.

$$o_b = (\mu_0 - \mu_b) + \lambda \cdot |\text{SPD}_b|, \quad \lambda \geq 0. \tag{2}$$

5.1.2 Policies based on Predictive Model

Our strategy for suggesting policies is evaluating a range of possible policies on a large number of generated application sets. As discussed in Section 4, we choose MLPC to model students' application behavior and thus to generate these application sets. To predict applications for 2017, we train the model on data from 2016; and to predict applications for 2016, we train on data from 2015. Once we have trained the model, we deploy it to sample a number n of possible application sets. As explained in Section 4, each application set consists of a sampled pool of students together with their predicted program preferences. We experiment with $n = 50$ and $n = 200$ sampled application sets to evaluate possible bonus policies.

5.1.3 Policies based on Historical Data

A simple approach to choose a bonus policy is to compute what bonus policy would have been optimal in the previous years. Such an approach has the disadvantage that we are limited to the number of years for which we have data for the program that we want to design a policy for. In addition, such historical data may be difficult to obtain. For example, it may be that, due to legal requirements, only aggregate statistics about student applications can be used or made public – in such cases, using a model built from such statistics would be necessary. Nevertheless, we also include this approach in our empirical evaluation, as a baseline. Specifically, we consider the design strategy that (in hindsight) computes the optimal bonus for the application sets of the past one, three or five years, and then uses the average of those bonus values as the bonus value for the upcoming application round.

5.2 Experimental Evaluation

We evaluate each strategy for both sensitive attributes – gender and income – and for the years 2016 and 2017. To find a fitting value for λ in Eq. 2, we calculate the median differences in grades and test scores between subgroups. We weigh the sum of the absolute differences with the median weights that programs give to these factors. The resulting values are equal for 2016 and 2017, so we optimize the objective function for gender subgroups with $\lambda = 23$ and for income subgroups with $\lambda = 28$ in both years.

We first utilize the strategies to suggest policies for all programs and evaluate their overall effect. As Section 3.5 demonstrated, bonus policies to equalize admission rates should only be used sparingly in order to avoid adverse effects. We therefore also evaluate the strategies when policies are only applied to programs that show consistent inequalities in admission rates over time. In both cases, we analyze the effects of each policy separately for each program, i.e., assuming that only the program under consideration enacts a bonus policy.

5.2.1 Policies for All Programs

We begin by comparing the objective function values resulting from applying the different strategies to the smallest achievable ones, i.e., the values that result from applying the ideal bonus policies. In Table 4, we show the mean and standard distribution (SD) of this difference for both sensitive attributes and the years 2016 and 2017. Note that, according to Table 4, policies that use more application sets for their suggestions generally lead to both smaller errors and less variance in the error.

Tables 5 and 6 split the findings from Table 4 into its two components: utility and SPD. We note that while of course no strategy can find a better value for the

Table 4: Error in objective function relative to ideal policies. Smaller values are better.

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	0.52	1.47	0.62	1.76
	Historical - 3 years	0.40	1.22	0.52	1.53
	Historical - 5 years	0.42	1.23	**0.49**	1.49
	Predictive - 50 sets	0.36	**1.09**	**0.49**	**1.40**
	Predictive - 200 sets	**0.35**	**1.09**	0.50	**1.40**
2017	Historical - 1 year	0.37	1.11	0.44	1.31
	Historical - 3 years	0.30	0.94	0.34	1.06
	Historical - 5 years	0.32	0.99	**0.33**	**0.98**
	Predictive - 50 sets	**0.28**	**0.90**	0.37	1.13
	Predictive - 200 sets	0.29	**0.90**	0.36	1.11

objective function than the ideal policy, it is possible for a strategy to suggest policies that lead to a lower loss in utility or a smaller gap in admission rates. Indeed, we find that on average all strategies produce a smaller utility loss than the ideal policies. However, the difference in admission rates is on average higher. In most cases, more application sets again lead to smaller and more robust errors.

Table 5: Difference in utility loss relative to ideal policies. Lower values are better.

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	-0.0182	0.2882	-0.0238	0.3247
	Historical - 3 years	-0.0465	0.2162	-0.0466	0.3092
	Historical - 5 years	-0.0481	0.2172	-0.0480	0.2845
	Predictive - 50 sets	-0.0561	**0.2108**	-0.0525	0.2864
	Predictive - 200 sets	**-0.0563**	0.2109	**-0.0530**	**0.2827**
2017	Historical - 1 year	-0.0220	0.2238	-0.0425	0.3506
	Historical - 3 years	-0.0340	**0.2077**	-0.0587	0.3196
	Historical - 5 years	-0.0358	0.2101	**-0.0602**	0.3191
	Predictive - 50 sets	**-0.0443**	0.2104	-0.0574	0.3256
	Predictive - 200 sets	-0.0438	0.2103	-0.0576	**0.3136**

While optimizing the objective function, it is also important to ensure that the gap in admission rates is decreased compared to when we do not intervene. Note that no intervention (i.e., using no bonus, $b = 0$) is almost guaranteed to lead to the lowest utility loss as utility is typically highest when no affirmative action policy is applied. To explore the effect on the admission rates gap, Table 7 compares both SPDs. Specifically, it compares the difference in the admission rate gaps (measured as the absolute value of SPD) with and without a bonus policy b: $|\mathrm{SPD}_b| - |\mathrm{SPD}_0|$. Negative values thus indicate a lower admission rate gap through the intervention

Table 6: Difference in absolute SPD relative to ideal policies. Lower values are better.

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	0.0233	0.0644	0.0230	0.0645
	Historical - 3 years	0.0196	0.0580	0.0203	0.0586
	Historical - 5 years	0.0202	0.0588	**0.0193**	0.0570
	Predictive - 50 sets	0.0180	**0.0542**	0.0194	**0.0549**
	Predictive - 200 sets	**0.0179**	**0.0542**	0.0196	0.0551
2017	Historical - 1 year	0.0169	0.0504	0.0171	0.0488
	Historical - 3 years	0.0146	0.0452	0.0143	0.0438
	Historical - 5 years	0.0156	0.0479	**0.0139**	**0.0416**
	Predictive - 50 sets	**0.0142**	**0.0451**	0.0151	0.0470
	Predictive - 200 sets	0.0143	**0.0451**	0.0147	0.0463

– which is desirable. In general, we can see that the values suggested through more application sets again exhibit less variance.

Table 7: Difference in absolute SPD relative to no intervention. Lower values are better.

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	0.0036	0.0378	0.0010	0.0373
	Historical - 3 years	0.0003	0.0309	-0.0013	0.0295
	Historical - 5 years	0.0010	0.0294	**-0.0022**	0.0309
	Predictive - 50 sets	-0.0013	0.0168	-0.0019	0.0268
	Predictive - 200 sets	**-0.0014**	**0.0167**	-0.0017	**0.0267**
2017	Historical - 1 year	0.0014	0.0339	-0.0013	0.0383
	Historical - 3 years	**-0.0005**	0.0269	-0.0034	0.0266
	Historical - 5 years	0.0005	0.0234	**-0.0037**	0.0270
	Predictive - 50 sets	-0.0003	**0.0124**	-0.0020	**0.0210**
	Predictive - 200 sets	-0.0002	0.0126	-0.0023	0.0215

The reason for this lies in the nature of the predictive approach which is more conservative in its suggestions. To see this, we compare the number of bonus points given to each program under the different strategies (see Table 8). What is evident is that the more application sets a strategy bases its suggestions on, the smaller the proposed bonus values become and the less variance they show across all programs. The bonus values suggested by the predictive approaches are thus closest to 0 and vary the least. The ideal policies for the same year are much higher. Despite its similar range in bonus values, the previous analysis has shown that the strategy based on last year's policies performs worse than the other strategies. This underlines the need for a more conservative design strategy.

Table 8: Comparison of bonus points for design strategies in 2016 and 2017.

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	2.32	6.95	2.56	6.93
	Historical - 3 years	1.61	3.83	1.92	4.77
	Historical - 5 years	1.53	3.19	2.08	4.56
	Predictive - 50 sets	0.96	2.96	1.31	3.96
	Predictive - 200 sets	0.92	2.63	1.28	3.93
	Ideal	2.19	6.18	2.40	6.39
2017	Historical - 1 year	2.31	6.33	2.49	6.49
	Historical - 3 years	1.85	4.06	1.90	4.58
	Historical - 5 years	1.50	3.25	1.77	4.28
	Predictive - 50 sets	0.97	2.70	1.25	3.96
	Predictive - 200 sets	0.95	2.66	1.24	3.94
	Ideal	1.76	6.24	2.21	6.78

5.2.2 Policies for Programs with Consistent Inequalities

We observe in Table 7 that some design strategies have averages above 0 and that the SDs are large compared to the means. At times the design strategies therefore increase the difference in admission rates compared to no policy implementation. This is largely due to the unpredictability of admission rates that Section 3.5 demonstrated.

In practice, affirmative action policies would not be deployed for all programs, but only sparingly for programs whose admission rates are consistently unequal. Therefore, in the following, we focus on programs that fulfill three conditions: (i) their admission rates were unequal for all of the three most recent years, (ii) the differences always negatively affected the same subgroup, and (iii) the admission rates were strongly unequal for two out of the three years. This filtering results in 9 and 12 programs to which a gender policy is applicable in 2016 and 2017, respectively. Income policies are applied to 34 programs in 2016 and 29 programs in 2017.

Tables 9, 10, 11 and 12 report the same measure as seen in the previous section, but only for the selected programs. The findings are similar to what we had previously observed for all programs. The predictive policy suggestions again exhibit lower variance – with the exception of the income policies suggested for 2017. In this case, the predictive policies lead to a notably larger error in the difference in admission rates (see Table 11). Compared to no intervention, the predictive policies on average still reduce the difference in admission rates (see Table 12).

Figure 5.1 illustrates the findings from Table 12. As we can see the filter left only a few cases where the absolute SPD is increased through the intervention. For

Table 9: Error in objective function relative to ideal policies. Smaller values are better.

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	0.80	1.16	2.36	2.06
	Historical - 3 years	1.01	1.10	1.88	1.91
	Historical - 5 years	1.68	2.21	**1.50**	1.69
	Predictive - 50 sets	**0.35**	**0.45**	1.91	1.79
	Predictive - 200 sets	0.39	0.48	1.87	**1.68**
2017	Historical - 1 year	1.31	1.88	1.79	1.63
	Historical - 3 years	1.02	1.81	**1.36**	1.62
	Historical - 5 years	1.22	1.20	1.63	**1.57**
	Predictive - 50 sets	**0.84**	1.16	2.28	2.12
	Predictive - 200 sets	0.91	**1.15**	2.14	1.97

Table 10: Difference in utility loss relative to ideal policies. Lower values are better.

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	-0.0467	**0.1491**	0.2732	1.0825
	Historical - 3 years	-0.019	0.2618	0.3094	1.1122
	Historical - 5 years	0.0421	0.3599	0.1482	0.8849
	Predictive - 50 sets	**-0.1034**	0.1541	-0.1447	0.8949
	Predictive - 200 sets	-0.0723	0.1931	**-0.1815**	**0.8404**
2017	Historical - 1 year	-0.0913	0.3613	**-0.6402**	**1.2614**
	Historical - 3 years	0.0378	0.2993	-0.4874	1.3413
	Historical - 5 years	-0.0322	0.3798	-0.5864	1.3225
	Predictive - 50 sets	**-0.138**	**0.281**	-0.5646	1.4039
	Predictive - 200 sets	-0.1197	0.2985	-0.5826	1.3181

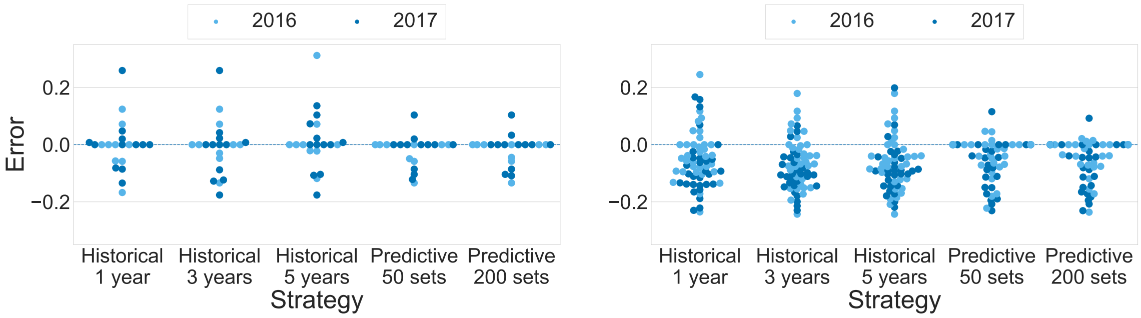
these remaining cases, it is important to consider whether the policy has a positive or negative effect on the students from the underrepresented group. In order to evaluate this, Figure 5.2 shows whether an increase in the absolute SPD occurs in favor of the underrepresented group. Negative values indicate that the intervention had a positive effect on the underrepresented group. We can thus see that an increase in the SPD to the disadvantage of the underrepresented group can almost entirely be avoided, in particular with the predictive strategies. Where this is not the case, a "flip" in the overrepresented group occurs between years: The group that has been in the majority for at least three years is in the minority in the next year.

Table 11: Difference in absolute SPD relative to ideal policies. Lower values are better.

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	0.0368	0.0539	0.0744	0.0635
	Historical - 3 years	0.0447	0.0504	0.0562	0.0524
	Historical - 5 years	0.0711	0.0945	**0.0484**	**0.0519**
	Predictive - 50 sets	**0.0196**	**0.0255**	0.0735	0.0637
	Predictive - 200 sets	0.0202	0.0258	0.0734	0.0643
2017	Historical - 1 year	0.0608	0.0788	0.0867	0.0558
	Historical - 3 years	0.0428	0.0702	**0.0658**	**0.0511**
	Historical - 5 years	0.0544	**0.0480**	0.0792	0.0618
	Predictive - 50 sets	**0.0424**	0.0561	0.1016	0.0869
	Predictive - 200 sets	0.0446	0.0555	0.0972	0.0851

Table 12: Difference in absolute SPD relative to no intervention. Lower values are better.

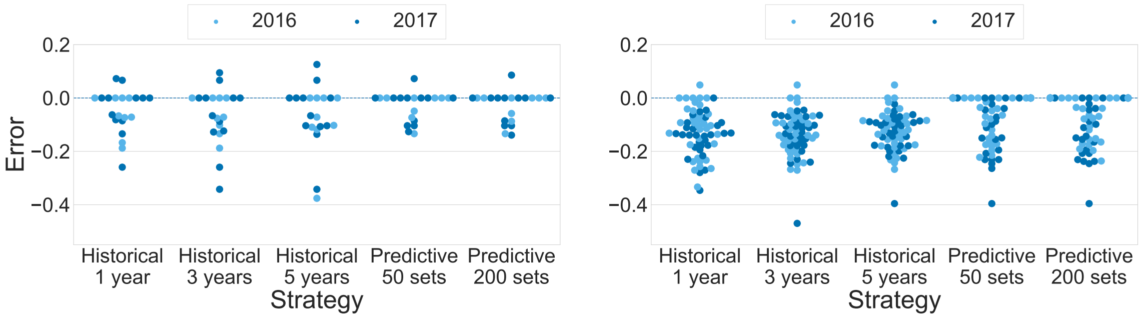
		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	-0.0096	0.0777	-0.0332	0.0877
	Historical - 3 years	-0.0017	0.068	-0.0514	0.0918
	Historical - 5 years	0.0247	0.1116	**-0.0592**	0.0937
	Predictive - 50 sets	**-0.0268**	0.0438	-0.0341	0.0612
	Predictive - 200 sets	-0.0263	**0.0435**	-0.0342	**0.0586**
2017	Historical - 1 year	0.0028	0.0922	-0.0774	0.1000
	Historical - 3 years	-0.0153	0.1069	0.0983	**0.0704**
	Historical - 5 years	-0.0037	0.0858	**-0.0849**	0.0827
	Predictive - 50 sets	**-0.0156**	0.0586	-0.0625	0.0822
	Predictive - 200 sets	-0.0135	**0.0575**	-0.0670	0.0817



(a) gender

(b) income

Figure 5.1: Difference in absolute SPD relative to no intervention. Lower is better.



(a) gender

(b) income

Figure 5.2: Difference in SPD relative to no intervention from the perspective of the subgroup that would be disadvantaged if no policy was implemented. Lower is better.

6 Policy Design Strategies for Multiple Sensitive Attributes

Section 5 introduced policy design strategies for a single sensitive attribute. This section discusses how these design strategies can be expanded to cover cases of affirmative action policies for multiple sensitive attributes. We will refer to policies for multiple sensitive attributes as *intersectional* policies.

A program might, for example, find that both low-income and female students are consistently disadvantaged and thus want to create a bonus point policy for each group. Such an extension might seem trivial at first: The program could simply first design a bonus policy for gender and then *independently* design a bonus policy for income with the previously introduced methods. In the case that a student belongs to the disadvantaged gender and disadvantaged income group, both bonuses could be added up. We refer to this strategy as independent optimization.

Upon further examination of this strategy, however, it becomes clear that this does not lead to the ideal combination of bonus point policies. The independent optimization of bonus policies minimizes Eq. 2 for a single sensitive attribute and does not take other attributes into account. It is thus unlikely that a policy designed to equalize the admission rates of women and men also equalizes the admission rates of low- and high-income students. This section hence demonstrates how policies can be optimized *jointly*.

6.1 Policy Design

We first expand the single-attribute objective function (see Eq. 2) to multiple sensitive attributes (§6.1.1). The design strategies that were described in Section 5 are then adapted, so that they optimize this function in an efficient manner (§6.1.2-§6.1.3).

6.1.1 Problem Definition

As proposed in [64], Eq. 3 extends Eq. 2 to multiple sensitive attributes in the following way: For every sensitive attribute $A_1, A_2, ..., A_m$, a bonus policy is to be implemented. Let b mark the combined application of all these bonuses $b_1, b_2, ..., b_m$. For the matching of students and programs produced by b, $\text{SPD}_{A_i,b}$ represents the difference in admission rates between the subgroups of sensitive attribute A_i. μ_b denotes the utility of the matching. μ_0 marks the utility when no bonus points are given. We assume that if a student is part of more than one disadvantaged group, the bonuses for these groups are added up.

$$o_b = (\mu_0 - \mu_b) + \sum_{i=1}^{m} \lambda_i \cdot |\mathrm{SPD}_{A_i,b}|, \quad \lambda_i \geq 0. \tag{3}$$

6.1.2 Extension of Design Strategies

Optimizing the changed objective function for a single application set is straightforward: Eq. 3 is simply evaluated for different combinations of bonus policies. The combination that minimizes the objective function is the optimal one. However, recall that most design strategies proposed in Section 5 base their suggestions for a policy on multiple application sets. We therefore have to decide which combination of bonus policies should be considered ideal based on the objective function values of multiple application sets. This is achieved by first evaluating Eq. 3 on each application set for a given combination of bonus policies. The average of these evaluations is the resulting objective function value for this combination of bonus policies and what is to be minimized.

6.1.3 Improved Efficiency through Bayesian Optimization

Eq. 3 lets us evaluate given combinations of bonus policies $b_1, b_2, ..., b_m$. However, it is unclear which combinations of policies should be tested and how those combinations are chosen. The optimization task is thus finding the combination of bonus policies that minimizes Eq. 3.

The search for the ideal combination of bonus policies can, similar to Section 5, be automated. Grid search – the evaluation of all combinations of possible policies in a predefined search space – is one of the simplest approaches to finding the ideal bonus policies. However, if k bonus policies are to be tested for each sensitive attribute, jointly optimizing the policies for m sensitive attributes would require the evaluation of k^m policy combinations. The number of policy combinations that have to be tested is thus exponential in the number of sensitive attributes. While such a grid search approach is easy to implement, it quickly becomes infeasible since testing a given combination of bonus policies requires (1) applying these bonuses to the student pool, (2) matching students with programs based on these new scores, and (3) evaluating Eq. 3 for the resulting matching.

A more efficient approach is the evaluation of only a predefined number of policy combinations. The simplest way of choosing these policy combinations is random sampling from the search space [12]. While this random approach is more efficient than the grid search approach, it has the drawback of not learning from the tested policy combinations. Consequently, if the number of tested policy combinations is limited, random search tends to be outperformed by more advanced hyperparameter optimization algorithms [14, 51].

Bayesian optimization [70] is such an advanced hyperparameter optimization algorithm. It was created to automatically and efficiently search for the hyperparameters that minimize a machine learning model's loss function [14]. Besides the loss function of a machine learning model, Bayesian optimization can be used to optimize any objective function for which no analytical solution is known [35]. It is particularly useful if the evaluation of this function is time-consuming [35]. In our case, we can thus use it to minimize Eq. 3. Similar to the random search, Bayesian optimization is carried out over a predefined number of iterations. However, different from the random search that does not remember the previous evaluations of parameter settings, Bayesian optimization is an automatic sequential optimization algorithm that learns from the sequence of previous evaluations: The evaluation history influences which parameter setting is suggested to be evaluated in the next iteration [14]. This is done by learning a function that approximates the objective function and is faster to evaluate than the true objective function [14]. In each iteration of the Bayesian optimization algorithm, a second function, which is referred to as the acquisition function, determines which parameter setting should be evaluated next [52]. This acquisition function balances the exploration of new parameter settings with the exploitation of variations of already tested parameter settings that produced good results.

Recall from Section 5 that we encode bonuses for male and high-income students as negative bonuses for female and low-income students, respectively. This simplifies the application of Bayesian optimization as the search space simply consists of one parameter for each sensitive attribute. In our case, the search space consists of a gender bonus and an income bonus, both of which can be negative. Depending on the library employed for the optimization, the search for each parameter can then, for example, be restricted to a certain range or be guided by a prior distribution.

Given a combination of bonus policies, we evaluate Eq. 3 for the resulting matching. After a predefined number of iterations, the combination of bonus policies that produced the lowest objective function value would be seen as the optimal one. However, it might happen that no combination of bonus policies performs better than no policy intervention. A simple example of this is a program that does not receive any applications. In such a case, all policy interventions result in the same matching and thus in the same objective function value. Under these circumstances, no bonus points should be awarded. The Bayesian optimization might, however, pick any combination of bonus points as they all lead to the same minimal objective function value. Therefore, comparing the lowest found objective function value to the value resulting from no intervention is essential. As the Bayesian optimization process may not even have tested the case of no intervention, we thus additionally evaluate this case once the Bayesian optimization is completed. In the case that the lowest found objective function is equal to the value resulting from no intervention, the suggested policy is not to intervene.

6.2 Experimental Evaluation

We begin by experimentally testing our assumption that the search for policies for multiple attributes requires a joint optimization (§6.2.1). To this end, we employ Bayesian optimization as previously described and select the combination of bonus policies that leads to the lowest value for Eq. 3. We compare these outcomes to the gender and income policies that have been optimized independently on Eq. 2. We then apply the design strategies introduced in Section 5 to the design of intersectional policies (§6.2.2) for a selection of programs that we expect to profit from such policies.

For each of these strategies, the suggested policies are found by running a Bayesian optimization with 1 000 steps. Note that while Bayesian optimization attempts to approximate the objective function, i.e., Eq. 3 in our case, the output is typically only the parameter setting which optimizes this function. With the library that we use in the implementation, `hyperopt` [13], the learned approximate objective function is therefore not accessible. The library was, however, chosen as it allows for the placement of a prior over each search space parameter. In this experimental evaluation, the prior for the bonus of each sensitive attribute is set to a normal distribution with mean 0 and SD 20. While we could reproduce the interval of $\{0, 1, ..., 120\}$ – which we evaluated in Section 5 – through a uniform distribution, a normal distribution allows us to place a higher weight on bonus policies which we expect to lead to better objective function values. The mean of 0 ensures that bonuses for both subgroups have an equal chance of being tested. The SD of 20, on the other hand, guides the search to fairly small bonus points, which Section 5 has shown to have beneficial effects. As in Section 5, we set $\lambda = 23$ for the sensitive attribute gender and $\lambda = 28$ for income in 2016 and 2017.

6.2.1 Necessity of Joint Optimization

This section empirically compares the objective function values resulting from an independent optimization and a joint optimization of bonus policies in the years 2016 and 2017. We show that with 1000 iterations of Bayesian optimization the joint optimization is almost guaranteed to do at least as well as the independent optimization.

In both years, the ideal combination of policies was found within 250 iterations for most programs. In 2016, after 1000 iterations, 19% of the programs had a better objective function value with the joint optimization than with the independent optimization. The joint optimization only produced a worse result for one program, indicating that more iterations would have been necessary in this case. In 2017, the joint optimization suggested policy combinations with a better objective function value in 15% of the cases and never a worse one. We therefore conclude that 1000 steps are sufficient for the optimization of the objective function in the scope of this work.

In a majority of cases, the joint optimization reaches a better objective function value because the resulting difference in admission rates is lower. The utility loss, on the other hand, is oftentimes higher for the joint optimization: In 2016 and 2017, 11-14% of the programs had a smaller utility loss with the independent optimization. This phenomenon can be understood by considering the effect of a random divergence from the optimal combination of bonus policies. Recall that the utility loss is 0 when no bonus points are given. Low utility losses can thus be achieved by policies that give few bonus points. Therefore, a divergence from the optimal policies is likely to reduce the utility loss if it decreases the number of bonus points given. Reducing the differences in admission rates is, however, more difficult as giving more bonus points does not guarantee smaller gaps in admission rates – it might favor the disadvantaged subgroup too strongly. The independent optimization thus tends to trade improved utility for an increased disparity in admission rates compared to the joint optimization.

In order to make this finding more tangible, we look at a concrete example: the prestigious engineering program with unequal admission rates that we already know from Section 1. Without an intervention, the difference in the admission rates of men and women is 14 percentage points in 2017. Moreover, low-income students' probability of being accepted is 21 percentage points lower than that of high-income students. When the gender and income policies are independently optimized, but applied at the same time, this results in a difference in admission rates of 2.6 percentage points for gender and 2.5 for income. While the differences are already notably lower, a joint optimization leads to a difference of 0.2 percentage points for gender and 0.1 for income. It thus almost perfectly equalizes admission rates. While the differences might already appear sufficiently reduced by the independent optimization, this is only the case because we have full access to the students' and programs' data. The reduction would likely be far lower if the policies were suggested based on historical data or generated application sets. As our policy design strategies can only aim at getting close to the ideal policies, setting the results of the independent optimization as the best achievable policies risks suggesting policies that are far from the true ideal combination of policies.

As an independent optimization of bonus policies is thus not sufficient, we will not just combine the independently optimized bonus policies, but optimize them jointly through Bayesian optimization.

6.2.2 Policies for Programs with Consistent Inequalities

Having demonstrated the need for joint optimization, we now apply Bayesian optimization as previously described to the design strategies introduced in Section 5. For practical reasons regarding the availability of computing resources, we only evaluate the predictive policy on 50 application sets as opposed to the 200 application sets used in Section 5. Due to the time-consuming nature of these extended strategies, we also only apply them to programs that show consistently unequal admission rates for both sensitive attributes, i.e., gender and income. Programs are filtered in the same way as in Section 5.2.2. As there are fewer programs that consistently exhibit (strongly) unequal admission rates with respect to both gender and income, this filtering only leaves 3 and 4 data points for the years 2016 and 2017, respectively.

Table 13 shows the mean error and its SD for the objective function as well as the objective function's two components, utility and disparity. In 2017, we find smaller errors and lower variance for strategies using more application sets, which corroborates the findings from Section 5. However, this is not the case for 2016 where the mean error and its SD tend to be higher for the predictive strategy.

Table 13: Error in objective function and difference in its two components, utility loss and disparity, relative to ideal policies. Lower values are better.

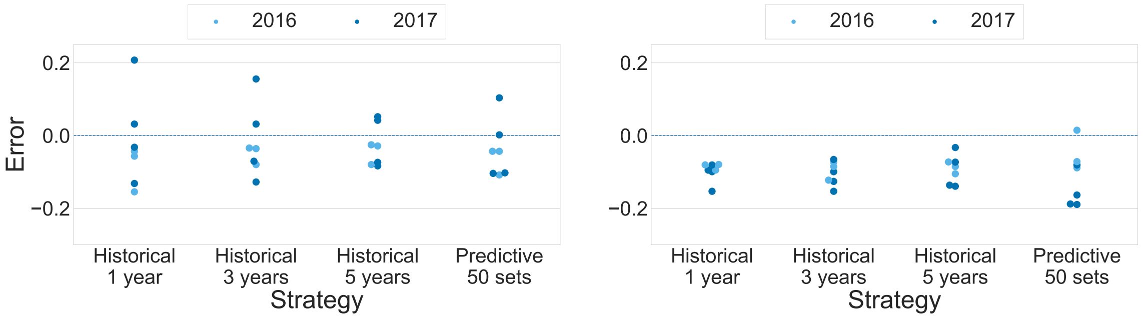
		Objective		Utility loss		Disparity	
	Strategy	Mean	SD	Mean	SD	Mean	SD
2016	Historical - 1 year	**1.08**	0.46	0.45	**0.44**	**0.63**	**0.10**
	Historical - 3 years	1.22	0.28	0.03	0.64	1.19	0.76
	Historical - 5 years	1.38	**0.12**	**-0.10**	0.56	1.48	0.54
	Predictive - 50 sets	2.97	1.77	0.87	0.71	2.11	1.71
2017	Historical - 1 year	3.14	1.56	-0.89	1.18	4.03	1.07
	Historical - 3 years	2.46	1.11	-0.97	0.69	3.42	0.96
	Historical - 5 years	2.42	1.09	**-1.14**	**0.52**	3.56	1.31
	Predictive - 50 sets	**1.47**	**0.91**	-0.20	**0.52**	**1.67**	**0.83**

Table 14 compares the differences in gender and income admission rates between the policies suggested by the different design strategies and the case of no intervention. The results show that in spite of the higher errors observed in Table 13 for 2016 the disparities in admissions are on average still reduced.

Table 14: Difference in absolute SPD relative to no intervention. Lower values are better.

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical - 1 year	**-0.0847**	0.0496	-0.0849	**0.0070**
	Historical - 3 years	-0.0499	**0.0211**	**-0.0933**	0.0212
	Historical - 5 years	-0.0445	0.0249	-0.0875	0.0134
	Predictive - 50 sets	-0.0648	0.0306	-0.0485	0.0452
2017	Historical - 1 year	0.0189	0.1235	-0.1070	**0.0275**
	Historical - 3 years	-0.0027	0.1077	-0.1110	0.0324
	Historical - 5 years	-0.0157	**0.0628**	-0.0953	0.0447
	Predictive - 50 sets	**-0.0251**	0.0859	**-0.1553**	0.0443

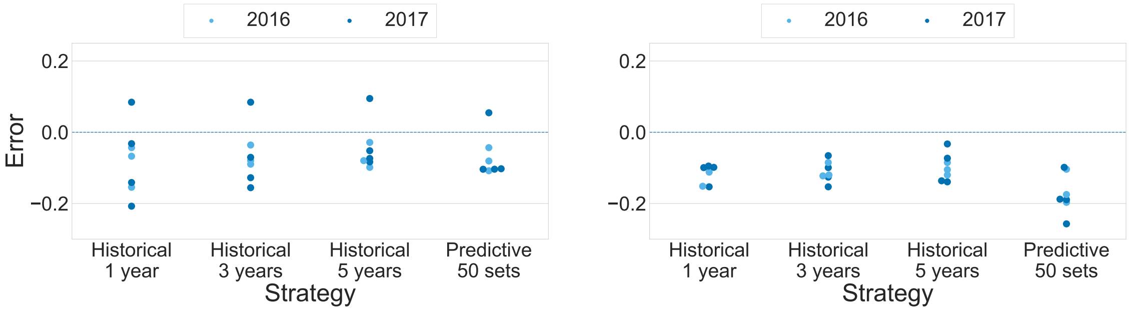
This is reflected in Figure 6.1, which shows the difference in absolute admission rates for both sensitive attributes relative to the case of no intervention. As already seen in Section 5, Figure 6.2 illustrates whether a change in admission rates occurred in favor of the underrepresented group. Negative values indicate that the underrepresented group is favored while positive values should be avoided as they indicate a favoring of the overrepresented group. We observe that positive cases were mostly prevented.



(a) gender

(b) income

Figure 6.1: Difference in absolute SPD relative to no intervention. Lower is better.



(a) gender

(b) income

Figure 6.2: Difference in SPD relative to no intervention from the perspective of the subgroup that would be disadvantaged if no policy was implemented. Lower is better.

Overall, the design strategies are beneficial to the reduction of gaps in the admission rates. While we find some evidence of more application sets leading to better results, we also see the opposite effect. This might be attributable to the fact that we were only able to test these strategies on a small number of programs. Moreover, we had to restrict the predictive approach to 50 application sets. While the differences between 50 and 200 application sets were small in the evaluations in Section 5, it is imaginable that more application sets are required to suggest robust intersectional policies due to their more complex nature. Further evaluations are therefore needed with respect to policies for intersectional groups.

7 Conclusions

In this study, we proposed a methodology for designing affirmative action policies in a central admission system. This methodology is based on a predictive approach that generates a large number of application sets over which a range of bonus policies are evaluated. We compared this predictive approach to simpler design strategies based on averaging retrospectively optimal bonuses over 1-5 years of historical data. These strategies were then extended to enable the joint optimization of intersectional policies. When testing each design strategy on real university admissions data, we were able to show that policies that are based on a few years of historical data are more likely to over-correct the differences in admission rates. Instead of equalizing admission rates, they at times increase the difference in admission rates. Our proposed predictive approach mostly avoids this pitfall through more conservative suggestions.

7.1 Limitations

There are two core assumptions that we made in this study: First, we assumed that the (unknown) applicant pool for which we want to implement a bonus policy is reasonably close to an average applicant pool; and second, we assumed that the goal of the affirmative action policy is the increase of the representation of the disadvantaged group through the equalization of the subgroups' admission rates.

If the first assumption does not hold and the next year's applicant pool is an outlier in the data, our predictive approach is unlikely to suggest effective policies. However, if the year is an outlier compared to the last year, too, it is unlikely that any of the simpler strategies based on historical data will suggest better policies, so the predictive strategy would likely still be preferable.

The second assumption means that our choice of fairness metric aims to avoid the case where the underrepresented subgroup is accepted at a higher rate than the over-represented subgroup. A higher admission rate of the underrepresented subgroup might, however, be desirable if the representation of the subgroups is to be increased without regard to admission rates. In this case, the metric used to measure disparity in Eq. 2 would have to be adapted.

Additional limitations stem from the data used for this case study. Its main limitation is the small number of programs for which the application of affirmative action policies is beneficial. This restricted the number of programs on which the introduced design strategies were tested and thus the reliability of the findings.

Another limitation with respect to data is the categorization of students as low- and high-income. As only household income categories and the number of household members were available, incomes were not easily comparable between students. Our chosen approach for splitting students into two socio-economic groups is limited in that it only compares the income per person within the application set, but not to the general population. As low-income students are less likely to attend university (see, e.g., [69]), it is improbable that they are sufficiently represented in the dataset. Labeling someone with an income below the median income of the dataset as "low-income" thus constitutes a fairly broad assignment of this label. Students that would actually be categorized as low-income compared to the general Chilean population are a subgroup of the low-income students in this study. We assume that this subgroup would have even lower admission rates which would increase the overall need for affirmative action policies. Having access to a more precise socio-economic classification would thus – at least to some extent – address the previously mentioned limitation by increasing the number of programs on which the design strategies can be tested.

7.2 Practical Implications

At first, it might seem questionable if predictive methods can help in designing robust bonus policies. One might assume that university administrators can easily determine policies that will work well for the next round of applications. However, we have shown that it is difficult to foresee how admission rates will change in the next year. Additionally, we have demonstrated that bonus policies interact in potentially unforeseeable ways and have to be optimized jointly. Given the number of possible policies, decision-makers by themselves cannot possibly assess the effect that each policy or combination of policies might have.

Our analysis shows that historical data – if used – has to be aggregated over several years in order to avoid the implementation of policies that have unexpected effects. Moreover, our experiments generally support our hypothesis that predictive methods can further reduce the risk of creating adverse effects as policies can be tested on more data. In practice, however, we have to consider that the predictive approach is more costly as its implementation is more complex than the strategies that are based on historical data. In light of this, we conclude that while the predictive strategy tends to be more robust than the simpler approaches, a simpler approach based on sufficient historical data (e.g., the last five years) might in practice be preferable. Predictive methods are, however, advantageous if historical data is only available for a few years or if only aggregate statistics are accessible.

We recommend applying either one of these design strategies under the supervision of human decision-makers. In this case, the decision-makers should be university administrators that know the programs for which bonus policies are to be implemented.

7.3 Future Work

Section 6 extended Section 5 to the case of intersectional bonus policies. A similar extension is possible to cover cases where multiple programs plan on implementing bonus policies in the same year and wish to coordinate their efforts. Coordination is important as potential interactions between policies have to be considered in their design. Adapting the objective function would again allow for a joint optimization (see [64] for a possible extension of Eq. 2 for this case).

Practitioners might also profit from the comparison of different types of affirmative action policies. While this study only explored different parameters for bonus policies, in practice, it could be interesting to compare the effects of the parameters of other types of policies. One could, for example, examine how giving different weights to standardized test scores compares to different bonus policies. Optimization techniques could be applied to optimize the objective function over the search space of the parameters of various types of affirmative action policies.

Another interesting aspect that is left for future research is the effect of the announcement of affirmative action policies on the application behavior. Existing research (see, e.g., [7, 33, 62]) suggests that the existence of affirmative action policies might encourage students from disadvantaged groups to apply in higher numbers. It is therefore imaginable that lower numbers of bonus points are sufficient to achieve the effects we observed in our experiments, which is another reason for preferring the more conservative predictive design strategies.

7.4 Final Thoughts

> *Talent is equally distributed,*
> *opportunity is not.*
>
> Leila Janah

Finally, let us consider this work in the broader context of social justice in education. With the emergence of the field of algorithmic fairness, the role of computing itself in the strive for social change has become a topic of discussion. In particular, computer science's role in reinforcing societal inequalities is seen as a danger to structural change. This reinforcement can in particular be seen in predictive models that learn from data from an unequal society. If such an issue is uncovered, its computational solution is oftentimes incremental. Instead of addressing the root cause of the issue, i.e., societal inequalities, the algorithm itself is incrementally improved [3, 40, 47]. Green thus instead argues for "holistic responses that promote egalitarian structures and outcomes in both the short and long term" [41, p. 595]. We see the application of computational methods to the design of affirmative action policies as at least part of such a holistic response. However, this alone is still only an incremental effort and does not address the root cause of the problem: the differences in opportunity that Leila Janah's quote at the beginning of this section pointed out. Affirmative action policies do not address this issue, they are only a remedy for the status quo. Abebe et al. [3] specifically discuss the issue of university admissions and note that equalizing admissions should not be an excuse not to address deeper issues, such as the need for more resources at schools that are primarily attended by low-income students. The prestigious engineering program with unequal admission rates mentioned in Section 1 is an example of this. Optimizing the parameters of a bonus policy can reduce this inequality in the next years, but cannot substitute other measures on the way to our main goal: increased social inclusion.

References

1 Atila Abdulkadiroğlu. College admissions with affirmative action. *International Journal of Game Theory*, 33(4):535–549, 2005.

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

3 Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 252–260, New York, NY, USA, 2020. ACM.

4 Kehinde F Ajayi. School choice and educational mobility: Lessons from secondary school applications in Ghana. *Institute for Economic Development Working Paper*, 259, 2013.

5 Hayri Alper Arslan. Preference estimation in centralized college admissions from reported lists. *Available at SSRN 3126391*, 2019.

6 Verne R Bacharach, Alfred A Baumeister, and R Michael Furr. Racial and gender science achievement gaps in secondary education. *The Journal of genetic psychology*, 164(1):115–126, 2003.

7 Loukas Balafoutas and Matthias Sutter. Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, 335(6068):579–582, 2012.

8 Cissy J Ballen, Shima Salehi, and Sehoya Cotner. Exams disadvantage women in introductory biology. *PLoS One*, 12(10), 2017.

9 Peter Barbrook-Johnson, Jennifer Badham, and Nigel Gilbert. Uses of Agent-Based Modeling for Health Communication: the TELL ME Case Study. *Health Communication*, 32(8):939–944, 2017.

10 María Cecilia Bastarrica, Nancy Hitschfeld, Maíra Marques Samary, and Jocelyn Simmonds. Affirmative action for attracting women to STEM in Chile. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, GE '18, pages 45–48, New York, NY, USA, 2018. ACM.

11 Rachel K E Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

12 James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305, 2012.

13 James Bergstra, Daniel L K Yamins, and David Daniel Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages I–115–I–123. JMLR.org, 2013.

14 James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.

15 Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

16 Marianne Bertrand, Rema Hanna, and Sendhil Mullainathan. Affirmative action in education: Evidence from engineering college admissions in India. *Journal of Public Economics*, 94(1-2):16–29, 2010.

17 Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larrañaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233, 2015.

18 Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

19 Paul Brest and Miranda Oshige. Affirmative action for whom? *Stanford Law Review*, pages 855–900, 1995.

20 Christopher JC Burges. From RankNet to LambdaRank to LambdaMART: An overview. *Learning*, 11(23-581):81, 2010.

21 Cristian Cabalin. Neoliberal education and student movements in Chile: Inequalities and malaise. *Policy Futures in Education*, 10(2):219–228, 2012.

22 Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

23 Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 129–136, New York, NY, USA, 2007. ACM.

24 Quinn Capers IV, Daniel Clinchot, Leon McDougle, and Anthony G Greenwald. Implicit racial bias in medical school admissions. *Academic Medicine*, 92(3): 365–369, 2017.

25 David Card and Alan B Krueger. Would the elimination of affirmative action affect highly qualified minority applicants? Evidence from California and Texas. *ILR Review*, 58(3):416–434, 2005.

26 Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

27 Yan Chen and Onur Kesten. Chinese college admissions and school choice reforms: A theoretical analysis. *Journal of Political Economy*, 125(1):99–139, 2017.

28 Faye J Crosby, Aarti Iyer, and Sirinda Sincharoen. Understanding affirmative action. *Annu. Rev. Psychol.*, 57:585–611, 2006.

29 Christine SM Currie, John W Fowler, Kathy Kotiadis, Thomas Monks, Bhakti Stephan Onggo, Duncan A Robertson, and Antuela A Tako. How simulation modelling can help reduce the impact of COVID-19. *Journal of Simulation*, pages 1–15, 2020.

30 Richard Davies. Why is inequality booming in Chile? Blame the Chicago Boys. *The Guardian*, November 2019. `https://www.theguardian.com/commentisfree/2019/nov/13/why-is-inequality-booming-in-chile-blame-the-chicago-boys`.

31 DEMRE. Puntaje ranking. https://psu.demre.cl/proceso-admision/factores-seleccion/puntaje-ranking. [Accessed on 10/28/2019].

32 Kevin C Desouza and Yuan Lin. Towards evidence-driven policy design: Complex adaptive systems and computational modeling. *Innovation Journal*, 16(1), 2011.

33 Lisa M Dickson. Does ending affirmative action in college admissions lower the percent of minority students applying to college? *Economics of Education Review*, 25(1):109–119, 2006.

34 Gary W Evans. The environment of childhood poverty. *American psychologist*, 59(2):77, 2004.

35 Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

36 Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

37 Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

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

39 Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

40 Ben Green. "Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*, 2019.

41 Ben Green. The false promise of risk assessments: Epistemic reform and the limits of fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 594–606, New York, NY, USA, 2020. ACM.

42 Anthony G Greenwald and Linda Hamilton Krieger. Implicit bias: Scientific foundations. *California Law Review*, 94(4):945–967, 2006.

43 Isa E Hafalir, M Bumin Yenmez, and Muhammed A Yildirim. Effective affirmative action in school choice. *Theoretical Economics*, 8(2):325–363, 2013.

44 Isa E Hafalir, Rustamdjan Hakimov, Dorothea Kübler, and Morimitsu Kurino. College admissions with entrance exams: Centralized versus decentralized. *Journal of Economic Theory*, 176:886–934, 2018.

45 James D Hamilton. *Time series analysis*, volume 2. Princeton New Jersey, 1994.

46 Peter Hinrichs. The effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities. *Review of Economics and Statistics*, 94(3):712–722, 2012.

47 Anna Lauren Hoffmann. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7): 900–915, 2019.

48 Jessica S Howell. Assessing the impact of eliminating affirmative action in higher education. *Journal of Labor Economics*, 28(1):113–166, 2010.

49 Caroline M Hoxby and Christopher Avery. The missing "one-offs": The hidden supply of high-achieving, low income students. Technical report, National Bureau of Economic Research, 2012.

50 Mala Htun. From "racial democracy" to affirmative action: Changing state policy on race in Brazil. *Latin American Research Review*, pages 60–89, 2004.

51 Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International conference on learning and intelligent optimization*, pages 507–523. Springer, 2011.

52 Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning*. Springer, 2019.

53 Muhammad Ibrahim and Mark Carman. Comparing pointwise and listwise objective functions for random-forest-based learning-to-rank. *ACM Trans. Inf. Syst.*, 34(4), August 2016.

54 Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

55 Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

56 Toshiji Kawagoe, Taisuke Matsubae, and Hirokazu Takizawa. The skipping-down strategy and stability in school choice problems with affirmative action: Theory and experiment. *Games and Economic Behavior*, 109:212–239, 2018.

57 Jared P Keller, Kevin C Desouza, and Yuan Lin. Dismantling terrorist networks: Evaluating strategic options using agent-based modeling. *Technological Forecasting and Social Change*, 77(7):1014–1036, 2010.

58 Fuhito Kojima. School choice: Impossibilities for affirmative action. *Games and Economic Behavior*, 75(2):685–693, 2012.

59 Dharma Kumar. The affirmative action debate in India. *Asian Survey*, 32(3): 290–302, 1992.

60 Nicholas Lemann. *The big test: The secret history of the American meritocracy.* Macmillan, 2000.

61 Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

62 Mark C Long. College applications and the effect of affirmative action. *Journal of Econometrics*, 121(1):319 – 342, 2004. Higher education (Annals issue).

63 Vincent T Maphai. Affirmative action in South Africa – a genuine option? *Social Dynamics*, 15(2):1–24, 1989.

64 Michael Mathioudakis, Carlos Castillo, Giorgio Barnabo, and Sergio Celis. Affirmative action policies for top-k candidates selection, with an application to the design of policies for university admissions, 2019.

65 Patrick J McEwan. The indigenous test score gap in Bolivia and Chile. *Economic development and cultural change*, 53(1):157–190, 2004.

66 Martin I Meltzer, Inger Damon, James W LeDuc, and J Donald Millar. Modeling potential responses to smallpox as a bioterrorist weapon. *Emerging infectious diseases*, 7(6):959, 2001.

67 Francisco Meneses and Javiera Toro Cáceres. Predicción de notas en Derecho de la Universidad de Chile: ¿sirve el ranking? *ISEES: Inclusión Social y Equidad en la Educación Superior*, (10):43–60, 2012.

68 Ministerio de Educación de Chile. Bases para una política de formación técnico-profesional en Chile. informe de la comisión para el estudio de la formación técnico-profesional en Chile, 2009. Executive Summary.

69 OECD. *Reviews of national policies for education: Tertiary education in Chile*. 2009.

70 Martin Pelikan, David E Goldberg, and Erick Cantú-Paz. BOA: The Bayesian optimization algorithm. In *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation - Volume 1*, GECCO'99, pages 525–532, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

71 Dana Pessach and Erez Shmueli. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*, 2020.

72 President of the United States. Executive Order No. 10925, Section 301, 1961.

73 Sean F Reardon. The widening income achievement gap. *Educational leadership*, 70(8):10–16, 2013.

74 Alvin C Rencher and William F Christensen. Chapter 10, Multivariate regression – Section 10.1, Introduction. *Methods of multivariate analysis, Wiley Series in Probability and Statistics*, 709:19, 2012.

75 Ignacio Ríos, Tomás Larroucau, Giorgiogiulio Parra, and Roberto Cominetti. College admissions problem with ties and flexible quotas. *Available at SSRN 2478998*, 2014.

76 Richard Rothstein. The racial achievement gap, segregated schools, and segregated neighborhoods: A constitutional insult. *Race and social problems*, 7(1): 21–30, 2015.

77 Shima Salehi, Sehoya Cotner, Samira M Azarin, Erin E Carlson, Michelle Driessen, Vivian E Ferry, William Harcombe, Suzanne McGaugh, Deena Wassenberg, Azariah Yonas, et al. Gender performance gaps across different assessment methods and the underlying mechanisms: The case of incoming preparation and test anxiety. In *Frontiers in Education*, volume 4, page 107. Frontiers, 2019.

78 Lloyd Shapley and Herbert Scarf. On cores and indivisibility. *Journal of mathematical economics*, 1(1):23–37, 1974.

79 Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.

80 Steven J Spencer, Claude M Steele, and Diane M Quinn. Stereotype threat and women's math performance. *Journal of experimental social psychology*, 35(1): 4–28, 1999.

81 Rhea E Steinpreis, Katie A Anders, and Dawn Ritzke. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex roles*, 41(7-8):509–528, 1999.

82 Marta Tienda, Kevin T Leicht, Teresa Sullivan, Michael Maltese, Kim Lloyd, et al. *Closing the gap?: Admissions and enrollments at the Texas public flagships before and after affirmative action.* Office of Population Research, Princton University, 2003.

83 Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

84 Universidad de Chile. Programa de Ingreso Prioritario de Equidad de Género (PEG). `http://www.uchile.cl/portal/presentacion/asuntos-academicos/pregrado/admision-especial/96722/ingreso-prioritario-de-equidad-de-genero-peg`. [Accessed on 10/28/2019].

85 Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

86 Alexander Westkamp. An analysis of the German university admissions system. *Economic Theory*, 53(3):561–589, 2013.

87 Gonca Telli Yamamoto. University evaluation-selection: A Turkish case. *International Journal of Educational Management*, 2006.

88 Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.