Volume 9(2), 121–137. https://doi.org/10.18608/jla.2022.7463

Designing Affirmative Action Policies under Uncertainty

Corinna Hertweck[1], Carlos Castillo[2], Michael Mathioudakis[3]

Abstract

We study university admissions under a centralized system that uses grades and standardized test scores to match applicants to university programs. In the context of this system, we explore affirmative action policies that seek to narrow the gap between the admission rates of different socio-demographic groups while still accepting students with high scores. Since there is uncertainty about the score distribution of the students who will apply to each program, it is unclear what policy would have the desired effect on the admission rates of different groups. We address this challenge by using a predictive model trained on historical data to help optimize the parameters of such policies. We find that a learned predictive model does significantly better than relying on the ideal parameters for the last year. At the same time, we also find that a large pool of historical data yields similar results as our predictive approach for our data. Due to the more complex nature of the predictive approach, we conclude that a simpler approach should be preferred if enough data is available (e.g., long-standing, traditional university programs), but not for newer programs and other cases in which our predictive strategy can prove helpful.

Notes for Practice

- Machine learning can be used to design affirmative action policies that take into account changes in applicants' behaviour from one admission cycle to the next.
- Simple strategies such as computing the optimal bonus policy based on the last few years of historical data perform well if enough data is available.
- For newer programs or in the absence of historical data, a machine learning method may be preferable.
- Due to fluctuations in applicant behaviour, an affirmative action policy that has a desirable effect in one year may have an undesirable effect in another year. Evidence-based policies should be based on multiple years of data, and uncertainty should be taken into account.
- When applied in a targeted manner, affirmative action policies may lead to a moderate increase in the admission rate of underrepresented groups, while having a minimal effect on the average admission score of admitted applicants.

Keywords
Algorithmic fairness, affirmative action, predictive analytics

Submitted: 27/03/2021 — **Accepted:** 16/03/2022 — **Published:** 31/08/2022

Corresponding author [1]*Email: corinna.hertweck@zhaw.ch Address: Institute for Data Analysis and Process Design, Zurich University of Applied Sciences, Winterthur, Switzerland, Department of Informatics, University of Zurich, Zurich, Switzerland. ORCID ID: https://orcid.org/0000-0002-7639-2771*

[2]*Email: chato@icrea.cat Address: ICREA Catalan Institution for Research and Advanced Studies, Barcelona, Spain, Department of Technologies of Information and Communication, Universitat Pompeu Fabra, Barcelona, Spain. ORCID ID: https://orcid.org/0000-0003-4544-0416*

[3]*Email: michael.mathioudakis@helsinki.fi Address: Department of Computer Science, University of Helsinki, Helsinki, Finland. ORCID ID: https://orcid.org/0000-0003-0074-3966*

1. Introduction

In centralized university admission systems, students submit their applications to a central institution, which then matches the applicants with the programs offered by various universities. This matching is typically automated using an algorithm and based on grades and standardized test scores. Since grades and test scores tend to differ across demographic groups (see, e.g., Bacharach et al., 2003; McEwan, 2004; Reardon, 2013; Rothstein, 2015), this kind of system can lead to large gaps in admission rates between groups. These gaps can be reduced by affirmative action policies, which are challenging to design because policies should be announced *before* the application period begins to provide candidates with all the information they need. Because universities want to balance inclusion goals with the goal of admitting the most promising students (usually the ones with the highest test scores), it becomes even more difficult to find the ideal policy. Computational methods can be used to evaluate a large range of alternatives, much more than can be evaluated manually, and to identify effective policies.

In this paper, we develop an approach for designing robust and effective affirmative action policies for university admissions. Specifically, we consider *bonus policies*, i.e., policies that add a number of bonus points to the scores of 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* (Mathioudakis et al., 2020), as we explain in Section 4.1. The technical problem we face is to choose the right allocation of bonuses so that the policy will have a consistently beneficial effect on the admission rate of the given group. We compare approaches of varying complexity on data from university admissions in Chile.

2. Related Work

The central question of this study is how matching algorithms (Subsection 2.1) can adopt affirmative action policies (Subsection 2.2). We thus discuss the existing literature in both areas.

2.1 Matching Algorithms

In many countries, the university admissions process is centralized, meaning that students send their university applications to a central institution. This institution then matches universities or university programs with students, usually based on the preferences of both the students and the universities or university programs. In centralized admission systems, algorithms are typically deployed to handle this matching. 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-many, two-sided matching problem first described in Gale and Shapley (1962) as the *college admissions problem*. To match students with programs, Gale and Shapley (1962) proposed the deferred acceptance (DA) algorithm, which they proved to have several desirable qualities. One of these is called *stability*. 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. The admission results in Chile are publicly visible, and hence stability is particularly important (Ríos et al., 2014). Unstable matchings could lead to legal challenges if students realize that someone with a lower score has been admitted to a program that they would have preferred over their own. It thus seems likely that the matching algorithm deployed in Chile is a version of the DA algorithm. While the matching algorithm used for Chilean university admissions has not been publicly revealed, the assumption that it is DA has been confirmed by experimental results (Ríos et al., 2014).

2.2 Affirmative Action

Affirmative action policies are typically implemented as a way to counter the effect of inequalities manifested through unequal grade and test score distributions.

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 admitted minority students. Kawagoe and colleagues (2018) evaluate this strategy for DA. Further affirmative action extensions of matching algorithms can be found in the literature (Abdulkadiroğlu, 2005; Abdulkadiroğlu & Sönmez, 2003; Hafalir et al., 2013; Kojima, 2012).

Previous work lacks a discussion of how the numerical parameters required by these policies (e.g., quotas or bonus points for an underrepresented group) should be chosen optimally, particularly under uncertainty. We may thus ask whether machine learning can be used to help find these numerical parameters. This seems to be a promising approach since previous research has already used analytics to inform the design of policies in the educational context. Computational methods have, e.g., been used to evaluate how the design of courses on learning management systems affects learning goals (Lancaster et al., 2020) or to predict students' progress on computer-assisted tasks (Faucon et al., 2020).

This paper will explore whether predictive analytics can be used to design affirmative action policies. While Mathioudakis and colleagues (2020) addressed the problem of the design of affirmative action policies, their work designed policies for a given set of applications and not under uncertainty. This study, on the other hand, discusses how historical data can be used to design affirmative action policies for university admissions when the applicants are unknown.

Using the terminology of Friedler and colleagues (2016), affirmative action is for the most part consistent with the "we're all equal" (WAE) worldview. In a given situation where two groups' grades are compared, assuming WAE means assuming that differences in grade averages between groups are not caused by one group being more talented or diligent—instead these differences are seen as a product of structural inequality, e.g., lack of resources (Evans, 2004). Contrary to WAE, "what you see is what you get" (WYSIWYG) is a diametrically opposing worldview, according to which the distribution of scores accurately reflects the talent, diligence, or aptitude of the applicants (Friedler et al., 2016). If universities were to adhere to WYSIWYG, they would simply admit the students with the highest scores.

The worldview that we take on in this paper falls between the WAE and WYSIWYG worldviews, i.e., we assume that scores do reflect student aptitude to some degree—but also that they are unfavourable to certain disadvantaged groups to some extent. Under such a worldview, universities would aim to balance admitting students with high scores, on one hand, and with similar admission rates across different student groups, on the other. In this balance, they should consider that above-average test scores do not guarantee good academic performance once admitted into an university, since test scores constitute, at most, a noisy estimate. In this paper, we do not determine which worldview universities should take on—instead, we allow for the possibility of trading off worldviews in the search for a fitting bonus point policy. This is done by allowing universities to decide how much of a loss in the expected average admission score of the admitted students they are willing to accept for more equal admission rates.

3. Dataset

Before we delve into the details of policy design and evaluation, this section presents the dataset we use for experimental validation and highlights characteristics of it that are relevant to our study. We begin the section by describing this dataset of university admissions (Subsection 3.1). Next, we briefly discuss affirmative action in Chile (Subsection 3.2). Subsequently, we present the choice of sensitive attributes, i.e., the demographic groups that we consider for affirmative action policies (Subsection 3.3). Then, we explore current disparities in admission rates for the chosen attributes and possible reasons for them (Subsections 3.4–3.6). In what follows, we assume that if affirmative action policies are to be implemented for the next year, the current year highly influences this decision—therefore, because Section 4 attempts to find bonus policies for the most recent available years in our dataset (2016 and 2017), this section mainly analyzes the data for the previous years (2015 and 2016).

3.1 Data Description

We analyze anonymized data from the central admission system of Chile, which is available under a research agreement[1]. This dataset contains information about all students who applied for university programs between 2004 and 2017, as well as the available programs. The number of students and programs has increased slowly across these years, with a sharper increase in the number of programs in 2012 because multiple universities joined the central admissions process in that year. In 2017, about 60,000 students applied to about 1,500 programs.

A variety of features are available for each student. Their *graduation year* indicates that most students (81% in both 2015 and 2016) apply for university right after graduation. Their *place of residence* shows that more than a third of applicants in the dataset come from the administrative region containing the country's capital, Santiago. The average *household size* is four people, including the student. Other features in the dataset include the *high school* that students attended and information about the students' families, such as their *parents' education* and *job status*. Additionally, we have access to students' *high school grades* and *standardized test scores* on four tests: mathematics, language, natural sciences, and social sciences.

After taking these standardized tests and obtaining their results, students submit an application consisting of a ranked list of university programs (including university name and degree) to a central institution. We will refer to this ranking of programs, which is part of our dataset, as the student's *preferences*. Students can list up to 10 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 set by each university for each of their programs independently. For example, engineering programs typically place more weight on the mathematics and natural sciences test scores than on the language and social sciences test scores, while law programs may have the opposite emphasis. The central institution knows the available seats of each program and matches students to programs based on these weighted scores through the DA algorithm (Gale & Shapley, 1962; Ríos et al., 2014). For this work, we also adopt DA as the matching algorithm. We note that between 2012 and 2017 the majority of programs received fewer applicants than the number of spots they offered; however, a few programs in top institutions receive many more applicants than the vacancies they offer.

3.2 Affirmative Action in Chile

Inequality in the education system of Chile has persisted for years (Cabalin, 2012). While higher education is no longer regarded as a privilege of the elites, multiple issues prevent equal access to higher education for all students (Davies, 2019, November 13). In 2009, the Chilean Ministry of Education found that a majority of the students that choose vocational training for their upper secondary education, instead of entering a university, come from a disadvantaged socio-economic background (Ministerio de Educación de Chile, 2009). Consequently, an OECD report (OECD & World Bank, 2009) found that students from low-income

[1] Department of Educational Evaluation, Measurement and Registration (DEMRE). Data requests through https://investigador.demre.cl/.

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 high schools, e.g., a transformed version of the high school grade score average has been introduced (DEMRE, n.d.). The transformation compares each student's grades to those of students who have studied at similar schools in the past years. This measure has been shown to help better judge students' academic abilities (Meneses & Cáceres, 2012).

Additionally, several universities in the country implement their own affirmative action policies in some programs. Through these policies, for instance, in one engineering program the people who are just below the last admitted person by DA enter a waiting list, and the first women on the waiting list are admitted; the same happens with the first men on the waiting list of a social work program (Universidad de Chile, n.d.). Bastarrica and colleagues (2018) evaluated the effect of the affirmative action policy in the engineering program and found a positive effect not only on the number of admitted women but also on the number of applications received from women.

3.3 Choice of Sensitive Attributes

To consider the effects of affirmative action policies, we need a definition of the demographic groups for which such policies will be considered. Commonly used sensitive attributes for affirmative action policies are race, gender, and income (Brest & Oshige, 1995; Crosby et al., 2006). We have no data on the race or ethnicity of students, but the data contains a binary gender variable and the students' household income. Therefore we focus on designing policies for gender and income. For gender, we consider the disparities of admission rates between two groups, i.e., *women* and *men* (gender is a binary feature in this data).

Income is given as a discrete variable whose range varies from year to year. For each student in the year's applicant pool, we use the average household income per household member, calculated from the total household income and household size features. Subsequently, we transform the household income variable into a binary variable: 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 the analysis that follows, we consider the disparities of admission rates between the two groups, i.e., low-income and high-income students.

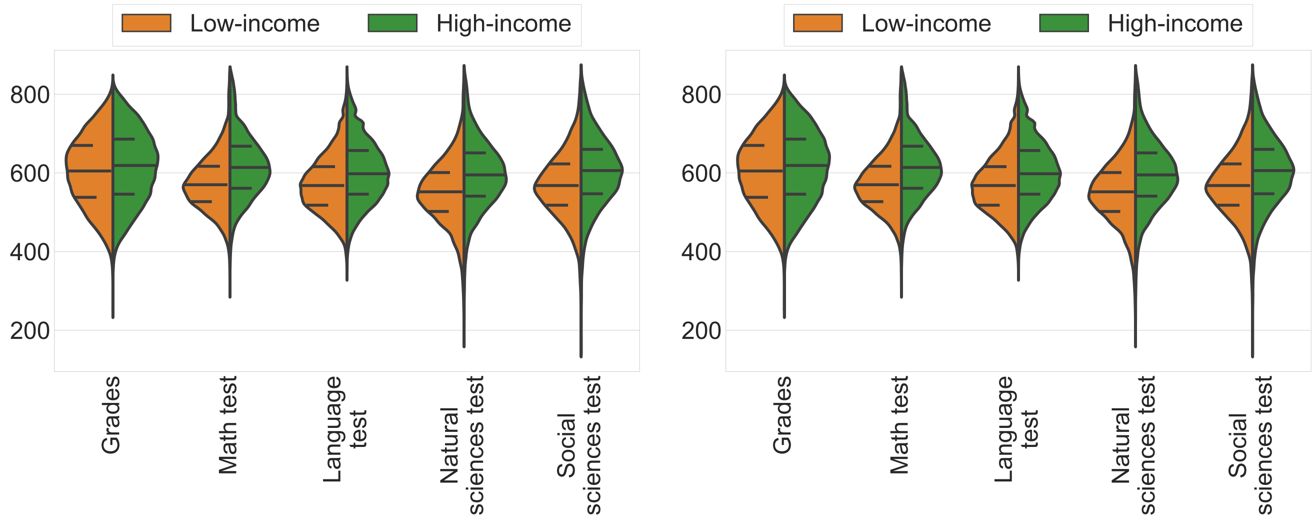
As a first observation of the properties of the two sensitive attributes, gender and income, we notice that while the share of female and male students is fairly stable between years, the share of low- and high-income students varies more (see Table 1). This is attributed to the fact that the dataset does not provide us with a continuous income distribution but rather discretized income values (i.e., for each individual, we know that their income belongs to a certain range). Depending on the year, a relatively high or low number of students share the median household-income-per-household-member, leading to a corresponding fluctuation in the number of students categorized as *high-income* or *low-income*.

Table 1. 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

3.4 Differences in Scores and Preferences

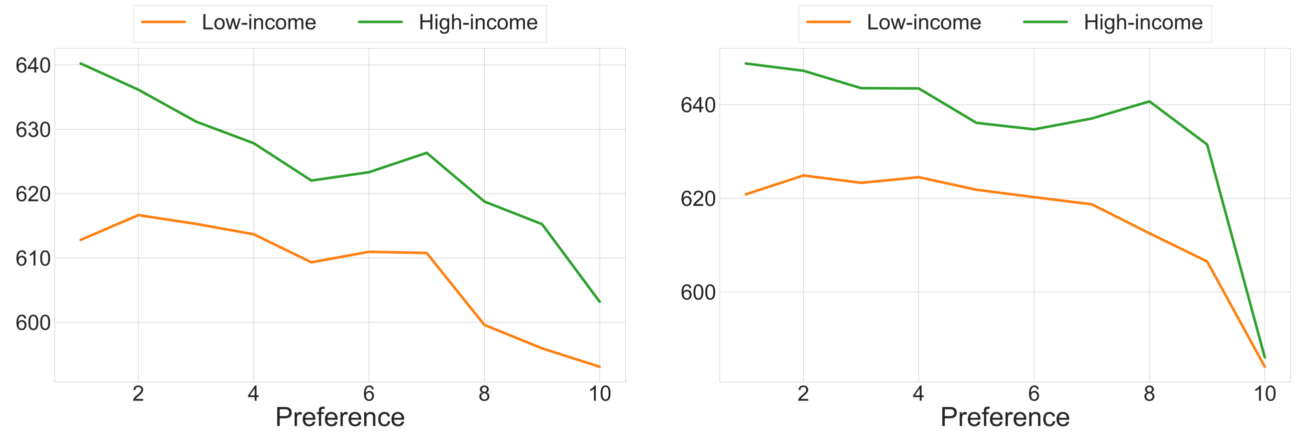
We find that there is a pronounced difference between the score distributions of the two income groups across all years, which we illustrate as announced with statistics from 2015 and 2016: high-income students have much higher scores across all standardized tests and slightly higher high school grades than low-income students (Figure 1). For instance, a high-income student with a math test score close to the median among high-income students would have been close to the top 25% if compared against low-income students. While there exist differences between genders, too, they are less pronounced in our dataset.



(a) 2015 **(b)** 2016

Figure 1. Distributions of Grades and Standardized Test Scores for Different Income Levels

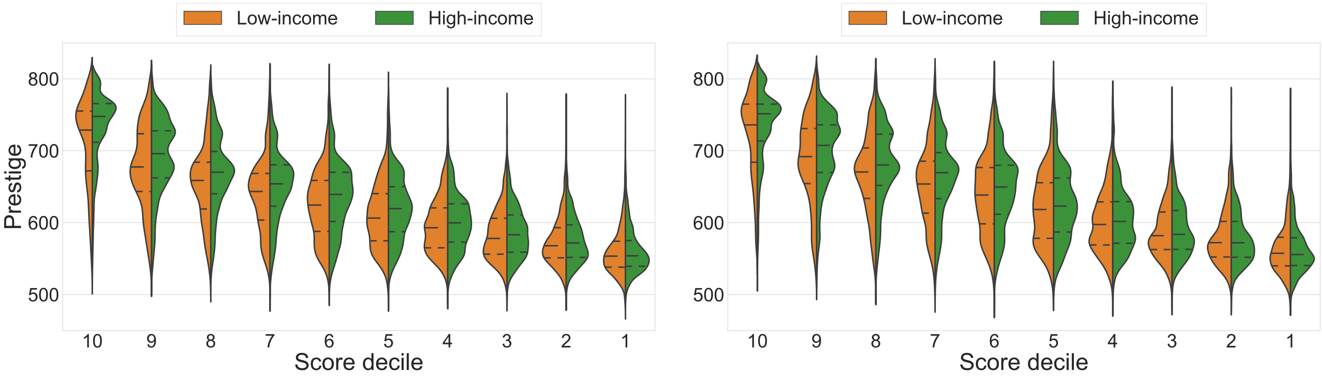
We also observe differences in the programs to which students apply. To describe these differences, we define the *prestige* of a program as the average score of its admitted students in the previous three years. We find that high-income students express preferences for programs of higher average prestige than do low-income students, particularly as their first preference (Figure 2).



(a) 2015 **(b)** 2016

Figure 2. Average Prestige of Students' Preferences across Income Levels

However, we note that high-income students tend to have higher scores than low-income students, so 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 shows, differences between low-income students and high-income students whose scores are above the median are notably lower, too. This suggests that application behaviour 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 (2012) 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, among others, 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. Average Prestige of First Choice by Income, Controlled for Score

3.5 Differences in Admission Rates

Affirmative action policies for university admissions generally aim at increasing the diversity of the student body. This idea can take on different forms, though. One concrete goal might be to equalize the admission rates between groups, where the admission rate is calculated as the number of accepted students of a particular group relative to the number of applicants from this group. Another goal would be to equalize the rates of accepted students compared to their share in the general population (and not just the applicant pool). If the applicant pool is fairly representative of the population, both goals lead to a similar outcome. If this is not the case, though, the second goal might require that the group that is underrepresented (relative to the general population) be accepted at an above-average rate (relative to the applicant pool). In this work, we will specifically consider the admission rates with respect to the applicant pool—although the methods we describe can also be used for differently defined admission rates (see Subsection 4.2 for more on this).

We measure differences in admission rates via what we will refer to as *statistical parity difference (SPD)*:

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

where $Y=1$ indicates being admitted into a program, A is a sensitive attribute (i.e., income or gender), and a marks a disadvantaged demographic group (in this work either low-income or female students). Probabilities $P(Y=1|A=a)$ and $P(Y=1|A\neq a)$ correspond to the event of admission ($Y=1$) conditional on the applicant's group ($A=a$ or $A\neq a$). Low absolute values of SPD are desired—and perfect equality of admission rates is achieved for a value of zero. As shown in Figure 4, we usually observe larger negative SPD values for the more prestigious programs, indicating differences in admission rates that place low-income students at a disadvantage.

Following thresholds recommended as defaults in tools for measuring algorithmic bias (Bellamy et al., 2018), we refer to values outside of the range $[-0.1,0.1]$ as *strongly unequal*.

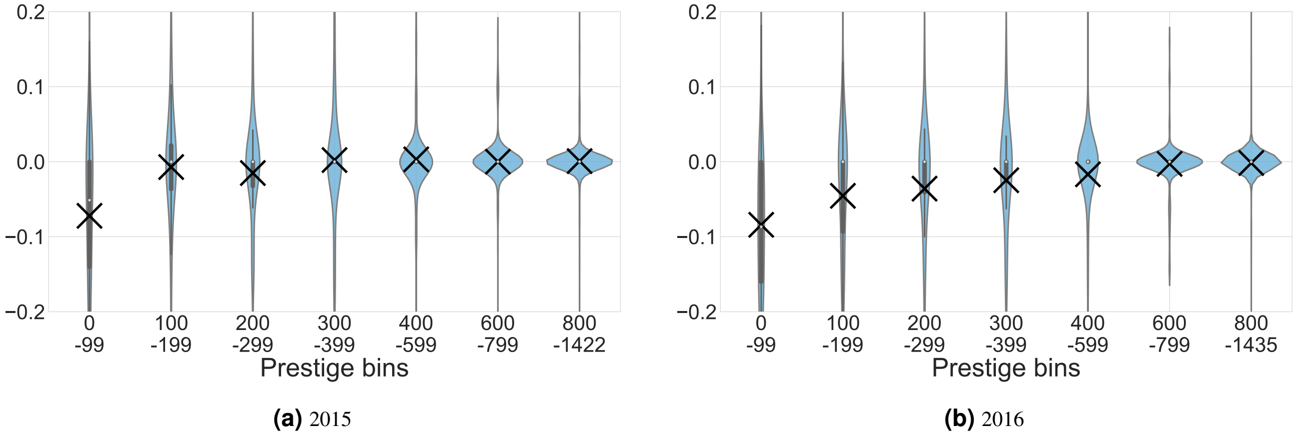
3.6 Variance in Admission Rates across Years

While the fraction of programs with strongly unequal admission rates is fairly constant across years (between 8% and 10%), there is large variation in the SPD of individual programs over the years. Because of this, designing policies to bring the SPD of programs closer to zero is challenging. In fact, we found that the SPD of one program in one year is a worse predictor for its SPD in the next year than simply predicting an SPD of zero. This is especially true for smaller programs where the composition of the applicant group looks very different between years. Thus, the applicants of the individual programs are unlikely to be a good representation of the total population. This observation is important in practice because we could be undercorrecting or overcorrecting inequalities if we assumed that differences in admission rates would not vary.

4. Policy Design

4.1 General Remarks

To reduce admission rate gaps, we consider using bonus policies, which award bonus points to students from disadvantaged groups. A bonus policy of b points concerns the applicants of a single program and simply increases the score of each applicant



(a) 2015 **(b)** 2016

Figure 4. Distribution of SPD for Income, with Crosses Marking the Mean*

*Programs have been ranked by prestige (higher to lower). Labels on the *x*-axis indicate the prestige ranks of programs in each bin. The *y*-axis corresponds to SPD values, calculated according to Equation (1).

that belongs to the group for which the policy is designed. For example, an income-related bonus policy of *b* points would add *b* points to each applicant of the program that belongs in the low-income group. Once the bonus policy has been applied, students are matched with programs by the DA algorithm based on their new scores (i.e., after the inclusion of bonuses). The application of the bonus policy generally results in a change in the admission rates of the analyzed socio-demographic groups (the fraction of applicants that are admitted, among the applicants of a given group) and the admission proportions (a.k.a. quotas, i.e., the proportion of applicants from each group among the admitted applicants).

Policies Based on Bonuses versus Quotas For this paper, we have opted to work with bonus policies, while in other cases affirmative action is implemented with policies based on quotas. Before we move forward, it is worth considering the relationship between the two.

In the setting we consider, quotas are equivalent to bonus policies in the following sense: for a fixed cohort of applications, for any bonus policy there is a corresponding quota policy that achieves the same outcome (i.e., set of admitted students), and for any quota policy there is a bonus policy that achieves the same outcome. To see this, consider the following. Suppose that a bonus policy with *b* bonus points for the disadvantaged group leads to an increase in the proportion of admitted disadvantaged applicants from $x\%$ (with no intervention) to $y\%$ (with the bonus policy). For example, suppose that giving $b = 10$ bonus points to lower-income students leads to an increase in the proportion of lower-income students from $x\% = 20\%$ to $y\% = 30\%$ among the admitted applicants. Then, exactly the same outcome (i.e., exactly the same set of admitted applicants) would have been achieved with a quota policy that required at least $y\% = 30\%$ of the admitted applicants to be from the lower-income group. The same holds in the opposite direction (given a quota policy, we can find a bonus that achieves the same outcome). The equivalence of the outcomes (i.e., sets of admitted applicants) holds because, either with a bonus policy or with a quota policy, it is always the set of applicants with the highest admission scores from each group that is admitted.

In our paper, of course, the cohort of applications is not fixed, but uncertain: there is a probability associated with every possible set of applications. However, the equivalence of bonus and quota policies still holds: for every set of applications sampled from such a probability distribution, we could have considered either a bonus policy with *b* bonus points or an equivalent quota policy with $y\%$ quota for the disadvantaged group.

Despite the equivalence between bonus and quota policies, there might be other considerations that favour one over the other. On the one hand, quota policies are easier to interpret from the point of view of a system administrator because they express the policy in terms of the final aggregate outcome, i.e., how many applicants of a given group will be admitted. On the other hand, bonus policies are easier to interpret from the point of view of individual applicants because individual applicants make their decisions based on their admission scores and bonus policies express directly how they affect an applicant's admission score.

In this paper, we describe bonus policies, keeping in mind that all of them have equivalent quota policies.

4.2 Problem Definition

In what follows, and unless otherwise specified, we will use the term *score* to refer to the score of applicants before the addition of bonus points. Accordingly, when we refer to the applicants that are admitted by a program, we will use the term *average admission score* to refer to the average of the pre-bonus scores with which they were admitted to the program.

We wish to design an admission policy that leads to a reduction in admission rate disparities while simultaneously admitting the students with the highest (pre-bonus) scores. We thus define the objective function of the bonus policy as a linear combination of two quantities: the inequality of admission rates, captured by SPD, and the aptitude of the admitted applicants, captured by their average pre-bonus score. Specifically, let μ_b be the average admission score when b bonus points are given, and, since average admission score might vary strongly between application years, consider the decrease in μ_b compared to μ_0, i.e., the average admission score when no bonus policy is implemented.

Given that the gap in admission rates should be small and the loss in average admission score should be minimized, we aim to minimize our objective function, which is formulated as follows:

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

We calculate the loss in average admission score using $\mu_0 - \mu_b$. Note that in the case that the bonus policy happens to improve the average admission score, $\mu_0 - \mu_b$ is negative. Because we aim to minimize the objective function, this would mean a desirable improvement in the objective function. Parameter λ is used to regulate the balance between the average admission score loss and the inequality in admission rates. Its value expresses how much loss in the average admission score can be tolerated for λ units of reduction in admission rate discrepancy. In practice, its value would be set by program administrators.

Note that, while this objective function fits the purposes of affirmative action, it is not the only possible one: variants of it could be defined and used, depending on the objectives of affirmative action in a particular setting. For example, as defined, the objective function in Equation (2) expresses inequality as the gap between the admission rates of two groups, and its minimization leads to policies where the groups have similar admission rates. An alternative definition would have inequality expressed as the disparity between the share of the protected group in the group of accepted students and the share of the protected group of the entire population in the country. In that case, minimizing the objective function would lead to admission policies that better represent the entire population. As another example, the objective function in Equation (2) considers the absolute value of SPD, assuming that disparities in either direction (i.e., either against or in favour of the protected group) are undesirable. A stronger notion of affirmative action, in which disparities in favour of the protected group are desirable, would use an alternative definition in which the raw value of SPD is considered rather than its absolute value.

In practice, designing a bonus policy requires a detailed examination of its objectives and possibly formalization of the objective function in a different way.

We aim to design a bonus policy that optimizes (i.e., minimizes) the objective function for a given program over multiple possible *application sets*, i.e., the possible applicant pools and their preferences. Notice that considering multiple possible application sets is necessary since the application set is not known with certainty at the time when the bonus policy is announced. Possible application sets can be collected in two ways, referred to as *strategies* hereafter. One way is to use application sets from historical data, i.e., from previous years. Another way is to resort to predictions, i.e., to create a sample of possible application sets based on historical data.

A conceptually straightforward approach to optimizing the objective function for a given program and sensitive group is to collect a number of possible application sets in one of the two ways described above, estimate the optimal bonus b for each of them and the average $\bar{b}$ over all of them, and finally suggest a bonus policy with $\bar{b}$ bonus points for the given program and sensitive group. The optimal bonus b for a given application set is determined by applying a range of bonus points, directly measuring the resulting value for the objective function, and picking the bonus that minimizes the objective function. In the rest of this section, we elaborate on the strategies we consider for collecting application sets, either from a statistical model learned from historical data or directly from historical data.

Effect on Average Admission Score. Notice that, whatever the number of bonus points b used by the given university program for a bonus policy, the loss in average admission score for the program can be at most equal to b. To see why, let X be the set of bonus recipients who were rejected under no intervention but are admitted with the bonus policy, and let Y be the set of applicants who were admitted under no intervention but are rejected with the bonus policy. Let x_i and y_i be the top-i candidates from X and Y, respectively. Intuitively, x_i replaces y_i as a consequence of the bonus policy. Then, the (pre-bonus) score of x_i is at most b points smaller than the (pre-bonus) score of y_i—otherwise, the applicants in X would not have high enough scores to replace Y. Therefore, the average admission score (i.e., the average pre-bonus score of admitted applicants) of this program can decrease by at most b points.

4.3 Policies Based on the Predictive Model

We build a multi-label probabilistic classifier to model students' application behaviour and use it to sample a collection of application sets. To predict applications for a certain year, e.g., 2017, we trained the model on data from the most recent year at that time, i.e., 2016. We deployed the trained model to create n possible application sets. Each application set consists of a sampled set of students together with their predicted program preferences. We experimented with $n = 50$ and $n = 200$ sampled application sets to evaluate possible bonus policies.

The input for the classifier that we built is a given student's features. These features are the pre-processed data described in Section 3 and thus include, e.g., the student's standardized high school grades and household size. One possibility would be to let the model output the programs to which the student is predicted to apply. This corresponds to a multi-label classification task (Tsoumakas & Katakis, 2008), 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. We thus let the model output the probability that the given student applies to each program. To select the student's predicted applications for the generated application set, we sampled *n* programs without replacement based on the application probabilities predicted for all programs. The sampled programs were then ranked by decreasing application probability.

The sampling of programs is necessary because the probabilities with which students are predicted to apply to smaller programs are always low simply because very few students apply to these programs. Hence, a ranking of the programs based on the predicted probabilities would always lead to bigger and more prestigious programs being placed at the top. Lower-ranked programs, on the other hand, would not receive any applications. The resulting matching through the DA algorithm would then not be realistic because in practice there are of course students applying to the lower-ranked programs. This means that the results for both the lower- and the higher-ranked programs would be skewed.

We evaluate this model against two baselines: *random* and *unigram*. For the random baseline, we predict each program with the same probability; for the unigram baseline, we predict every program p with $\frac{\text{\# applications to } p}{\text{\# total applications}}$. We use the well-known normalized discounted cumulative gain (nDCG) (Järvelin & Kekäläinen, 2002) as the performance measure. nDCG is a metric designed to evaluate rankings of results from search engines. It is useful in our context because it places a higher weight on the predicted first choice, while 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. Values range from 0.0 to 1.0, with 1.0 representing a perfect prediction. 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 trained a random forest classifier (Breiman, 2001) and tuned its hyperparameters with Bayesian optimization (Pelikan et al., 1999) over 100 iterations to reach an nDCG score of 0.43 on the testing data.

We compare this approach to two other approaches with which we achieved nDCG scores of 0.44 and 0.35. Due to its strong performance compared to the other models and its practical advantages, we used the multi-label probabilistic classifier to sample application sets.

4.4 Policies Based on Historical Data

A simple approach to choosing a bonus policy is to compute what bonus policy would have been optimal in the previous years. This 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, it would be necessary to use a model built from such statistics. 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 these bonus values as the bonus value for the upcoming application round.

5. Experimental Evaluation

We evaluate each strategy for both sensitive attributes—gender and income—for the years 2016 and 2017[2]. To find a fitting value for λ in Equation (2), we calculated the weighted median differences in grades and test scores between subgroups according to the following formula:

$$\lambda = \sum_{\text{score } s} (\text{median}(W_s) \times |\text{median}(S_{A=a}) - \text{median}(S_{A\neq a})|),$$

where the summation is over the scores (high school grade, math score, etc.) that contribute to the admission score, $median(\ldots)$ refers to the median of a set of values, W_s is the set of weights that a given score s is associated with over all programs, and $S_{A=a}$ and $S_{A\neq a}$ are the score values for the two groups (male/female or high-/low-income) over all programs. 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 used the strategies to suggest a policy for each program and provide aggregate statistics on its effect (Subsection 5.1). Subsequently, we discuss the effects on two subsets of university programs. One subset is the set of the most prestigious programs (Subsection 5.2), where we explore whether the bonus policy has a large and potentially undesirable effect on their average admission score. The other subset is that of programs that show consistent inequalities in admission rates over time

[2] The code for this is publicly available on GitHub: https://github.com/hcorinna/university-admissions.

(Subsection 5.3). We focus on them because, as Subsection 3.6 demonstrated, bonus policies to equalize admission rates should only be used sparingly in order to avoid adverse effects due to variance across years. In all 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.

5.1 All Programs

We begin by comparing the objective function values resulting from applying the different strategies with the ideal bonus policies that are designed in hindsight, i.e., assuming that the application set for each program is already known with certainty. In Table 2, 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 2, policies that use more application sets for their suggestions generally lead to both smaller errors and smaller variance in the error.

Table 2. Error in Objective Function Relative to Ideal Policies*

		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

*Smaller values are better.

Tables 3 and 4 split the findings from Table 2 into its two components: average admission score and SPD. We note that while of course no strategy can find a better value for the objective function than the ideal policy, it is possible for a strategy to suggest policies that lead to a lower loss in average admission score or a smaller gap in admission rates. Indeed, we find that on average all strategies produce a smaller average admission score 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 3. Average Admission Score Loss Relative to Ideal Policies*

		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**

*Lower values are better.

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—by definition of Equation (2)—guaranteed to lead to no average admission score loss.

An average admission score loss is, however, expected when intervening with a bonus policy. How big it is for the different strategies can be seen in Table 5. What is once again evident is that the average admission score loss is smaller for the strategies that use more application sets.

To explore the effect on the admission rates gap, Table 6 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: $|\text{SPD}_b| - |\text{SPD}_0|$. Negative values thus indicate a lower admission rate gap with the intervention, which is desirable. This is more likely to occur for strategies based on multiple application sets. In general, we can also see that the values suggested by more application sets again exhibit smaller variance.

Table 4. Difference in Absolute SPD Relative to Ideal Policies*

		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

*Lower values are better.

Table 5. Average Admission Score Loss Relative to No Intervention*

		Gender		Income	
	Strategy	Mean	SD	Mean	SD
2016	Historical: 1 year	0.0462	0.2587	0.0616	0.3444
	Historical: 3 years	0.0162	0.0803	0.0373	0.2672
	Historical: 5 years	0.0144	0.0800	0.0356	0.2245
	Predictive: 50 sets	0.0063	0.0450	0.0299	0.1961
	Predictive: 200 sets	**0.0061**	**0.0442**	**0.0293**	**0.1864**
2017	Historical: 1 year	0.0287	0.1461	0.0518	0.2530
	Historical: 3 years	0.0151	0.0832	0.0322	0.2333
	Historical: 5 years	0.0132	0.0816	0.0303	**0.1997**
	Predictive: 50 sets	**0.0045**	**0.0406**	0.0300	0.2276
	Predictive: 200 sets	0.0049	0.0413	**0.0298**	0.2191

*Lower values are better, in most cases, less than 1 point out of 850 points.

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 7). 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 zero 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 in our case.

5.2 Effects on Prestigious Programs

In competitive programs, the effect of bonus policies will be much more visible than in programs that get just enough applications to fill all spots. Therefore, the effect of these bonus policies is particularly interesting for high-prestige programs because they are typically very competitive. One might wonder whether a highly prestigious program would still be considered prestigious if such a bonus policy led to a stark decrease in average admission score. We thus evaluate how the average admission score changes under different strategies. Moreover, we are interested in how the gap in admission rates changes. Tables 8 and 9 show these results for the years 2016 and 2017 for gender-based and income-based bonus policies for the four most prestigious programs, i.e., as ranked based on the average scores of the students they admitted from 2013 to 2015. We note that all four demonstrated programs are for medicine and that we observed similar results for the top 10 programs—however, we show results only for the top four for brevity.

It is evident that the bonus policies barely affect the average admission score of the programs. Most changes are below one point—a negligible change considering that the maximum average admission score is 850 points. We observe stronger changes in the gap in admission rates. In order to better understand how that change in the admission rate gap affects the composition of the student body, the tables also include the share of accepted women and low-income students. This data shows that the bonus policies have a sizable effect on the share of women and, in particular, on the share of low-income students. For example, Table 9 shows that for 2016, no intervention would mean that only 7% of the admitted students to the #1 program are low-income, while the share of low-income students rises to 15% or more for the identified bonus policies.

Table 6. Difference in Absolute SPD Relative to No Intervention*

		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

*Lower values are better.

Table 7. 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

Notice that for program #3, women are already in the majority when no bonus policy is implemented. However, the negative SPD shows that women still have a lower admission rate than men. Since our objective function tries to minimize the gap in admission rates, i.e., the absolute SPD, this implies that bonus points have to be given to women, which further increases their share in the program. This phenomenon, which may or may not be desirable, is a direct result of the definition of our objective function. If the goal is to admit a student body that is more representative of the applicants, then the objective function has to be adapted accordingly.

5.3 Policies for Programs with Consistent Inequalities

We observe in Table 6 that some design strategies have averages above zero 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 discussed in Subsection 3.6.

In practice, affirmative action policies aimed at reducing gaps in admission rates should be deployed not for all programs (remember that indeed most programs have fewer applicants than vacancies) but only sparingly for programs whose admission rates are consistently unequal along a protected attribute. 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 (i.e., the absolute difference in admission rates was at least 0.1). 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 10, 11, and 12 report the same measure as seen in Subsection 5.1, 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).

Table 8. Effects of Different Strategies for Gender-Based Bonus Policies on the Top Four Programs*

		#1			#2			#3			#4		
Year	Strategy	Avg. adm. score	SPD in pp	%	Avg. adm. score	SPD in pp	%	Avg. adm. score	SPD in pp	%	Avg. adm. score	SPD in pp	%
2016	No intervention	801.8	−8.03	41	790.6	−6.19	47	776.1	−0.43	60	772.5	−6.59	44
	Ideal policy	801.2	−0.49	55	790.3	−0.39	55	776.1	−0.43	60	772.1	0.02	52
	Historical: 1 year	801.7	−5.02	47	790.2	0.50	56	776.0	4.03	67	772.1	0.75	53
	Historical: 3 years	801.6	−4.26	48	790.1	1.39	57	775.3	11.89	78	771.6	3.62	56
	Historical: 5 years	801.5	−2.75	51	789.8	4.07	61	775.3	11.89	78	771.5	4.35	57
	Predictive: 50 sets	801.5	−2.75	51	790.4	−1.28	53	775.3	11.89	78	772.1	0.75	53
	Predictive: 200 sets	801.6	−4.26	48	790.3	−0.39	55	775.4	10.78	77	771.8	2.15	55
2017	No intervention	805.3	−3.93	50	794.5	−12.40	41	777.6	−2.87	57	776.8	−2.63	53
	Ideal policy	805.2	−0.75	53	793.8	0.24	53	777.6	−1.46	58	776.7	−0.56	55
	Historical: 1 year	804.9	4.55	59	794.2	−4.25	49	777.6	−2.87	57	776.5	4.62	59
	Historical: 3 years	805.1	1.37	56	794.2	−3.61	50	777.6	−1.46	58	776.3	6.83	61
	Historical: 5 years	805.0	3.49	58	794.0	−1.69	52	777.1	6.04	67	776.0	9.93	64
	Predictive: 50 sets	805.2	−1.81	52	793.8	0.24	53	777.2	4.51	65	776.5	4.62	59
	Predictive: 200 sets	805.2	−0.75	53	793.7	1.52	55	777.2	4.51	65	776.4	5.94	60

*“SPD in pp” shows the difference in admission rates between women and men in percentage points. “%” refers to the share of women in the program.

Table 9. Effects of Different Strategies for Income-Based Bonus Policies on the Top Four Programs*

		#1			#2			#3			#4		
Year	Strategy	Avg. adm. score	SPD in pp	%	Avg. adm. score	SPD in pp	%	Avg. adm. score	SPD in pp	%	Avg. adm. score	SPD in pp	%
2016	No intervention	801.8	−9.94	7	790.6	−12.31	13	776.1	−8.48	22	772.5	−9.62	24
	Ideal policy	800.7	−0.84	16	789.3	−0.07	25	775.3	−0.06	33	771.6	−0.04	35
	Historical: 1 year	800.4	0.46	18	788.6	2.84	28	775.8	−3.71	28	770.7	3.98	40
	Historical: 3 years	797.6	9.57	27	789.0	1.10	27	775.2	1.13	35	770.0	6.39	43
	Historical: 5 years	798.8	5.67	23	789.3	−0.07	25	775.4	−0.06	33	770.7	3.98	40
	Predictive: 50 sets	801.0	−2.14	15	787.9	5.16	31	772.3	13.00	52	770.3	5.59	42
	Predictive: 200 sets	801.0	−2.14	15	787.9	5.16	31	773.0	10.62	48	770.3	5.59	42
2017	No intervention	805.3	−19.41	8	794.5	−24.14	15	777.6	−21.65	25	776.8	−25.36	22
	Ideal policy	803.5	0.13	23	791.4	−1.87	34	775.4	−2.29	47	773.1	−0.19	44
	Historical: 1 year	804.0	−2.88	20	793.4	−10.22	27	777.0	−11.32	37	775.8	−13.80	32
	Historical: 3 years	802.5	4.64	26	792.8	−7.44	29	777.2	−12.80	35	774.7	−7.46	38
	Historical: 5 years	802.8	3.13	25	793.3	−9.52	27	777.0	−11.32	37	774.7	−7.46	38
	Predictive: 50 sets	804.2	−4.38	19	792.2	−4.65	32	776.3	−6.71	42	773.6	−2.27	43
	Predictive: 200 sets	804.0	−2.88	20	792.2	−4.65	32	776.3	−6.71	42	773.6	−2.27	43

*“SPD in pp” shows the difference in admission rates between low- and high-income students in percentage points. “%” refers to the share of low-income students in the program.

Table 10. Error in the Objective Function Relative to Ideal Policies*

		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

*Smaller values are better.

Table 11. Values Relative to Ideal Policies*

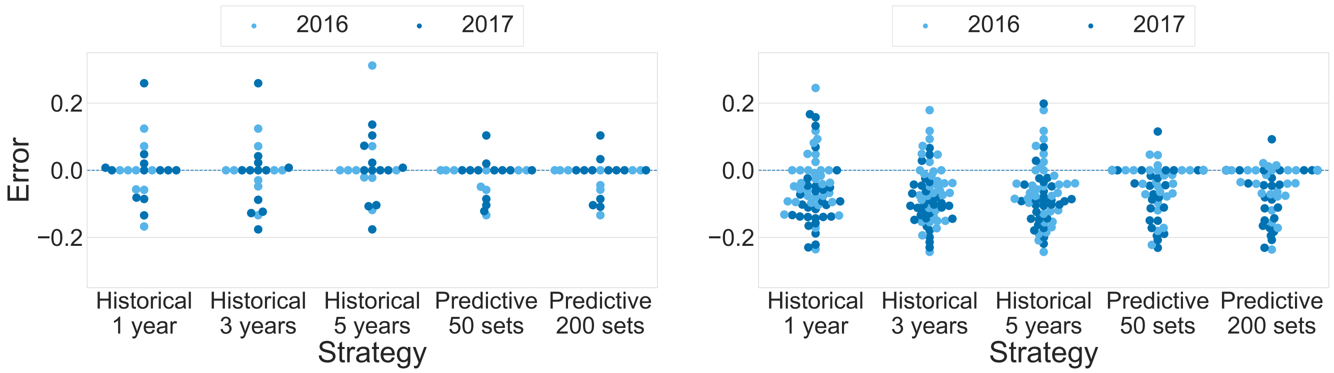
		Average admission score loss				Difference in absolute SPD			
		Gender		Income		Gender		Income	
	Strategy	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2016	Historical: 1 year	−0.0467	**0.1491**	0.2732	1.0825	0.0368	0.0539	0.0744	0.0635
	Historical: 3 years	−0.019	0.2618	0.3094	1.1122	0.0447	0.0504	0.0562	0.0524
	Historical: 5 years	0.0421	0.3599	0.1482	0.8849	0.0711	0.0945	**0.0484**	**0.0519**
	Predictive: 50 sets	**−0.1034**	0.1541	−0.1447	0.8949	**0.0196**	**0.0255**	0.0735	0.0637
	Predictive: 200 sets	−0.0723	0.1931	−0.1815	**0.8404**	0.0202	0.0258	0.0734	0.0643
2017	Historical: 1 year	-0.0913	0.3613	−0.6402	**1.2614**	0.0608	0.0788	0.0867	0.0558
	Historical: 3 years	0.0378	0.2993	−0.4874	1.3413	0.0428	0.0702	**0.0658**	**0.0511**
	Historical: 5 years	−0.0322	0.3798	−0.5864	1.3225	0.0544	**0.0480**	0.0792	0.0618
	Predictive: 50 sets	**−0.138**	**0.281**	−0.5646	1.4039	**0.0424**	0.0561	0.1016	0.0869
	Predictive: 200 sets	−0.1197	0.2985	−0.5826	1.3181	0.0446	0.0555	0.0972	0.0851

*Lower values are better.

Table 12. Values Relative to No Intervention*

		Average admission score loss				Difference in absolute SPD			
		Gender		Income		Gender		Income	
	Strategy	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2016	Historical: 1 year	0.19	0.29	0.95	1.31	-0.0096	0.0777	−0.0332	0.0877
	Historical: 3 years	0.22	0.32	0.99	1.27	−0.0017	0.068	−0.0514	0.0918
	Historical: 5 years	0.28	0.38	0.83	1.01	0.0247	0.1116	**−0.0592**	0.0937
	Predictive: 50 sets	**0.13**	**0.21**	0.53	0.88	**−0.0268**	0.0438	-0.0341	0.0612
	Predictive: 200 sets	0.17	0.25	**0.50**	**0.81**	−0.0263	**0.0435**	−0.0342	**0.0586**
2017	Historical: 1 year	0.17	**0.20**	**0.80**	**0.68**	0.0028	0.0922	−0.0774	0.1000
	Historical: 3 years	0.30	0.32	0.95	1.19	−0.0153	0.1069	0.0983	**0.0704**
	Historical: 5 years	0.23	0.31	0.85	0.95	−0.0037	0.0858	**−0.0849**	0.0827
	Predictive: 50 sets	**0.12**	**0.20**	0.87	1.19	**−0.0156**	0.0586	−0.0625	0.0822
	Predictive: 200 sets	0.14	0.24	0.86	1.13	−0.0135	**0.0575**	−0.0670	0.0817

*Lower values are better.

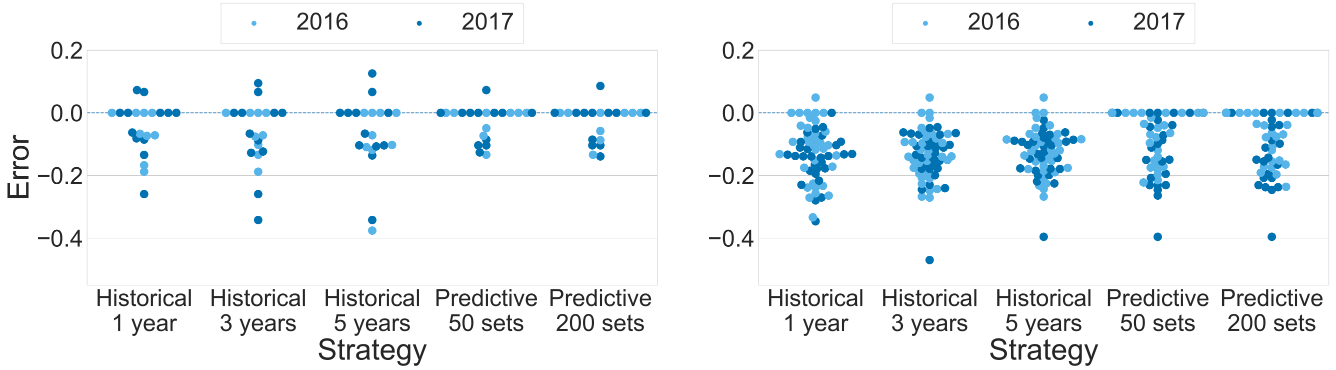


(a) gender **(b)** income

Figure 5. Difference in Absolute SPD Relative to No Intervention*

*Lower is better.

Figure 5 illustrates the findings for the difference in absolute SPD from Table 12. As we can see, the filter left only a few cases where the absolute SPD is increased through the intervention. For 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 6 shows whether an increase in the absolute SPD occurs in favour 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 be almost entirely 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.



(a) gender **(b)** income

Figure 6. 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. Conclusions

The goal of this paper was to use historical data to design bonus policies that reduce the gap in admission rates between two socio-demographic groups while simultaneously still admitting students with high grades and test scores. We compared different strategies that could be used to determine bonus policies. Specifically, we compared a predictive approach to simpler design strategies based on averaging retrospectively optimal bonuses over one to five years of historical data. In doing so, we found that simply using the previous year's ideal bonus policy is likely to overcorrect the differences in admission rates. Basing the bonus policy suggestion on more data, e.g., through our proposed predictive approach, mostly avoids this pitfall by using more conservative suggestions. With the predictive strategy achieving similar results as the simpler approaches that use data from several application years, the simpler approach based on sufficient historical data (e.g., the last five years) appears to be preferable in practice due to its faster and less error prone implementation. Predictive methods are, however, advantageous if historical data is only available for a few years or if only aggregate statistics are accessible.

Our work has also led to practical findings that we hope will be significant for practitioners, such as university administrators, who aim to implement their own affirmative action policies for their programs, but also for the understanding of the general public. One finding of our work is that, for this dataset, designing policies that balance applicant scores with admission rates leads to effects that are limited or moderate rather than extensive. Another finding is that affirmative action should be used sparingly and should preferably be used with programs that exhibit consistent admission rate discrepancies between protected groups; otherwise, the affirmative action may "overshoot" and lead to discrepancies in the opposite direction. Moreover, in the case of the most prestigious programs, we saw that, even though the effects appeared minimal in terms of the objective function and the average admission score, the policies led to a non-trivial positive effect on the admission rate of the protected group. And, finally, it was generally found that the effects are more pronounced for income-based policies (where the gap in admission rates is larger) than for gender-based policies (where the gap in admission rates is comparatively smaller). Note that these findings are limited to the case where universities want to trade off average admission score against the gap in admission rates (instead of, e.g., the representation of underrepresented groups).

This formalization of the objective function is the main limitation of our work because it embodies specific objectives that we discussed in Subsection 4.2. If universities want to, e.g., make the demographic distribution of the students in their programs more representative of the general population (instead of the applicant pool), they should use a somewhat different objective function, possibly leading to higher bonus points than with the conservative bonus point policies identified in this work. Our empirical findings are thus limited to cases that share the set of assumptions underlying our objective function. Further evaluations are necessary to understand how the tested approaches compare for other objective functions (e.g., when the goal is to increase the representation of an underrepresented group).

An aspect that is left for future research is the effect of the announcement of affirmative action policies on application behaviour. Existing research (e.g., Balafoutas & Sutter, 2012; Dickson, 2006; Long, 2004) suggests that the mere existence of affirmative action policies might encourage students from disadvantaged groups to apply; empirical evidence in Chile also shows this (Bastarrica et al., 2018). It is therefore imaginable that fewer bonus points may suffice to achieve the desired effect, which is one more reason to prefer conservative strategies, assuming that the goal of the policies is to reduce gaps in admission rates.

Acknowledgements

We thank Sergio Celis for insightful discussions, and our anonymous reviewers for their helpful feedback. Corinna Hertweck developed this work as part of her MSc Thesis at the University of Helsinki (Hertweck, 2020).

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This study was financially supported by the University of Helsinki; the HUMAINT programme (Human Behaviour and Machine Intelligence), Joint Research Centre, European Commission; and "la Caixa" Foundation (ID 100010434), under the agreement LCF/PR/PR16/51110009.

References

Abdulkadiroğlu, A. (2005). College admissions with affirmative action. *International Journal of Game Theory*, *33*(4), 535–549. https://doi.org/10.1007/s00182-005-0215-7

Abdulkadiroğlu, A., & Sönmez, T. (2003). School choice: A mechanism design approach. *American Economic Review*, *93*(3), 729–747. https://doi.org/10.1257/000282803322157061

Bacharach, V. R., Baumeister, A. A., & Furr, R. M. (2003). Racial and gender science achievement gaps in secondary education. *Journal of Genetic Psychology*, *164*(1), 115–126. https://doi.org/10.1080/00221320309597507

Balafoutas, L., & Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, *335*(6068), 579–582. https://doi.org/10.1126/science.1211180

Bastarrica, M. C., Hitschfeld, N., Marques Samary, M., & Simmonds, J. (2018). Affirmative action for attracting women to STEM in Chile. *Proceedings of the Workshop on Gender Equality in Software Engineering* (GE 2018), 27 May–3 June 2018, Gothenburg, Sweden (pp. 45–48). ACM. https://www.computer.org/csdl/proceedings-article/ge/2018/573801a045/13l5NXW7OuR

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. https://doi.org/10.48550/arXiv.1810.01943

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brest, P., & Oshige, M. (1995). Affirmative action for whom? *Stanford Law Review*, *47*(5), 855–900. https://doi.org/10.2307/1229177

Cabalin, C. (2012). Neoliberal education and student movements in Chile: Inequalities and malaise. *Policy Futures in Education*, *10*(2), 219–228. https://doi.org/10.2304/pfie.2012.10.2.219

Crosby, F. J., Iyer, A., & Sincharoen, S. (2006). Understanding affirmative action. *Annual Review of Psychology*, *57*, 585–611. https://doi.org/10.1146/annurev.psych.57.102904.190029

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

DEMRE. (n.d.). *Puntaje Ranking* [Accessed on 10/28/2019]. https://psu.demre.cl/proceso-admision/factores-seleccion/puntaje-ranking

Dickson, L. M. (2006). Does ending affirmative action in college admissions lower the percent of minority students applying to college? *Economics of Education Review*, *25*(1), 109–119. https://doi.org/10.1016/j.econedurev.2004.11.005

Evans, G. W. (2004). The environment of childhood poverty. *American Psychologist*, *59*(2), 77. https://doi.org/10.1037/0003-066X.59.2.77

Faucon, L., Olsen, J. K., Haklev, S., & Dillenbourg, P. (2020). Real-time prediction of students' activity progress and completion rates. *Journal of Learning Analytics*, *7*(2), 18–44. https://doi.org/10.18608/jla.2020.72.2

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (Im)possibility of Fairness. *arXiv:1609.07236*. https://doi.org/10.48550/arXiv.1609.07236

Gale, D., & Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, *69*(1), 9–15. https://doi.org/10.1080/00029890.1962.11989827

Hafalir, I. E., Yenmez, M. B., & Yildirim, M. A. (2013). Effective affirmative action in school choice. *Theoretical Economics*, *8*(2), 325–363. https://doi.org/10.3982/TE1135

Hertweck, C. (2020). *Designing Affirmative Action Policies under Uncertainty* (Master's thesis). University of Helsinki. Helsinki, Finland. http://urn.fi/URN:NBN:fi:hulib-202005202223

Hoxby, C. M., & Avery, C. (2012). *The Missing "One-Offs": The Hidden Supply of High-Achieving, Low Income Students* (tech. rep.). National Bureau of Economic Research. https://www.brookings.edu/wp-content/uploads/2016/07/2013a_hoxby.pdf

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 422–446. https://doi.org/10.1145/582415.582418

Kawagoe, T., Matsubae, T., & Takizawa, H. (2018). The skipping-down strategy and stability in school choice problems with affirmative action: Theory and experiment. *Games and Economic Behavior*, *109*, 212–239. https://doi.org/10.1016/j.geb.2017.12.012

Kojima, F. (2012). School choice: Impossibilities for affirmative action. *Games and Economic Behavior*, *75*(2), 685–693. https://doi.org/10.1016/j.geb.2012.03.003

Lancaster, A., Moses, S., Clark, M., & Masters, M. C. (2020). The positive impact of deliberate writing course design on student learning experience and performance. *Journal of Learning Analytics*, *7*(3), 48–63. https://doi.org/10.18608/jla.2020.73.5

Long, M. C. (2004). College applications and the effect of affirmative action. *Journal of Econometrics*, *121*(1), 319–342. https://doi.org/10.1016/j.jeconom.2003.10.001

Mathioudakis, M., Castillo, C., Barnabo, G., & Celis, S. (2020). Affirmative action policies for top-k candidates selection: With an application to the design of policies for university admissions. *Proceedings of the 35th Annual ACM Symposium on Applied Computing* (SAC 2020), 30 March–3 April 2020, Brno, Czechia (pp. 440–449). ACM. https://doi.org/10.1145/3341105.3373878

McEwan, P. J. (2004). The indigenous test score gap in Bolivia and Chile. *Economic Development and Cultural Change*, *53*(1), 157–190. https://doi.org/10.1086/423257

Meneses, F., & Cáceres, J. T. (2012). Predicción de notas en Derecho de la Universidad de Chile: ¿sirve el ranking? *ISEES: Inclusión Social y Equidad en la Educación Superior*, *2012*(10), 43–60. https://www.fundacionequitas.cl/publicaciones/isees/10/4420036.pdf

Ministerio de Educación de Chile. (2009). *Bases para una política de formación técnico-profesional en Chile. Informe de la Comisión para el Estudio de la Formación Técnico-Profesional en Chile* [Executive Summary]. https://ciperchile.cl/wp-content/uploads/documento-link-4.pdf

OECD & World Bank. (2009). *Reviews of National Policies for Education: Tertiary Education in Chile*. https://openknowledge.worldbank.org/bitstream/handle/10986/10219/527530BRI0En1B10Box345577B01PUBLIC1.pdf

Pelikan, M., Goldberg, D. E., & Cantú-Paz, E. (1999). BOA: The Bayesian optimization algorithm. In W. Banzhaf, J. M. Daida, A. E. Eiben, M. H. Garzon, & V. Honavar (Eds.), *Proceedings of the First Annual Conference on Genetic and Evolutionary Computation—Volume 1* (GECCI 1999), 13–17 July 1999, Orlando, FL, USA (pp. 525–532). Morgan Kaufmann Publishers. https://dl.acm.org/doi/10.5555/2933923.2933973

Reardon, S. F. (2013). The widening income achievement gap. *Educational Leadership*, *70*(8), 10–16. https://www.ascd.org/el/articles/the-widening-income-achievement-gap

Ríos, I., Larroucau, T., Parra, G., & Cominetti, R. (2014). *College Admissions Problem with Ties and Flexible Quotas*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2478998

Rothstein, R. (2015). The racial achievement gap, segregated schools, and segregated neighborhoods: A constitutional insult. *Race and Social Problems*, *7*(1), 21–30. https://doi.org/10.1007/s12552-014-9134-1

Tsoumakas, G., & Katakis, I. (2008). Multi-label classification: An overview. In J. Wang (Ed.), *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 64–74). IGI Global. https://www.igi-global.com/chapter/multi-label-classification/7632

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