Towards Data-Driven Affirmative Action Policies under Uncertainty

Corinna Hertweck,[1*] Carlos Castillo,[2] Michael Mathioudakis[3]
[1] Zurich University of Applied Sciences, [2] Universitat Pompeu Fabra, [3] University of Helsinki

arXiv:2007.01202v1 [cs.CY] 2 Jul 2020

ABSTRACT

In this paper, we study university admissions under a centralized system that uses grades and standardized test scores to match applicants to university programs. We consider affirmative action policies that seek to increase the number of admitted applicants from underrepresented groups. Since such a policy has to be announced before the start of the application period, there is uncertainty about the score distribution of the students applying to each program. This poses a difficult challenge for policy-makers. We explore the possibility of using a predictive model trained on historical data to help optimize the parameters of such policies.

1. INTRODUCTION

In centralized university admission systems, students submit their applications to a central institution, which matches the applicants with the programs offered by various universities. This matching is typically automated through an algorithm and based on grades and standardized test scores. As grades and test scores tend to differ across demographic groups (see, e.g., [2, 15, 18, 19]), this kind of system can lead to large gaps in admission rates between groups. These gaps can be reduced through affirmative action policies, which are challenging to design. First, such policies should reward merit and lead to the admission of the best applicants. Second, policies should be announced *before* the application period begins to provide candidates with all the information they need. Computational methods can be leveraged to evaluate a large range of alternatives, much more than can be evaluated manually, and identify effective policies.

In this paper, we develop an approach for designing robust and effective affirmative action policies for university admissions. 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 [14]. The technical problem we face is to choose the right number of bonus points so that the policy will have a consistently beneficial effect on the admission rate of the given group.

Related Work

Using the terminology of Friedler et al. [9], our work and affirmative action more generally are consistent with the *we are all equal* (WAE) worldview. According to WAE, higher grades do not necessarily imply more talent or more diligence, but are seen as a product of structural inequality (e.g., lack of resources for subgroups of the student population [8]). Affirmative action policies (e.g., bonus policies) are thus implemented as a way to counter the effect of inequalities being reinforced through – in our case – unequal grade and test score distributions. The beneficial effects of affirmative action policies on disadvantaged groups have both been shown in studies based on real-world implementations of specific policies (e.g., [5, 6]) and in laboratory settings (e.g., [3]). Other studies discuss how different types of affirmative action policies can be incorporated into the algorithm that matches students with university programs (see, e.g., [1, 10, 11]). However, existing literature lacks a discussion of how the numerical parameters in these policies (e.g., the number of bonus points given to an underrepresented group) affect the resulting matching. The usage of computational methods in policy design has, for example, already been researched in the fields of public health [4] or counter-terrorism [12, 16]. In the area of university admissions, recent work [14] addressed the problem of the design of affirmative action policies. However, the study designed policies for a given set of applications and not under uncertainty. 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.

2. DATASET

We analyze data from the university admission system of Chile, which contains information on all applicants and university programs between 2004 and 2017. For 2017, this amounts to 60 000 students and about 1 500 programs.

After taking standardized tests and obtaining their scores, students submit their *preferences* as a ranked list of up to ten programs. Students are admitted to programs according to an admission score, which is a weighted average of

*Work done while at the University of Helsinki.

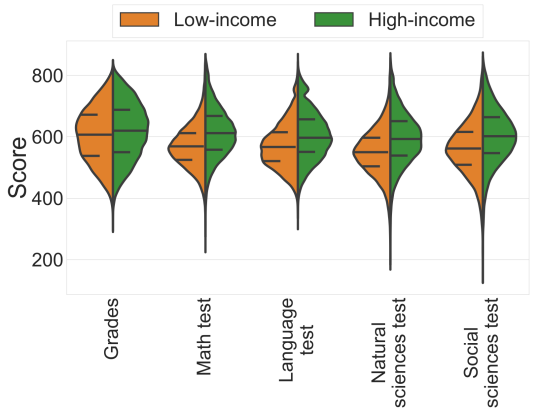

Figure 1: Distributions of grades and standardized test scores for different income levels.

high school grades and test scores – with different weights used by different programs. The data contains these high school grades, test scores, and applicant preferences. It also includes a (binary) gender variable as well as the household income and size. We binarize the student's income as *low-income* (bottom five household-income-per-household-member deciles), and *high-income* (top five deciles).

Differences in scores and preferences. According to statistics of the year 2016, with similar observations for other years, we find that there is a pronounced difference between the score distributions of different income groups: high-income students have slightly higher high school grades than low-income students, and much higher scores across all standardized tests (Figure 1). While there exist differences between genders, too, they are less pronounced in the data.

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 preference for programs of higher average prestige compared to low-income students, particularly in their first preference (Figure 2). The difference can largely be attributed to the different scores of the two groups (Figure 1) – once we control for admission scores, the differences mostly disappear (figure omitted for brevity).

Differences in admission rates. To measure differences in admission rates, we use the *Statistical Parity Difference (SPD)* measure:

$$P(Y = 1|A = a) - P(Y = 1|A \neq a), \quad (1)$$

where $Y = 1$ indicates being admitted into a program, A is a sensitive attribute, and a marks a demographic group (in this work either low-income or female students). Low absolute values of SPD are desired – and perfect equality is achieved for a value of 0. Following thresholds provided in tools for measuring algorithmic bias [17], we accept values inside $[-0.1, 0.1]$ and reject values outside of this range as *strongly unequal*. As shown in Figure 3, we tend to measure larger negative SPD values for the more prestigious programs, indicating differences in admission rates that place low-income students at a disadvantage.

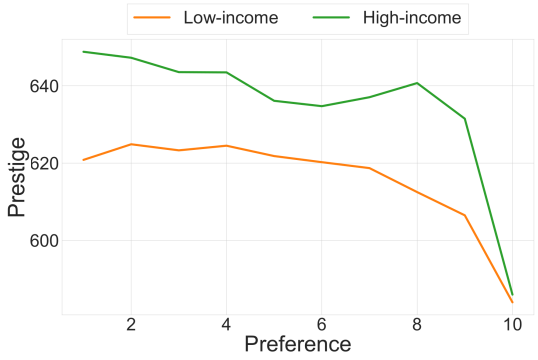

Figure 2: Average prestige of students' preferences across income levels.

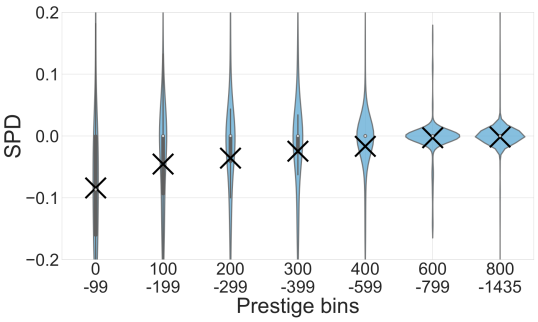

Figure 3: Distribution of SPD for income. Programs are ranked by prestige (higher to lower); x-axis indicates program ranks in each bin; crosses mark the means.

Variance in admission rates across years. We seek to bring the SPD of programs closer to zero, which is challenging because while the share of programs with strongly unequal admission rates is fairly constant across years, there is large variation in the SPD of individual programs over the years. 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 0. This observation is important in practice, as we could be under-correcting or over-correcting inequalities if we assume differences in admission rates will not vary.

3. POLICY DESIGN

To reduce admission rate gaps, we consider the use of bonus policies, which award bonus points to students from disadvantaged groups.

Problem definition. We wish our admission policy to lead to the admission of the students with the highest scores, and to a reduction in admission rate disparities. We thus define the objective function of the bonus policy as a linear combination of both the equality of admission rates, i.e., the SPD, and *utility*, i.e., the scores of the accepted students. 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.

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

The optimization is executed by finding the optimal bonus policy for multiple *application sets*, which are forecasts of potential applicants and their preferences based on historical data. The optimal bonus for a single application set is determined by applying a range of bonus points and directly measuring the objective function. The output of this procedure is simply the average of the optimal bonuses over all application sets that are evaluated. The strategies we consider next either sample application sets from a statistical model learned from historical data; or optimize directly on historical data.

Application sets generated by a predictive model. We build a multi-label probabilistic classifier to model students' application behavior. To predict applications for 2017, we train the model on data from the at the time most recent year, i.e., 2016. We deploy 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 experiment with $n = 50$ and $n = 200$ sampled application sets to evaluate possible bonus policies.

Baseline: directly use historical data. A simpler approach is to compute what bonus policy would have been optimal in 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. Nevertheless, we also include this approach in our empirical evaluation as a baseline.

4. EXPERIMENTAL EVALUATION

We compare strategies based on application sets generated by a predictive model (50 and 200 application sets) and strategies that average the (in hindsight) optimal policies for historical data (1 year, 3 years, and 5 years). We consider two sensitive attributes, gender and income, for the year 2017. To find an appropriate value for λ in Eq. 2, we calculate the median differences in grades and test scores between subgroups: 23 points for gender groups and 28 points for income groups. First, we evaluate the strategies for suggesting policies for all programs and evaluate their overall effect. Second, we evaluate the strategies when policies are only applied to programs which show consistent inequalities in admission rates over time.

All programs. In Table 1, we compare the objective function values resulting from applying the different strategies to the ideal objective function values for all programs. In the table itself, we show the mean and standard distribution (SD) of this difference for both sensitive attributes. Table 1 shows the error in the objective function relative to the smallest achievable ones, i.e. the values that result from applying the ideal bonus policies. Note that, based on Table 1, policies that use more application sets for their suggestions generally lead to both smaller errors and less variance in the error.

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., no bonus, $b = 0$) is almost guaranteed to lead to the lowest utility loss, as utility is typically largest when no affirmative action policy is applied. Table 2 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 through the intervention – which is desirable. In general, we can see that the values suggested through more application sets again exhibit less variance.

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 3). 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 optimal policy in general gives more bonus points. Even though the strategy based on last year's historical data has similar statistics, the previous analysis has shown that this strategy often performs worse than the other strategies. This underlines the need for a more conservative design strategy.

Programs with consistent inequalities. We observe in Table 2 that some design strategies have averages above 0 and that the SDs are large compared to the means. Therefore, sometimes the strategies *increase* the gap of admission rates compared to having no policy – largely due to the unpredictability of admission rates (see Section 2).

In practice, affirmative action policies should not be deployed for all programs, but only sparingly for programs that have consistently unequal admission rates. Hence, we consider programs that (i) have unequal admission rates in each of the three most recent years, where (ii) the admission rate is always lower for the same subgroup, and (iii) the differences are above 0.1, i.e., strong differences, for at least two out of the three years. This results in 12 programs to which a gender policy is applicable and 29 programs for income policies. This means that income tends to generate more consistent and stronger admission rate disparities than gender. Anecdotally, we note that programs with consistent inequalities tend to be large programs with many high-scoring applicants, such as medicine or engineering.

Table 1 and 2 include the values for the filtered programs in the columns on the right. 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.

In general, we note that strategies based on predictive models are less likely to harm an underrepresented group in cases where we observe "flips" in the underrepresentation that sometimes happen from one year to the next one. These changes can to some extent, but not entirely, be anticipated and addressed with more conservative policy suggestions.

Table 1: Error in objective function relative to ideal policies. Smaller values are better.

	All programs				Consistently unequal			
	Gender		Income		Gender		Income	
Strategy	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Historical - 1 year	0.37	1.11	0.44	1.31	1.31	1.88	1.79	1.63
Historical - 3 years	0.30	0.94	0.34	1.06	1.02	1.81	**1.36**	1.62
Historical - 5 years	0.32	0.99	**0.33**	**0.98**	1.22	1.20	1.63	**1.57**
Predictive - 50 sets	**0.28**	**0.90**	0.37	1.13	**0.84**	1.16	2.28	2.12
Predictive - 200 sets	0.29	**0.90**	0.36	1.11	0.91	**1.15**	2.14	1.97

Table 2: Difference in absolute SPD relative to no intervention. Lower values are better.

	All programs				Consistently unequal			
	Gender		Income		Gender		Income	
Strategy	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Historical - 1 year	0.0014	0.0339	-0.0013	0.0383	0.0028	0.0922	-0.0774	0.1000
Historical - 3 years	**-0.0005**	0.0269	-0.0034	0.0266	-0.0153	0.1069	0.0983	**0.0704**
Historical - 5 years	0.0005	0.0234	**-0.0037**	0.0270	-0.0037	0.0858	**-0.0849**	0.0827
Predictive - 50 sets	-0.0003	**0.0124**	-0.0020	**0.0210**	**-0.0156**	0.0586	-0.0625	0.0822
Predictive - 200 sets	-0.0002	0.0126	-0.0023	0.0215	-0.0135	**0.0575**	-0.0670	0.0817

Table 3: Comparison of the number of bonus points given by different strategies.

	Gender		Income	
Strategy	Mean	SD	Mean	SD
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. CONCLUSIONS

We compared a predictive approach to simpler design strategies based on averaging retrospectively optimal bonuses over 1-5 years of historical data. The latter are more likely to over-correct the differences in admission rates. Our proposed predictive approach mostly avoids this pitfall through more conservative suggestions. In practice, 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.

An 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., [3, 7, 13]) suggests that the mere existence of affirmative action policies might encourage students from disadvantaged groups to apply. It is therefore imaginable that less bonus points may suffice to achieve the desired effect: one more reason for preferring conservative strategies.

6. REFERENCES

[1] A. Abdulkadiroğlu. College admissions with affirmative action. *Int. J. of Game Theory*, 33(4):535–549, 2005.

[2] V. R. Bacharach, A. A. Baumeister, and R. M. Furr. Racial and gender science achievement gaps in secondary education. *J. of Genetic Psy.*, 164(1):115–126, 2003.

[3] L. Balafoutas and M. Sutter. Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, 335(6068):579–582, 2012.

[4] P. Barbrook-Johnson, J. Badham, and N. Gilbert. Uses of Agent-Based Modeling for Health Communication. *Health Communication*, 32(8):939–944, 2017.

[5] M. C. Bastarrica, N. Hitschfeld, M. M. Samary, and J. Simmonds. Affirmative action for attracting women to STEM in Chile. In *Proc. Workshop on Gender Equality in Software Engineering*, GE '18, pages 45–48. ACM, 2018.

[6] D. Card and A. B. Krueger. Would the elimination of affirmative action affect highly qualified minority applicants? *ILR Review*, 58(3):416–434, 2005.

[7] L. M. Dickson. Does ending affirmative action in college admissions lower the percent of minority students applying to college? *Econ. of Educ. Rev.*, 25(1):109–119, 2006.

[8] G. W. Evans. The environment of childhood poverty. *American psychologist*, 59(2):77, 2004.

[9] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *arXiv:1609.07236*, 2016.

[10] I. E. Hafalir, M. B. Yenmez, and M. A. Yildirim. Effective affirmative action in school choice. *Theoretical Economics*, 8(2):325–363, 2013.

[11] T. Kawagoe, T. Matsubae, and H. 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.

[12] J. P. Keller, K. C. Desouza, and Y. Lin. Dismantling terrorist networks. *Technological Forecasting and Social Change*, 77(7):1014–1036, 2010.

[13] M. C. Long. College applications and the effect of affirmative action. *J. of Econometrics*, 121(1):319 – 342, 2004.

[14] M. Mathioudakis, C. Castillo, G. Barnabo, and S. Celis. Affirmative action policies for top-k candidates selection, 2019.

[15] P. J. McEwan. The indigenous test score gap in Bolivia and Chile. *Econ. Dev. and Cult. Change*, 53(1):157–190, 2004.

[16] M. I. Meltzer, I. Damon, J. W. LeDuc, and J. D. Millar. Modeling potential responses to smallpox as a bioterrorist weapon. *Emerging infectious diseases*, 7(6):959, 2001.

[17] R. K. E. Bellamy et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

[18] S. F. Reardon. The widening income achievement gap. *Educational leadership*, 70(8):10–16, 2013.

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