Adaptive Priority Mechanisms

Oğuzhan Çelebi[*] Joel P. Flynn[†]
MIT MIT

April 18, 2023

Abstract

How should authorities that care about match quality and diversity allocate resources when they are uncertain of the market they face? We propose a new class of *adaptive priority mechanisms* (APM) that prioritize agents as a function of both scores (reflecting match quality) and the socioeconomic characteristics of the assigned agents (reflecting diversity). For a single authority with separable preferences over scores and characteristics, we derive an APM that is optimal, generates a unique outcome, and depends solely on the authority's preferences. By contrast, the ubiquitous priority and quota mechanisms are optimal if and only if the authority is risk-neutral or extremely risk-averse over diversity, respectively. With many authorities, it is dominant for each of them to use the optimal APM, and each so doing implements the unique stable matching. Using data from Chicago Public Schools, we benchmark the gains from adopting APM and find that they may be considerable.

[*] MIT Department of Economics, 50 Memorial Drive, Cambridge, MA 02142. Email: ocelebi@mit.edu

[†] MIT Department of Economics, 50 Memorial Drive, Cambridge, MA 02142. Email: jpflynn@mit.edu

This paper supersedes our previous work circulated as "Priorities *vs.* Quotas." We are grateful to Daron Acemoglu, Mohammad Akbarpour, George-Marios Angeletos, Jonathan Cohen, Viola Corradini, Roberto Corrao, Mert Demirer, Glenn Ellison, Aytek Erdil, Arda Gitmez, Ömer Karaduman, Stephen Morris, Anh Nguyen, Parag Pathak, Charlie Rafkin, Karthik Sastry, Tayfun Sönmez, Bertan Turhan, Alexander Wolitzky, Bumin Yenmez and seminar participants at the 2022 INFORMS Workshop on Market Design, the 2022 Society for the Advancement of Economic Theory Conference, Iowa State University and the MIT Theory Lunch for helpful comments. We thank Chicago Public Schools for graciously sharing their data and Eryn Heying, Talia Gerstle and Jim Shen for invaluable administrative support. The views expressed here are those of the authors and do not reflect the views of Chicago Public Schools.

1 Introduction

Authorities that allocate resources such as school seats, university places, and medical supplies often face conflicting objectives. On the one hand, they want to maximize match quality or appear fair by allocating resources to the highest-scoring agents according to various criteria such as academic attainment, mortality risk, or distance. On the other hand, they want to achieve diversity across a range of socioeconomic attributes including race, religion, and gender. Resolving this conflict is complicated, especially in new markets, due to uncertainty regarding the distribution of individuals' scores, characteristics, and preferences.

To balance these trade-offs, when the use of prices is seen as infeasible or unethical, authorities have broadly used two classes of policies: *quotas*,[1] where a certain portion of the resource is set aside for given groups; and *priorities*, where individuals in given groups receive higher scores. These policies have been applied across many different markets in many different countries, for example: the Indian government reserves some government jobs for disadvantaged groups; Chicago Public Schools employs quotas for students from different socioeconomic groups at its competitive exam schools; the University of California, Davis instituted a quota system for minority students; many countries gave differential priority to healthcare workers in the receipt of Covid-19 vaccines; church-run schools in the UK give explicit priority points to students from various religious groups; and the University of Michigan and the University of Texas have used different priority scales for minority students.

But what mechanism *should* such an authority use? Despite the practical importance of this question, we do not know if (and under what circumstances) an authority should use a priority mechanism, a quota mechanism, or something else entirely.

In this paper, we formulate and solve the optimal mechanism design problem of an authority that allocates a resource to agents who are heterogeneous in their individual scores and belong to different groups. The authority cares about individuals' *scores*, through some aggregate index, and *diversity*, through the numbers of agents from different groups who are allocated the resource.[2] Moreover, they are uncertain about the market they face and have some beliefs about the joint distribution of scores and groups in the population.

[1] We use *quota* as a general term that includes the widely used reserve policies (see Definition 4).

[2] This diversity preference can be interpreted more generally as encoding a preference of the authority over the composition of assigned agents across a range of attributes, *e.g.,* when allocating medical resources, the authority may care about ensuring that frontline medical workers are treated. Moreover, when scores represent individuals' property rights over objects (*e.g.,* higher-scoring students *deserve* better schools), we can interpret the preference for higher scores as a preference for procedural fairness. Whenever a lower-scoring agent obtains the resource while a higher-scoring agent does not, the latter agent has *justified envy* towards the former. In the two-sided matching literature, justified envy is often seen as inimical to fairness (see *e.g.,* Balinski and Sönmez, 1999).

We propose a new class of *adaptive priority mechanisms* (APM) that adjust agents' scores as a function of the number of assigned agents with the same characteristics and that allocate the resource to the set of agents with the highest adjusted scores. With a single authority, we derive an APM that is optimal, implements a unique outcome, and can be specified solely in terms of the *preferences* of the authority (*i.e.,* it is optimal regardless of their beliefs). By contrast, we show that priorities and quotas are optimal if and only if risk aversion over diversity is extremely low or high, respectively. Moreover, optimally set priority and quota policies depend on both the preferences and beliefs of the authority. Thus, the optimal APM both improves outcomes and requires less information. When there are many authorities, it is dominant for each of them to implement this APM and this leads to the unique stable outcome.

Single-Authority Model. We begin our analysis by studying a setting with a single authority that has some amount of a homogeneous resource (*e.g.,* seats at a school, medical resources) that it can allocate to a continuum of agents.[3] Agents differ in their scores (*e.g.,* exam score, clinical need) and discrete attributes (*e.g.,* socioeconomic status, if they are a frontline health worker). The authority cares separably about some index of the score distribution (*e.g.,* the average score) of those to whom it allocates the resource and the numbers of agents from different groups. As a result, the preferences of the authority over agents depend on the joint distribution of agents' scores and groups. We assume that this distribution is potentially unknown and varies arbitrarily across states of the world. The authority's problem is to design an allocation mechanism that is optimal regardless of their beliefs, a property that we call *first-best optimality.*

Adaptive Priority Mechanisms. To this end, we introduce the class of adaptive priority mechanisms (APM), which proceed in two steps. First, each agent is given an *adaptive priority* that is a function of their own score and the number of agents from the same group to whom the resource is assigned. Second, APM allocate the resource to agents in order of adaptive priorities, subject to fully allocating the available amount. This class of mechanisms allows the implicit preference for agents from different groups to depend upon the ultimate allocation. The allocation under an APM is defined as the fixed point of the above operation: an allocation is implemented by APM if the adaptive priority of all agents who are allocated the resource (evaluated at the allocation) is higher than those who are not allocated the resource. When an agent's adaptive priority is increasing in their own score and decreasing in the number of agents with the same attributes that are assigned the resource – a property we call *monotonicity* – the APM implements a unique allocation. Moreover,

[3] In Appendix G, we generalize our analysis and results to a setting with discrete agents.

this allocation can be computed greedily by prioritizing agents according to their adaptive priority, evaluated at the number of higher-scoring agents in their group.

Most importantly, we derive a particular, monotone APM that is first-best optimal. Under this optimal APM, an agent's priority is equal to the contribution of their own score plus their marginal contribution to diversity utility. Intuitively, this mechanism equates the benefits and costs of allocating to the marginal agent, regardless of the ultimate joint distribution of agents' scores and groups. Moreover, this APM can be described solely as a function of the authority's preferences, without any reference to its beliefs.

(Sub)Optimality of Priorities and Quotas. We next establish that priority and quota mechanisms are generally dominated by APM. We do so by characterizing the conditions on the preferences of the authority such that priorities and quotas attain first-best optimality. Concretely, we find that priorities and quotas are first-best optimal if and only if (i) the authority is risk-neutral over diversity, in which case priorities are optimal, or (ii) the authority is extremely risk-averse over diversity, in which case quotas are optimal. Hence, outside of extreme cases, APM deliver strict improvements relative to the *status quo*.

A Price-Theoretic Intuition. To both illustrate and develop the intuition behind these results, we study a detailed example that allows for a closed-form comparison of priorities, quotas, and the optimal adaptive priority mechanism. We do this in the spirit of the seminal analysis of Weitzman (1974), who compares price and quantity regulation in product markets. In the example, the resource corresponds to seats at a school and there are two groups of students (minority and majority students). The authority is uncertain over the relative scores of minority and majority students, and has linear-quadratic preferences over the scores of admitted students and the number of minority students admitted to the school.

The preference of the authority between priority and quota mechanisms is governed by its risk aversion over the number of admitted minority students: there is a cutoff value such that quotas are preferred when risk aversion exceeds this threshold and priorities are otherwise preferred. On the one hand, by mandating a minimal level of minority admissions, quotas *guarantee* a level of diversity. On the other hand, as relatively more minority students receive the resource in the states in which they have relatively higher scores, priorities *positively select* minority students. Adaptive priority mechanisms optimally exploit the guarantee effects of quotas and the positive selection effects of priorities, and are always optimal.

Dominance and Stability with Multiple Authorities. While the single-authority model is relevant for studying settings with a single resource, in many markets there are multiple authorities who control heterogeneous resources (*e.g.,* school seats) over which agents have heterogeneous preferences. We generalize our analysis to this setting and show that

APM arise under both cooperative (stability) and non-cooperative (dominant-strategy equilibrium) solution concepts. We first show that there is a unique stable allocation. This constitutes a methodological contribution as we establish the uniqueness of stable allocations in continuum economies in which authorities have preferences between agents that are *endogenous* to the ultimate allocations (extending Theorem 1.1 of Azevedo and Leshno, 2016). Moreover, we characterize this unique stable allocation and show that a mechanism is consistent with stability if and only if it coincides with the single-authority-optimal APM. Furthermore, when authorities sequentially admit agents, each authority using its single-authority-optimal APM is a dominant strategy and implements the unique stable matching. Thus, APM are not only privately optimal but also inherent to the structure of stability. These results imply that one could advise authorities to use APMs with confidence that outcomes will be stable and that they could do no better under any alternative mechanism.

Benchmarking the Gains from APM in Chicago Exam Schools. To obtain a sense of the practical benefit of using APM, we benchmark the improvements from APM using application and admission data from 2013-2017 on the selective exam schools of Chicago Public Schools (CPS), a setting also empirically studied by Angrist, Pathak, and Zárate (2019) and Ellison and Pathak (2021). CPS uses a reserve system to increase the admissions of underrepresented groups. In this system, as we later detail, academic scores and the socioeconomic tiers of the census tracts in which students live determine the schools that students can attend. Estimating preference parameters to best rationalize the pursued reserve policy, we find that the gains from using the optimal APM are equivalent to eliminating 37.5% of the loss to CPS' payoffs from failing to admit a diverse class of students. This gain is 2.3 times larger than the estimated gain from a 2012 policy change that increased the size of all reserves. This proof-of-concept exercise shows both that APM could be practically implemented and that the gains from so doing may be considerable.

Related Literature. The market design literature has largely studied the comparative statics and axiomatic foundations of mechanisms. In this context, our paper relates to the literature on matching with affirmative action concerns initiated by Abdulkadiroğlu and Sönmez (2003) and Abdulkadiroğlu (2005). For example, in the study of quotas, Kojima (2012) shows how affirmative action policies that place an upper bound on the enrollment of non-minority students may hurt all students, Hafalir, Yenmez, and Yildirim (2013) introduce the alternative and more efficient minority reserve policies, Ehlers, Hafalir, Yenmez, and Yildirim (2014) generalize reserves to accommodate policies that have floors and ceilings for minority admissions, and Doğan (2016) shows that stronger affirmative action can (weakly) harm all minority students under reserve policies and proposes a new rule that fixes this issue. The quota policies studied in this paper are a special case of the slot-specific priorities introduced in Kominers and Sönmez (2016). Further related papers study quota policies in university admissions in India (Aygün and Turhan, 2020; Sönmez and Yenmez, 2022a,b), in Germany (Westkamp, 2013) and in Brazil (Aygun and Bó, 2021). Kamada and Kojima (2017, 2018) and Goto, Kojima, Kurata, Tamura, and Yokoo (2017) study stability and efficiency in more general matching-with-constraints models. Echenique and Yenmez (2015) characterize a class of substitutable choice rules under diversity preferences, Erdil and Kumano (2019) study tie-breaking rules under substitutable priorities under stable matching mechanisms and distributional constraints, and Imamura (2020) presents axioms to compare the meritocracy and diversity of different choice rules and characterizes reserves and quotas. Çelebi (2022) studies when affirmative action policies, including quota policies, can be rationalized by diversity preferences.

In this paper, we instead pursue the methodological approach of mechanism design and welfare economics by analyzing optimal mechanisms from the perspective of an authority with some given preferences over allocations. Chan and Eyster (2003) share this perspective in their analysis of the costs and benefits of banning affirmative action.[4] In this vein, we have previously analyzed the narrower problem of how to optimally coarsen agents' scores into priorities (Çelebi and Flynn, 2022) in a continuum matching market framework in the style of Abdulkadiroğlu, Che, and Yasuda (2015) and Azevedo and Leshno (2016). This analysis nevertheless restricted authorities to use a priority mechanism that does not consider agents' characteristics and implement only allocations that are stable with respect to these priorities. Thus, our focus on comparing priorities, quotas, and optimal mechanisms distinguishes our analysis from our prior work and the previous literature, which study the properties of each policy in isolation and without an explicit treatment of uncertainty.

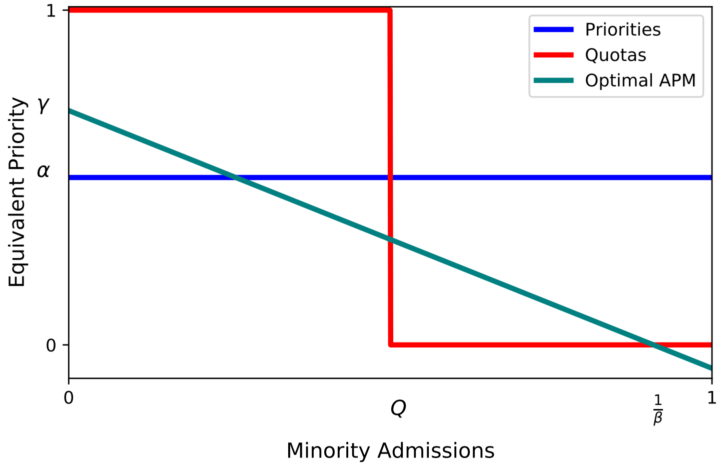
Outline. Section 2 exemplifies our main results. Section 3 studies optimal mechanisms with a single authority. Section 4 studies equilibrium mechanisms with many authorities. Section 5 quantifies the gains from APM using data from Chicago Public Schools. Section 6 concludes. The proofs of all Theorems are in Appendix A and the proofs of Propositions are in Appendix B.

2 Comparing Mechanisms: An Example

The Setting. A single school has capacity q. Students are of unit total measure, have scores in $[0, 1]$, and are either minority or majority students. The authority has linear-

[4] Other analyses of this issue include Epple, Romano, and Sieg (2008) and Temnyalov (2023).

Figure 1: How Priorities, Quotas, and APM Prioritize Minority Students



Note: Illustration of the equivalent priority given to a minority student as a function of the measure of admitted minority students under: the optimal APM (see Proposition 1), a priority mechanism α, and a quota mechanism Q.

quadratic preferences $\xi : \mathbb{R}^2 \to \mathbb{R}$ over students' total scores $\bar{s}$ and the measure of admitted minority students x:

$$\xi(\bar{s}, x) = \bar{s} + \gamma \left(x - \frac{\beta}{2} x^2 \right) \tag{1}$$

where $\gamma \geq 0$ indexes their general concern for admitting minority students relative to ensuring high scores and $\beta \geq 0$ indexes the degree of risk aversion regarding the measure of admitted minority students.

The minority students are of measure κ and have scores that are uniform over $[0, 1]$. The majority students are of measure $1 - \kappa$ and all have common underlying score $\omega \in [\underline{\omega}, \overline{\omega}] \subseteq [0, 1]$ with distribution Λ. The score of the majority students, ω, parameterizes how well the majority students score relative to the minority students. Finally, we assume that the affirmative action preference is neither too small nor too large with the following: $\min\{\kappa, q\} > \frac{1+\gamma-\underline{\omega}}{\frac{1}{\kappa}+\gamma\beta} + \kappa(\overline{\omega} - \underline{\omega})$, $\kappa(1 - \underline{\omega}) < \frac{1+\gamma-\overline{\omega}}{\frac{1}{\kappa}+\gamma\beta}$. These conditions ensure that optimal affirmative action policies will neither be so large as to award all slots to minority students in some states nor so small that there is no affirmative action in some states.

The authority can implement an APM, a priority mechanism, or a quota mechanism. An APM awards a score boost of $A(y)$ to a minority student when measure y other minority students are admitted, and allocates the seats to the highest-scoring minority students who have transformed scores higher than ω.[5] An (additive) priority mechanism $\alpha \in \mathbb{R}_+$ increases uniformly the scores of minority students: the score used in admissions becomes uniform

[5] Formally, this happens when $s(x(\omega)) + A(x(\omega)) = \omega$, where $s(x(\omega))$ denotes the score of the marginal minority student when the highest-scoring $x(\omega)$ minority students are admitted.

over $[\alpha, 1+\alpha]$. The authority then admits the highest-scoring measure q students. A quota policy $Q \in [0, \min\{\kappa, q\}]$ sets aside measure Q of the capacity for the minority students. The measure Q highest-scoring minority students are first allocated to quota slots, and all other agents are then admitted to the residual $q - Q$ places according to the underlying score.[6]

We illustrate how these three policies prioritize minority students in Figure 1. Priority mechanisms award a constant score boost of α. Quota mechanisms give enough points to always ensure admission until measure Q is reached and then give no advantage. APM allow any pattern of prioritization as a function of minority admissions (we plot only the optimal APM, which turns out to be linear in this context).

Comparing Mechanisms. Let the authority's expected utility be V^* under any optimal (expected utility maximizing) mechanism, V_A under an optimal adaptive priority mechanism, V_P under an optimal priority mechanism, and V_Q under an optimal quota mechanism. The following proposition characterizes the relationships between these mechanisms:

Proposition 1. *The following statements are true:*

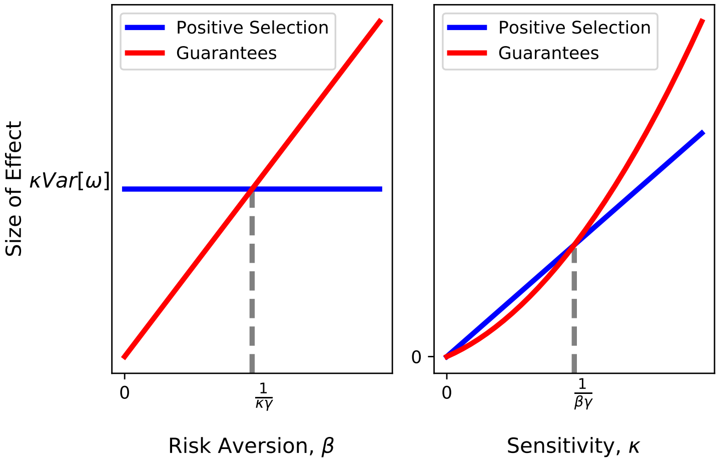
1. *The APM $A(y) = \gamma(1 - \beta y)$ is optimal, $V^* = V_A$*
2. *The comparative advantage of priorities over quotas is given by:*
$$\Delta \equiv V_P - V_Q = \frac{\kappa}{2}\left(1 - \kappa\gamma\beta\right) \mathrm{Var}[\omega] \tag{2}$$
3. *The comparative advantage of APM over priorities and quotas is given by:*
$$\Delta^* \equiv \min\{V^* - V_P, V^* - V_Q\} = \begin{cases} \frac{1}{2}\left(\kappa\gamma\beta\right)^2 \frac{\kappa \mathrm{Var}[\omega]}{1+\kappa\gamma\beta}, & \kappa\gamma\beta \leq 1, \\ \frac{1}{2}\frac{\kappa \mathrm{Var}[\omega]}{1+\kappa\gamma\beta}, & \kappa\gamma\beta > 1. \end{cases} \tag{3}$$

We now develop intuition for the comparative advantage of priorities over quotas. First, observe that a quota of Q admits measure Q minority students in all states of the world under our assumptions. However, a priority policy induces variability in the measure of admitted minority students across states of the world. We call the gain to quota policies in eliminating this variation the *guarantee effect* and find mathematically that it is equal to $\frac{\kappa}{2}\left(1 + \kappa\gamma\beta\right) \mathrm{Var}[\omega]$ in payoff terms.

Second, the optimal priority policy admits more minority students when minority students score relatively well and fewer when minority students score relatively poorly. To demonstrate this, we show that minority admissions in state ω under the optimal priority policy are $x(\alpha, \omega) = \bar{x}(\alpha) + \varepsilon(\omega)$ where $\bar{x}(\alpha) = \kappa(1 + \alpha - \mathbb{E}[\omega])$ and $\varepsilon(\omega) = \kappa\left(\mathbb{E}[\omega] - \omega\right)$. Thus, in the states where minority students score relatively better ($\omega < \mathbb{E}[\omega]$), we have that

[6] This corresponds to a precedence order that processes quota slots first. We discuss the importance of precedence orders in Section 2.1 and in Appendix C.3.

Figure 2: Comparative Statics for the Positive Selection and Guarantee Effects



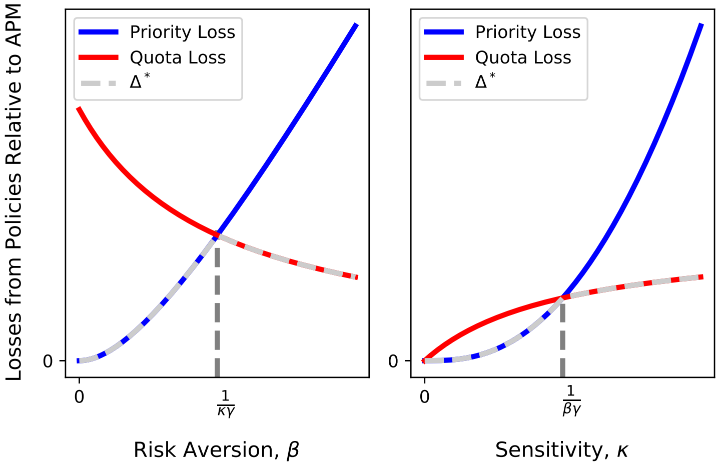
Note: Illustration of the comparative statics for the trade-offs between priority and quota mechanisms. Positive Selection plots the positive selection effect, $\kappa\mathrm{Var}[\omega]$, and Guarantee plots the guarantee effect, $\frac{\kappa}{2}\left(1+\kappa\gamma\beta\right)\mathrm{Var}[\omega]$. As per Equation 2 in Proposition 1, priorities dominate quotas if and only if $1 \geq \kappa\gamma\beta$, where the point of indifference is denoted by the dashed grey line.

$\varepsilon(\omega) > 0$ and $x(\alpha,\omega) > \bar{x}(\alpha)$. We call this effect the *positive selection* effect and find that this benefits a priority policy by $-\mathrm{Cov}[\omega,\varepsilon(\omega)] = \kappa\mathrm{Var}[\omega]$ in payoff terms.

The ultimate preference between priority and quota mechanisms is determined by which of the guarantee and positive selection effects dominates. This is itself determined by the extent to which the authority values diversity γ, the risk preferences of the authority β, and the measure of minority students κ. We illustrate how risk aversion and the measure of minority students affect the sizes of the positive selection and guarantee effects in Figure 2. If the authority is close enough to risk-neutral (*i.e.*, $\frac{1}{\kappa\gamma} > \beta$), then priorities are strictly preferred as positive selection dominates guarantees. If the authority is sufficiently risk-averse (*i.e.*, $\frac{1}{\kappa\gamma} < \beta$), then quotas are strictly preferred as the guarantee effects dominate positive selection. The threshold for risk aversion scales inversely with the measure of minority students κ. Because minority students' scores are uniform, κ corresponds to the density of minority students' scores. Hence, the change in minority admissions from a small change in their priority equals κ. Thus, κ indexes the *sensitivity* of minority admissions to the state under priority policies. As a result, higher κ favors quota policies by increasing the magnitude of the guarantee effect relative to the positive selection effect. Finally, the extent of uncertainty $\mathrm{Var}[\omega]$ may intensify an underlying preference but never determines which regime is preferred.

An APM is optimal and overcomes the limitations posed by both priorities and quotas. In this case, the optimal APM is linear in the measure of admitted minority students, with slope given by the authority's risk aversion over minority admissions, awarding each minority

Figure 3: Comparative Statics for the Losses from Priorities and Quotas



Note: Illustration of the comparative statics for the losses from optimal priority and quota policies relative to the optimal APM (as presented in Equation 3 in Proposition 1). The lower envelope of the losses, Δ^*, corresponds to the comparative advantage of the optimal APM over priorities and quotas. The point of indifference between priorities and quotas is denoted by the dashed grey line.

student a subsidy equivalent to their marginal contribution to the diversity preferences of the authority. This allows the adaptive priorities to optimally balance the positive selection and guarantee effects, and implement the first-best allocation in every state. In Figure 3, we illustrate how the losses from priority mechanisms and quota mechanisms vary with risk aversion and sensitivity. As risk aversion moves, the loss from priority and quota policies relative to the optimum is greatest when the authority is indifferent between the two regimes. The loss from restricting to priority or quota policies is zero when the authority is risk-neutral or there is no uncertainty regarding relative scores, and decreases as the authority becomes extremely risk-averse. As sensitivity increases, the scope for affirmative action increases and so the gains from APM also increase. Thus, we should expect there to be large gains from switching to APM precisely when authorities have intermediate levels of risk aversion and/or the scope for implementing affirmative action is significant.

Finally, optimal APM have an advantage with respect to priority and quota mechanisms that we have not yet highlighted: they depend only on the authority's preferences, γ and β, and not their beliefs about ω, Λ. This contrasts with the optimal priority and quota policies, which depend on Λ.[7] As a result, APM improve outcomes while using *less* information.

[7] The optimal quota policy is given by $Q^* = \frac{1+\gamma-\mathbb{E}[\omega]}{\frac{1}{\kappa}+\gamma\beta}$, while the optimal priority policy sets the expected measure of minorities to Q^*. The policies depend on Λ through $\mathbb{E}[\omega]$.

2.1 Discussion

Before moving to the general analysis, we discuss three additional findings that emphasize the broader economics and scope of these results.

A Price-Theoretic Intuition. This comparison of *priorities vs. quotas* echoes the comparison of *prices vs. quantitities* by Weitzman (1974). We show in Appendix C.1 that there is a formal mapping between the two.[8] Intuitively, the positive selection effect is equivalent to the effect that price regulation gives rise to the greatest production in states where the firm's marginal cost is lowest. Moreover, the guarantee effect is equivalent to the ability of quantity regulation to stabilize the level of production. An APM corresponds in the Weitzman (1974) setting to a regulator setting neither a price nor a quantity, but completely specifying the optimal demand curve. As in Weitzman's analysis, this allows the authority to implement the optimal allocation, regardless of the firm's realized marginal costs. Thus, in this context, the comparison of mechanisms for allocating goods without prices boils down to similar trade-offs between well-understood price-based mechanisms for goods allocation.

Medical Resource Allocation. In Appendix C.2, we apply this model to understand the trade-offs between priority and quotas in the context of medical resource allocation. This topic received enormous attention during the Covid-19 pandemic (see *e.g.,* Pathak, Sönmez, Unver, and Yenmez, 2021). Our analysis provides a formal justification for the idea that priorities may lose out relative to quotas from ignoring some groups or ethical values in the allocation of scarce resources (the guarantee effect). However, we also uncover a benefit of priorities that was not previously understood: they induce positive selection. Thus, if we care mostly about treating the neediest (β is low), priorities may yet be optimal.[9]

Optimal Precedence Orders. We have modelled quotas by first allocating minority students to quota slots and then allocating all remaining students according to the underlying score. However, we could have done the opposite. The orders in which quotas are processed are called *precedence orders* in the matching literature and their importance has been the subject of a growing literature (see *e.g.,* Dur, Kominers, Pathak, and Sönmez, 2018; Dur, Pathak, and Sönmez, 2020; Pathak, Rees-Jones, and Sönmez, 2022). In Corollary 1 in

[8] Mapping Weitzman's curvature of production costs $C''^{-1} \mapsto \kappa$, curvature of benefits to consumers $B'' \mapsto -\gamma\beta_m$, and uncertainty over marginal costs $\mathrm{Var}[\alpha(\theta)] \mapsto \mathrm{Var}[\omega]$, we have that Weitzman's Δ coincides with our own. See Appendix C.1 for more details.

[9] Inspired by the uncertainty over the number of medical professionals and the general population who were expected to need scarce medical resources at the onset of the Covid-19 pandemic, we consider an extension where there is also uncertainty over the number of medical professionals who get sick, κ. We show that an increase in the uncertainty regarding the need of frontline workers $\mathrm{Var}[\kappa]$ leads to a greater preference for quotas. This can be seen diagrammatically in Figure 2 (and is formally shown in Appendix C.2), as the guarantee effect is convex in κ.

Appendix C.3, we show that processing quotas second is equivalent to using a priority policy in this setting. Thus, processing quotas first is better than processing them second if and only if $1 \leq \kappa\gamma\beta$. We emphasize that this equivalence merely illustrates the similarity between priority policies and processing quotas second and does not hold in the more general model we study in the remainder of the paper. This notwithstanding, the main aspect of this conclusion is robust: in Theorem 2, we show that for any quota policy to be optimal in the presence of uncertainty, it must process quotas first. In the absence of uncertainty, in Appendix H, we show that priority and quota mechanisms are equivalent and use this to quantify the impact of changes in precedence orders for US H1-B visa allocation.

3 Optimal Mechanisms with a Single Authority

We begin our general analysis by studying the resource allocation problem of a single authority. In this context, we define APM and derive an optimal APM that attains the first-best. We moreover provide necessary and sufficient conditions for the optimality of the ubiquitous priority and quota mechanisms and find that they are extremely restrictive, implying that there are likely gains from switching to APM.

3.1 Model

An authority allocates a single resource of measure $q \in (0, 1)$. Agents differ in their type $\theta \in \Theta = [0, 1] \times \mathcal{M}$ comprising their scores $s \in [0, 1]$ and the group to which they belong, $m \in \mathcal{M}$, where their score denotes their suitability for the resource and $\mathcal{M}$ is a finite set comprising potential attributes such as race, gender, or socioeconomic status. We denote the score and group of any type θ by $s(\theta)$ and $m(\theta)$, respectively. The true distribution of types is unknown to the authority. The authority's uncertainty is parameterized by $\omega \in \Omega$, where Ω is the set of all distributions over Θ that admit a density. The authority believes that ω has distribution $\Lambda \in \Delta(\Omega)$. In state of the world ω, we denote the measure of types by F_ω with density f_ω.[10] In Appendix G, we translate our analysis and results to the discrete context.

An allocation $\mu : \Theta \to \{0, 1\}$ specifies for any type $\theta \in \Theta$ whether they are assigned to the resource.[11] Two allocations μ and μ' are *essentially the same* if they coincide up to a measure zero set. The set of possible allocations is $\mathcal{U}$. An allocation is feasible if it allocates no more than measure q of the resource. A mechanism is a function $\phi : \Omega \to \mathcal{U}$ that returns a feasible allocation for any possible measure of types.

[10] Formally, we mean that $f_\omega(s, m) = \frac{\partial}{\partial s} F_\omega(s, m)$ exists for all $s \in [0, 1]$ and $m \in \mathcal{M}$.

[11] Formally, μ is a measurable function with respect to the Borel $\sigma-$algebra of the product topology in Θ.

As motivated, authorities often have preferences over scores and diversity. To model this, we define the aggregate score index of any allocation as:

$$\bar{s}_h(\mu,\omega) = \int_\Theta \mu(s,m)h(s)\mathrm{d}F_\omega(s,m) \tag{4}$$

for some continuous, strictly increasing function $h : [0,1] \to \mathbb{R}_+$, which determines the extent to which the authority values agents with higher scores. To capture diversity, we compute the measure of agents of each group allocated the resource $x(\mu,\omega) = \{x_m(\mu,\omega)\}_{m\in\mathcal{M}}$ as:

$$x_m(\mu,\omega) = \int_{[0,1]} \mu(s,m)f_\omega(s,m)\mathrm{d}s \tag{5}$$

To separate the roles of scores and diversity, we impose that their utility function over these dimensions $\xi : \mathbb{R}^{|\mathcal{M}|+1} \to \mathbb{R}$ satisfies the following separability assumption:[12]

Assumption 1. *The authority's utility function can be represented as:*

$$\xi\left(\bar{s}_h, x\right) \equiv g\left(\bar{s}_h + \sum_{m\in\mathcal{M}} u_m(x_m)\right) \tag{6}$$

for some continuous, strictly increasing function $g : \mathbb{R} \to \mathbb{R}$ and differentiable and concave functions $u_m : \mathbb{R} \to \mathbb{R}$ for all $m \in \mathcal{M}$.

We also assume that the authority always prefers to allocate the entire resource.[13] The preference of the authority is a monotone transformation of a quasi-linear utility index comprised of scores and a diversity preference. Intuitively, u_m determines the preference for assigned agents of group m, with its concavity following from a preference for diversity.[14] The function g determines their risk preferences over their utility over scores and diversity across states of the world.

As we later show, this assumption allows for particularly simple functional forms for optimal mechanisms. We explore the robustness of our results to relaxing this assumption in Appendix F. We show that our results are essentially unchanged when preferences are non-separable over diversity, *i.e.,* when $\sum_{m\in\mathcal{M}} u_m(x_m)$ is replaced with $u(x)$. The essential

[12] The assumption of separable preferences over scores and diversity is common in the literature on affirmative action concerns (see e.g., Athey, Avery, and Zemsky, 2000; Chan and Eyster, 2003; Ellison and Pathak, 2021).

[13] A necessary and sufficient condition for this is: $h(0) + u'_m(q) \geq 0$ for all $m \in \mathcal{M}$. This condition is clearly sufficient, the lowest utility the authority can get from allocating the resource is always positive. It is also necessary: if $h(0) + u'_m(q) < 0$ for some m, in the state of the world where there are only measure q of group m agents with uniform score distribution, the authority would prefer not to allocate a portion of the resource to the lowest-scoring agents.

[14] Note that u_m depends on m, so our specification allows the designer to have different preferences for allocating the resource to agents from different groups. For example, this allows for a designer with affirmative action motives who prefers to assign the resource to some particular group m: $u'_m(x) > u'_{m'}(x)$ for all x or a designer who prefers a balanced composition of allocated agents: $u'_m(x) = u'_{m'}(x)$ for all $m \in \mathcal{M}$.

assumption for our results is the separability of diversity and score preferences. When this fails, it is no longer possible to specify optimal mechanisms without explicitly conditioning the allocation on the realized distribution of agents. As a result, our analysis will not apply to situations in which there is complementarity or substitutability in preferences over match quality and diversity.

We define the value of a mechanism ϕ under distribution Λ as the authority's expected utility of the allocations induced by that mechanism:

$$\Xi(\phi, \Lambda) = \int_{\Omega} \xi(\bar{s}_h(\phi(\omega), \omega), x(\phi(\omega), \omega)) \mathrm{d}\Lambda(\omega) \tag{7}$$

We say that a mechanism is first-best optimal if it maximizes the authority's expected utility for all possible *distributions of* measures of agents' characteristics.

Definition 1 (First-Best Optimality)**.** *A mechanism ϕ^* is first-best optimal if:*

$$\Xi(\phi^*, \Lambda) = \sup_{\phi} \Xi(\phi, \Lambda) \tag{8}$$

for all $\Lambda \in \Delta(\Omega)$.

This is a demanding property for a mechanism to possess. Indeed, as the example from Section 2 shows, priority and quota mechanisms can fail to be first-best optimal while APM can attain first-best optimality. In the remainder of this section, we formally define APM, show that (when suitably designed) they are first-best optimal, and characterize the conditions under which priorities and quotas are first-best optimal.

3.2 Adaptive Priority Mechanisms

Toward deriving a first-best optimal mechanism, we introduce APMs. To this end, we first introduce an *adaptive priority policy* $A = \{A_m\}_{m \in \mathcal{M}}$, where $A_m : \mathbb{R} \times [0, 1] \to \mathbb{R}$. The adaptive priority policy assigns priority $A_m(y_m, s)$ to an agent with score s in group m when measure y_m of agents of the same group is allocated the object. Given an adaptive priority policy, an APM implements allocations in the following way:

Definition 2 (Adaptive Priority Mechanism)**.** *An adaptive priority mechanism, induced by an adaptive priority A, implements an allocation μ in state ω if the following are satisfied:*

1. *Allocations are in order of priorities: $\mu(\theta) = 1$ if and only if for all θ' with $\mu(\theta') = 0$, we have that:*

$$A_{m(\theta)}(x_{m(\theta)}(\mu, \omega), s(\theta)) > A_{m(\theta')}(x_{m(\theta')}(\mu, \omega), s(\theta')) \tag{9}$$

2. The resource is fully allocated:

$$\sum_{m \in \mathcal{M}} x_m(\mu, \omega) = q \tag{10}$$

With some abuse of terminology, we will often refer to an APM as the adaptive priority A that induces it. By way of illustration, we provide a simple example of the flexibility of APM to act like a hybrid of priority and quota policies.

Example 1. Let $\mathcal{M} = \{m, n\}$. Both groups have measure 0.5 and capacity is $q = 0.5$. We consider the adaptive priority policy $A = \{A_m, A_n\}$ given by:

$$A_m(x, s) = s, \qquad A_n(x, s) = \begin{cases} s + 1 & \text{if } x \leq 0.1 \\ s + 0.1 & \text{if } x \in (0.1, 0.25) \\ s & \text{if } x \geq 0.25 \end{cases} \tag{11}$$

This leaves the score of group m agents unchanged and gives agents of group n a score boost of: 1 if less than measure 0.1 group n agents is assigned, 0.1 if between measure 0.1 and 0.25 group n agents is assigned, and no score boost at all if measure greater than 0.25 group n agents is assigned.

To understand the properties of this adaptive priority policy, observe that the highest-scoring measure 0.1 group n agents is guaranteed the resource, even in states where they score badly. Therefore, A_n practically embeds a quota of 0.1. For admissions levels between 0.1 and 0.25, the APM acts like a priority policy and boosts the scores of group n agents by 0.1, increasing the admissions of group n when group n agents score moderately well. For admissions levels beyond 0.25, group n agents are given no further advantage. Thus, when diversity is attained, this APM "phases out" and no longer advantages any group. △

At this point, we have not established that a given APM implements any allocation at all, or that it implements a unique allocation (*i.e.,* it may not even be a mechanism). However, there is a natural subclass of APM that do implement a unique allocation: those that are monotone. An APM A is *monotone* when (i) $A_m(\cdot, s)$ is a decreasing function for all $m \in \mathcal{M}, s \in [0, 1]$ and (ii) $A_m(y_m, \cdot)$ is a strictly increasing function for all $m \in \mathcal{M}, y_m \in \mathbb{R}$.[15]

Proposition 2. *Any Monotone APM A implements an essentially unique allocation.*

Moreover, the unique outcome of a monotone APM can be implemented "greedily:"[16]

[15] Observe that monotone adaptive priority mechanisms are fair in the sense that they preserve the ranking of agents within any group and assign higher priority to an agent whenever there are fewer agents from her group who are allocated the resource.

[16] Formally, when we consider an Adaptive Priority *Mechanism*, we are studying any selection from the set of allocations that the APM implements. As monotone APMs implement an essentially unique allocation, this is without loss of optimality. When we refer to the "unique" allocation, we refer to the cutoff allocation defined in the proof of Proposition 2 and which is implemented by Algorithm 1.

Algorithm 1 (Greedy Algorithm for Implementation of APM)**.** *The greedy APM algorithm proceeds in the following four steps:*

1. *For each θ, define*
$$\overline{x}(\theta) = \int_{s(\theta)}^{1} f_\omega(s, m(\theta))\mathrm{d}s \tag{12}$$
as the measure of agents who have higher scores than θ and belong to same group.
2. *Construct a ranking of the agents as*
$$R(\theta) = A_{m(\theta)}(\overline{x}(\theta), s(\theta)) \tag{13}$$
3. *Define the cutoff ranking for the agents as $\overline{R}_\omega$ by*
$$\int_{\Theta} \mathbb{I}\{R(\theta) \geq \overline{R}_\omega\}\mathrm{d}F_\omega(\theta) = q \tag{14}$$
4. *Allocate the resource to all θ with $R(\theta) \geq \overline{R}_\omega$.*

Intuitively, this algorithm works by ranking all agents by their score within each group m and assigning agents in order of their transformed scores evaluated at the measure of *already assigned* agents of the same group, conditional on their admission. Informally, the algorithm greedily moves down the ranking of agents until the resource is exhausted.

3.3 Adaptive Priority Mechanisms Achieve the First-Best

Having shown that monotone APM implement a unique allocation and provided an algorithm to compute this allocation, we now show that a certain, monotone APM is first-best optimal:

Theorem 1. *The following APM is monotone and first-best-optimal:*
$$A^*_m(y_m, s) \equiv h^{-1}(h(s) + u'_m(y_m)) \tag{15}$$
Moreover, if a mechanism is first-best-optimal, then it implements essentially the same allocations as A^*.

Observe that A^* is not only uniquely first-best optimal, it also requires only that the authority knows its preferences over scores h and diversity u_m. Importantly, it need have no knowledge of the underlying distribution of agents and can be fully specified even without any knowledge of the nature or extent of uncertainty, Λ. Moreover, this mechanism does not depend at all on the authority's across-state risk preferences, g. This is because it achieves the *ex post* optimal allocation in all states and so there is no need to trade-off gains and losses across states (the preferences over which are exactly determined by g).

To gain intuition for the form of this mechanism, suppose that the authority has linear utility over scores $h(s) \equiv s$. In this case, $A^*_m(y_m, s) = s + u'_m(y_m)$, so an agent in group m is awarded a boost of $u'_m(y_m)$ when there are y_m higher-scoring agents of the same group, their direct marginal contribution to the diversity preferences of the authority. This is optimal because this boost precisely trades off the marginal benefit of additional diversity with the marginal costs of reduced scores. Moreover, failing to award this precise level of boost would result in a suboptimal allocation. Thus, any optimal mechanism must be essentially identical to the optimal APM we have characterized. To generalize this beyond linear utility of scores, consider the following observation: we can map agents' scores from s to $h(s)$, and consider the optimal boost in this space. As h is monotone, this preserves the ordinal structure of the optimal allocation, and the authority has linear preferences over $h(s)$. Thus, in this transformed space, the optimal boost remains additive and given by $u'_m(y_m)$. To find the optimal transformed score in the original space, we simply invert the transformation h and apply it to the optimal score in the transformed space, yielding the formula for the optimal mechanism in Theorem 1.

3.4 (Sub)Optimality of Priorities and Quotas

We have shown that APM are optimal. However, the primary classes of mechanisms that have been used in practice are priority and quota mechanisms. Therefore, it is important to understand whether (and when) these mechanisms are also optimal. We now establish that APM generally provide a strict improvement over priority and quota mechanisms.

We first formally define priority and quota mechanisms. A *priority policy* $P : \Theta \to [0, 1]$ awards an agent of type $\theta \in \Theta$ a priority $P(\theta)$.

Definition 3 (Priority Mechanisms)**.** *A priority mechanism, induced by a priority policy P, allocates the resource in order of priorities until measure q has been allocated, with ties broken uniformly and at random.*

We define a *quota policy* as (Q, D), where $Q = \{Q_m\}_{m\in\mathcal{M}}$ and $D : \mathcal{M} \cup \{R\} \to \{1, 2, \ldots, |\mathcal{M}| + 1\}$ is a bijection. The vector Q reserves measure of the capacity Q_m for agents in group m, with residual capacity $Q_R = q - \sum_{m\in\mathcal{M}} Q_m$ open to agents of all types. The bijection D (the precedence order) gives the order in which the groups are processed.

Definition 4 (Quota Mechanisms)**.** *A quota mechanism, induced by a quota policy (Q, D), proceeds by allocating the measure $Q_{D^{-1}(k)}$ agents from group $D^{-1}(k)$ (if there are sufficient agents from this group) to the resource in ascending order of k, and in descending order of score within each k. If there are insufficiently many agents of any group to fill the quota, the residual capacity is allocated to a final round in which all agents are eligible.*

We now characterize when priority and quota mechanisms are (sub)optimal. To do this, we first provide some definitions. Authority preferences are *non-trivial* if for all $m, n \in \mathcal{M}$, we have that:

$$h(1) + u'_n(0) > h(0) + u'_m(q) \tag{16}$$

Intuitively, the authority's preferences are non-trivial when their concerns for representation of certain groups do not always outweigh the consideration of scores.[17] The authority is *risk-neutral* over diversity if for all $m \in \mathcal{M}$, $u'_m : [0, q] \to \mathbb{R}$ is constant, *i.e.,* there are constant marginal returns to admitting more agents from all groups. If there are decreasing marginal returns, then the authority's preferences feature risk aversion. To define extremely risk-averse preferences, let $\tilde{u}$, $\tilde{h}$ and $\{x_m^{\text{tar}}\}_{m \in \mathcal{M}}$ be as follows: (i) $\tilde{u}'_m(x_m) = 0$ for all $x_m > x_m^{\text{tar}}$ (ii) $\tilde{u}'_m(x_m) \geq \tilde{h}(1) - \tilde{h}(0)$ for $x_m \leq x_m^{\text{tar}}$ and (iii) $\sum_{m \in \mathcal{M}} x_m^{\text{tar}} \leq q$. Intuitively, an authority whose preferences are represented by $\tilde{u}$ and $\tilde{h}$ is very risk-averse as the condition that $\tilde{u}'_m(x_m) \geq \tilde{h}(1) - \tilde{h}(0)$ implies that the loss from being below the target level for a group x_m^{tar} dominates any benefit from increased scores. Thus, they are infinitely risk-averse to failing to meet this target. We say that the authority is *extremely risk-averse* if the authority's preferences over the optimal allocations can be represented by $(\tilde{u}, \tilde{h})$.[18]

Theorem 2. *Suppose that the authority has non-trivial preferences. The following statements are true:*

1. *There exists a first-best optimal priority mechanism if and only if the authority is risk-neutral. Moreover, this mechanism is given by $P(s, m) = h^{-1}(h(s) + u'_m)$.*
2. *There exists a first-best optimal quota mechanism if and only if the authority is extremely risk-averse. Moreover, this mechanism is given by $Q_m = x_m^{tar}$ and $D(R) = |\mathcal{M}| + 1$.*

Theorem 2 provides precise conditions on preferences such that the inability of priorities and quotas to adapt to the state is not problematic. That risk-neutrality and high risk aversion are sufficient for the optimality of priority and quota mechanisms is intuitive. On the one hand, if the authority is risk-neutral over the measure of agents from different groups, then they can perfectly balance their score and diversity goals without regard for the state of the world. This is because, under risk-neutrality, there is a constant "exchange rate" between the two: how the authority compares any two agents does not depend on the final allocation

[17] Note that failure of non-triviality means there exists m and n such that $h(1) + u'_n(0) \leq +h(0) + u'_m(q)$, *i.e.,* a group n agent with the maximum score is less preferred than a group m agent with the minimum score even when all of the entire capacity is allocated to group m agents.

[18] More formally, this means that there exists $(\tilde{u}, \tilde{h})$ such that the optimal allocation under (u, h) is also optimal under $(\tilde{u}, \tilde{h})$ for all $\omega \in \Omega$.

and thus can be specified *ex ante* by a priority policy. On the other hand, if the authority is extremely risk-averse as to the prospect of failing to assign x_m^{tar} agents from group m, then a quota allows them to always achieve this target level of allocation in all states of the world while minimally sacrificing scores. It is less obvious that risk-neutrality and high risk aversion are necessary. We prove this result by constructing certain adversarial measures of agents that render any priority or quota mechanism suboptimal unless the authority is risk-neutral or extremely risk-averse, respectively. Importantly, this result also shows that the only optimal quota mechanisms are those that process open slots last.

This result highlights the fragility of priority mechanisms to uncertainty absent the strong assumption of risk-neutrality over diversity. Intuitively, this is because they feature no guarantees as to how many agents of different groups will be assigned. Indeed, the unfortunate interaction between priority mechanisms and unforeseen market realizations has led to public backlash against priority mechanisms. For example, in the Vietnamese university admissions system, which combines exam scores with priority boosts for students from disadvantaged groups, a year of unexpectedly easy exams led to "top-scoring students missing out on the opportunity to attend their university of choice" and generated backlash against the system (Tuoi Tre News, 2017). Moreover, in the Boston Public Schools system, a priority policy that is set to award students bonus points for high school admissions based on the level of disadvantage of their middle school has made it impossible for students from certain middle schools to get into certain high schools, no matter their grades. This led the Boston Herald (2022) to write that "in Boston, hard work and good grades will only set your child back." As APM dynamically adjust priorities, they have the potential to remedy these practical deficiencies of priority mechanisms.

Moreover, our result highlights that quota mechanisms similarly fail to achieve the first-best away from high levels of risk aversion as they do not take advantage of the potential for positive selection. Our quantitative analysis in Section 5 in the context of quota mechanisms in Chicago Public Schools suggests that the variation in the distribution of characteristics across years generates meaningful welfare gains from switching to APM.

To formalize the connection between uncertainty and the importance of the adaptability of APM, we consider a setting with *no uncertainty*, where Λ is a Dirac measure. In this context, we say that a mechanism is optimal without uncertainty if it is a utility maximizer.

Proposition 3. *If there is no uncertainty, then there exist optimal priority and quota mechanisms.*

This result shows that if an authority is certain about the market, then appropriately constructed priority and quota mechanisms would be optimal. This formalizes the idea that the suboptimality of priority and quota mechanisms stems from their inability to adapt to the state. Of course, in practice, an authority is always somewhat uncertain of the market they face. Thus, absent the strong conditions on authority preferences that we have characterized in Theorem 2, APM dominate priority and quota mechanisms.

4 Equilibrium Mechanisms with Multiple Authorities

The single-authority model is relevant for many resource allocation contexts, such as the medical resource allocation problem of a hospital. However, in other settings such as school or university admissions, multiple authorities must decide upon their admissions policies and rules. In this section, we generalize our single authority model to a setting with multiple authorities. We define *stability* in this setting and show that there is a unique stable allocation. Moreover, we show that a mechanism is consistent with stability if and only if it coincides with the single-authority-optimal APM. We then consider a model where agents sequentially apply to the authorities, who decide which agents to admit. We show that the optimal APM is a dominant strategy. Moreover, we show that in any equilibrium in which authorities use the optimal APM, the resulting allocation corresponds to the unique stable matching of the economy. Taken together, our results provide cooperative (stability) and non-cooperative (dominance) foundations for recommending the use of APM in multi-authority settings.

4.1 The Multi-Authority Model

There are authorities $c \in \mathcal{C} = c_0 \cup \bar{\mathcal{C}} = \{c_0, c_1, \ldots, c_{|\mathcal{C}|-1}\}$ with capacities q_c, where c_0 is a dummy authority that corresponds to an agent going unmatched. The agents differ in their authority-specific scores, the group to which they belong, and their preferences over the authorities, $\succ$. We index agents by their type $\theta = (s, m, \succ) \in [0,1]^{|\mathcal{C}|} \times \mathcal{M} \times \mathcal{R} = \Theta$, where $\mathcal{R}$ is set of all complete, transitive, and strict preference relations over $\mathcal{C}$ such that c_0 is less preferred than all $c \in \bar{\mathcal{C}}$. For each type θ, $s_c(\theta)$ denotes the score of θ at authority c and $m(\theta)$ denotes the group of θ. From now, to economize on notation, we suppress indexing states by $\omega \in \Omega$ and let the measure of types be F, with density f.[19] We assume that f has full support over Θ (*i.e.,* $f > 0$) and that $F(\Theta)$ is less than the capacity of c_0 and greater than the capacity of $\bar{\mathcal{C}}$.

Each authority has preferences over the agents they are assigned of the form introduced

[19] Formally, this density is given by $f(s, m, \succ) = \frac{\partial}{\partial s} F(s, m, \succ)$.

in the previous section:

$$\xi_c(\bar{s}_{h_c}, x_c) = g_c\left(\bar{s}_{h_c} + \sum_{m\in\mathcal{M}} u_{m,c}(x_{m,c})\right) \tag{17}$$

where the extent to which they care about risk g_c, scores h_c, and diversity $\{u_{m,c}\}_{m\in\mathcal{M}}$ are potentially specific to each authority.

A matching is a function $\mu : \mathcal{C} \cup \Theta \to 2^\Theta \cup \mathcal{C}$ where $\mu(\theta) \in \mathcal{C}$ is the authority any type θ is assigned and $\mu(c) \subseteq \Theta$ is the set of agents assigned to authority c.[20] Given a matching μ, $\bar{s}_{h_c,c}(\mu)$ and $x_c(\mu) = \{x_{m,c}(\mu)\}_{m\in\mathcal{M}}$ denote the score indices and measures of agents from different groups matched to c at μ. We say that c *prefers* μ to μ', which we denote by $\mu \succ_c \mu'$, if $\xi_c(\bar{s}_{h_c,c}(\mu), x_c(\mu)) > \xi_c(\bar{s}_{h_c,c}(\mu'), x_c(\mu'))$.

Definition 5. *A matching μ is a cutoff matching if there exist cutoffs $S = \{S_{m,c}\}_{m\in\mathcal{M},c\in\mathcal{C}}$ such that $\mu(\theta) = c$ if (i) $s_c(\theta) \geq S_{m(\theta),c}$ and (ii) for all c' with $c' \succ_\theta c$, $s_{c'}(\theta) < S_{m(\theta),c'}$.*

Given S, the *demand* of an agent θ is their favorite authority among those for which they clear the cutoff:

$$D^\theta(S) = \{c : s_c(\theta) \geq S_{m(\theta),c} \text{ and } c \succeq_\theta c' \text{ for all } c' \text{ with } s_{c'}(\theta) \geq S_{m(\theta),c')}\} \tag{18}$$

The *aggregate demand* for authority c is the set of agents who demand it $D_c(S) = \{\theta : D^\theta(S) = c\}$, while $\tilde{D}_c(S_{-c}) = D_c((0, \ldots, 0), S_{-c})$ returns the set of all agents who would demand c if offered admission admission when other authorities' cutoffs are S_{-c}.

4.2 Characterization of Stable Allocations

We first characterize the set of stable allocations. Our context presents two challenges in this regard. First, the priorities which are typically used to define stability are not primitives of our model. Therefore, to define stability, we will use the preferences of the authorities induced by Equation 17. Second, unlike discrete models, a single agent does affect the preferences of an authority. Therefore, we need to consider a positive mass of agents to define blocking.

For each matching μ, authority $c \neq c_0$, and two sets of agents $\tilde{\Theta}$ and $\hat{\Theta}$, we let $\hat{\mu}_{(\hat{\Theta},\tilde{\Theta},c,\mu)}$ denote the matching that maps $\hat{\Theta}$ to c and $\tilde{\Theta}$ to c_0 and otherwise coincides with μ.[21] A set of

[20] The mathematical definition of a matching for the continuum economy we study follows Azevedo and Leshno (2016) and requires that μ satisfies the following four properties: (i) for all $\theta \in \Theta$, $\mu(\theta) \in \mathcal{C}$; (ii) for all $c \in \mathcal{C}$, $\mu(c) \subseteq \Theta$ is measurable and $F(\mu(c)) \leq q_c$; (iii) $c = \mu(\theta)$ iff $\theta \in \mu(c)$; (iv) (open on the right) for any $c \in \mathcal{C}$, the set $\{\theta \in \Theta : c \succ_\theta \mu(\theta)\}$ is open.

[21] Formally,

$$\hat{\mu}_{(\hat{\Theta},\tilde{\Theta},c,\mu)}(\theta) = \begin{cases} c_0 & \text{if } \theta \in \tilde{\Theta} \\ c & \text{if } \theta \in \hat{\Theta} \\ \mu(\theta) & \text{otherwise} \end{cases} \tag{19}$$

agents $\hat{\Theta}$ *blocks* matching μ at authority c by $\tilde{\Theta}$ if (i) for all $\theta \in \hat{\Theta}$, $c \succ_\theta \mu(\theta)$, (ii) $\tilde{\Theta} \subseteq \mu(c)$, (iii) $F(\tilde{\Theta}) = F(\hat{\Theta})$, and (iv) $\hat{\mu}_{(\hat{\Theta},\tilde{\Theta},c,\mu)} \succ_c \mu$. A matching μ is *not blocked* if there does not exist such a $(\hat{\Theta}, \tilde{\Theta}, c)$. A matching μ satisfies *within-group fairness* if for all $\theta, \theta' \in \Theta$ such that $m(\theta') = m(\theta)$ and $s_{\mu(\theta)}(\theta') > s_{\mu(\theta)}(\theta)$, it holds that $\mu(\theta') \succeq_{\theta'} \mu(\theta)$.[22] A matching μ is *stable* if it satisfies within-group fairness, is not blocked, and all non-dummy authorities fill their capacity. The following result establishes that there exists a unique stable matching and that this is a cutoff matching.

Theorem 3. *There is a unique stable matching. This matching is a cutoff matching.*

This result extends existing results on the uniqueness of stable allocations in continuum economies (see Theorem 1.1 of Azevedo and Leshno, 2016) to our setting, in which the preferences of authorities are not exogenously fixed and rather depend on the composition of the admitted agents. The stable allocation is characterized by cutoffs S, where each agent θ is matched to $D^\theta(S)$.

To gain intuition for this result, first imagine that there is only one group of agents $|\mathcal{M}| = 1$, so that authorities' preferences are determined by the scores of the agents. Given a set of cutoffs S_{-c}, a cutoff t_c clears the market for c if $F(D_c(t_c, S_{-c})) = q_c$. When $|\mathcal{M}| = 1$, for a given S_{-c}, there is a unique t_c that clears the market since a smaller cutoff will exceed the capacity while a larger one will leave a positive measure of the capacity empty. Define $T = \{T_c\}_{c \in \mathcal{C}}$, where $T_c(S)$ is the function that maps each S to the market-clearing cutoff t_c under S_{-c}. The result then follows from (i) showing the fixed points of T correspond to market-clearing cutoffs of stable matchings, (ii) establishing that T is monotone, (iii) applying Tarski's fixed point theorem to show that the set of market-clearing cutoffs is a lattice, and (iv) showing that there can only be one market-clearing cutoff as, if there were two, one would strictly exceed the capacities of at least one authority.

When $|\mathcal{M}| > 1$, there is a potential continuum of cutoffs that would clear the market for authority c. A selection from this set is provided by the cutoffs induced by the optimal APM, $A^*_{m,c}(y_m, s) \equiv h_c^{-1}(h_c(s) + u'_{m,c}(y_m))$. We show that the APM cutoffs are unique among the market-clearing cutoffs in being compatible with stability. This is because, for any other t'_c, there is a set $\hat{\Theta}$ of agents (with positive measure) who have scores lower than the cutoff for their group and a set $\tilde{\Theta}$ of agents (with positive measure) who have scores higher than the cutoff for their group, but the authority is strictly better off by admitting $\hat{\Theta}$ and rejecting $\tilde{\Theta}$. We define $T_c(S)$ as the market-clearing cutoffs induced by the optimal APM, show that

[22] Within-group fairness simply requires an authority not to reject a agent if it is admitting a agent from the same group with lower score. Under our assumption that authorities prefer higher scores (h_c is strictly increasing), within-group fairness is satisfied if there is no blocking in discrete models.

the fixed points of T_c correspond to market-clearing cutoffs of stable matchings, and follow the same steps as above to demonstrate uniqueness.

This hints at a connection between the stable allocation and the allocation induced by all authorities pursuing the optimal APM, which we now make explicit. The demand set of c at μ, $D_c(\mu)$, is the set of agents who prefer c to their allocation under μ. A mechanism is *consistent with stability* if for all F with stable matching μ_F, it chooses $\mu_F(c)$ from $D_c(\mu_F)$. In other words, evaluated at the set of agents who demand an authority, this mechanism chooses the set of agents with which the authority is already matched. Moreover, we say that a mechanism ϕ is *equivalent* to ϕ' if it chooses the same agents under all full support measures. We now establish that single-authority-optimal APMs (and equivalent mechanisms) comprise the full set of mechanisms that are consistent with stability.

Theorem 4. *A mechanism ϕ is consistent with stability if and only if it is equivalent to A^*_c.*

Thus, not only is the optimal APM A^* inherent to the structure of stable allocations, but it also characterizes stability in this setting in the sense that any deviation from A^* would result in a violation of stability.

4.3 APM Are Dominant Under Decentralized Admissions

To study which mechanisms are optimal under decentralized admissions, we now consider a setting in which the agents apply sequentially to the authorities, who then decide which agents to admit. We index the stage of the game by $t \in \mathcal{T} = \{1, \ldots, |\mathcal{C}| - 1\}$. Each stage corresponds to a (non-dummy) authority $I(t)$, where $I : \mathcal{T} \to \mathcal{T}$. At each stage t, any unmatched agents choose whether to apply to authority $I(t)$. Given the set of applicants, authority $I(t)$ chooses to admit a subset of these agents, who are then matched to the authority. Given this, histories are indexed by the path of the measure of agents who have not yet matched, $h^{t-1} = (F, F_1, \ldots, F_{t-1}) \in \mathcal{H}^{t-1}$. Given each history h^{t-1} and set of applicants $\Theta_c^A \subseteq \Theta$, a strategy for an authority returns a set of agents $\Theta_c^G \subseteq \Theta$ whom they will admit such that $\Theta_c^G \subseteq \Theta_c^A$ and $F_t(\Theta_c^G) \leq q_c$ for each time at which they could move $t \in \mathcal{T}$, $a_{c,t} : \mathcal{H}^{t-1} \times \mathcal{P}(\Theta) \to \mathcal{P}(\Theta)$, where $\mathcal{P}(\Theta)$ is the power set over Θ.[23] A strategy for an agent returns a choice of whether to apply to authorities at each history and time for all agent types $\theta \in \Theta$, $\sigma_{\theta,t} : \mathcal{H}^{t-1} \to [0, 1]$.

Within this context, our notion of equilibrium is that of subgame perfect equilibrium:

[23] Formally, so that $F_t(\Theta_c^G)$ is well defined, we require that authorities' strategies be measurable in the Borel sigma algebra over Θ.

Definition 6 (Equilibrium)**.** *A strategy profile* $\Sigma = \{\{a_{c,t}\}_{c\in\bar{\mathcal{C}}}, \{\sigma_{\theta,t}\}_{\theta\in\Theta}\}_{t\in\mathcal{T}}$ *is a subgame perfect equilibrium if* $a_{c,t}$ *maximizes authority utility given* Σ *for all* $c \in \bar{\mathcal{C}}$ *and* $t \in \mathcal{T}$ *and* $\sigma_{\theta,t}$ *is maximal according to agent preferences for all* $\theta \in \Theta$ *and* $t \in \mathcal{T}$.

We moreover say that a strategy $a_{\tilde{c},t}$ for an authority $\tilde{c}$ at time t is *dominant* if it maximizes authority utility regardless of the strategies of all other authorities and agents, $\{\{a_{c,t}\}_{c\in\bar{\mathcal{C}}\setminus\{\tilde{c}\}}, \{\sigma_{\theta,t}\}_{\theta\in\Theta}\}_{t\in\mathcal{T}}$, and the order in which authorities admit agents, I. Moreover, an equilibrium Σ is in *dominant strategies* if $a_{c,t}$ is dominant for all $c \in \bar{\mathcal{C}}$ and $t \in \mathcal{T}$. We denote the unique probabilistic allocation of agents to authorities induced by Σ as $\mu_\Sigma : \Theta \to \Delta(\mathcal{C})$. A probabilistic allocation μ_Σ is deterministic if $\mu_\Sigma(\theta)$ is a Dirac measure on some authority $c \in \mathcal{C}$ for all $\theta \in \Theta$. A deterministic allocation μ_Σ corresponds to a matching μ if $\mu_\Sigma(\theta)$ is a Dirac measure on $\mu(\theta)$ for all $\theta \in \Theta$.

We now establish that the single-authority optimal APM characterizes dominance.

Theorem 5. *A mechanism implements a dominant strategy for an authority if and only if it implements essentially the same allocations as* A_c^*.

The intuition behind this result is that each authority takes as given the set of agents that will accept it. Thus, given this measure of agents, they can do no better than to employ the same APM that a single authority would, which is A_c^* by Theorem 1.

Theorem 5 provides a powerful rationale for focusing on APMs in decentralized markets at both positive and normative levels. Normatively, this result allows an analyst to advise an authority regarding how it should conduct its admissions. This is important because any policy that does not coincide with the APM we derive — such as the popular priority and quota mechanisms outside of the cases delimited by Theorem 2 — will disadvantage an authority. Positively, this result allows a sharp prediction that the equilibrium matching between agents and authorities will be the unique stable matching (as per Theorem 3):

Proposition 4. *For all equilibria* Σ^* *where authorities use* A^*, *the allocation* μ_{Σ^*} *is deterministic and corresponds to the unique stable matching of this economy.*

The intuition is that if an equilibrium matching under A^* was not the unique stable matching, then it must be that some agents are applying suboptimally and failing to select the most preferred authority that they can attend, which contradicts that the outcome is consistent with equilibrium. Proposition 4 also shows that the allocation implemented in the dominant strategy equilibrium is the same for all possible orderings of authorities, I.

4.4 Inefficiency of APM and an Efficient Mechanism

In Appendix D, we study the efficiency properties (from the perspective of authorities) of APM. We find that decentralized outcomes when all authorities use APM are generally inefficient (Proposition 5). This is because authorities do not internalize the "pecuniary externalities" they generate by over-admitting agents that have a preference for them but that the other authorities value more. To remedy this, we propose a centralized allocation mechanism, an *Adaptive Priority Mechanism with Quotas* (APM-Q). An APM-Q first constructs a fictitious aggregate authority, decides the aggregate levels of admissions of each group according to an optimal APM, and then allocates these groups across authorities according to optimally set quotas. We show that an optimally set APM-Q yields efficient outcomes (Theorem 6). Intuitively, this mechanism fixes the pecuniary externalities by creating a pseudo-market in which each group has a "price" and authorities are assigned agents only if they would be willing to "pay" for them.

5 Benchmarking the Quantitative Gains from APM

We have so far shown theoretically that APM outperform conventional priority and quota mechanisms. In this section, we attempt to benchmark the magnitude of the gains from implementing APM relative to the reserve system employed by Chicago Public Schools (CPS). To do this, we use application and admission data from CPS for the 2013-2017 academic years. Estimating preference parameters to best rationalize the pursued reserve policy, we find that the gains from using the optimal APM are equivalent to removing 37.5% of the loss to CPS' payoffs from failing to admit a diverse class of students. Our analysis therefore suggests that the gains from APM are considerable.

5.1 Institutional Detail on Chicago Public Schools

We first describe the institutional context of CPS. Under current policy, CPS admits students to its selective exam schools based on two criteria. First, CPS ranks students according to a score which combines the results of a specialized entrance exam, prior standardized test scores, and grades in prior coursework. This composite score ranges from 0 to 900 and higher-scoring students are admitted before lower-scoring ones, reflecting CPS's desire to allocate seats in exam schools to the students with the best academic standing. In our model, these are the students' scores. Second, CPS divides the census tracts in the city into four *tiers* based on socioeconomic characteristics.[24] Tier 1 tracts are the most disadvantaged, while

[24] Concretely, 800 census tracts are divided into four tiers based on six characteristics of each census tract: (i) median family income, (ii) percentage of single-parent households, (iii) percentage of households where English is not the first language, (iv) percentage of homes occupied by the homeowner, (v) adult educational attainment, and (vi) average Illinois Standards Achievement Test scores for attendance-area schools. These characteristics are then combined to construct the socioeconomic score for the tract. Finally, the tracts are ranked according to socioeconomic scores and partitioned into 4 tiers with approximately the same number

Table 1: Admissions Cutoff Scores for Payton

Cutoff Score	2013	2014	2015	2016	2017
Tier 1	801	838	784	769	771
Tier 2	845	840	831	826	846
Tier 3	871	883	877	853	875
Tier 4	892	896	891	890	894

Note: The table reports the score of the lowest-scoring student that was admitted to Payton in each of the four tiers and five years.

Tier 4 tracts are the most advantaged. This is reflected in the composite scores of students from Tier 1, who represent 25% of the city's population but comprise relatively few of the high-scoring students (Ellison and Pathak, 2021). As a result, Tier 1 students would have a very small share in the city's top exam schools without affirmative action. To ensure more equal representation across socioeconomic status in these schools, between 2013-2017 CPS implemented a quota policy that reserves 17.5% of the seats for each tier, yielding a total of 70% reserve seats and 30% merit slots that are open to students from all tiers. CPS allocates the seats by first assigning the highest-scoring students (regardless of their tier) to the merit slots and then the highest-scoring students from each tier to the 17.5% reserve seats.

We focus on the most selective CPS school, Walter Payton College Preparatory High School (Payton), which has the highest cutoffs for each tier in each year in our data and would have very few tier 1 students without affirmative action.[25] Table 1 presents the cutoff scores of each tier (*i.e.,* the composite score of the last admitted student from each tier).

We make two observations. First, the cutoff students from less advantageous tiers face is lower than the cutoff for students from more advantageous tiers. Therefore, CPS has a *revealed* preference (and not merely a stated preference) for a diverse student body. Second, cutoff scores vary across years. This implies that the distribution of applicant characteristics varies from year to year. Given this uncertainty and the fact that CPS uses a policy that processes quotas after open slots, we know by Theorem 2 that CPS' baseline policy cannot be rationalized as optimal (even if they are extremely risk-averse). Nevertheless, it is always possible that the gains from APM could be small.

of school-age children. See Ellison and Pathak (2021) for a more detailed account of the CPS system.

[25] This approach follows the analysis in Ellison and Pathak (2021), who focus on the two most competitive schools, Northside College Preparatory High School (Northside) alongside Payton. In the years we study, the cutoff scores for Northside are below some other schools frequently, which is why we restrict attention to Payton.

5.2 Preferences and Estimation Methodology

We perform our analysis in two steps: establishing a parametric framework for evaluating welfare gains and losses and then estimating its parameters.

Preferences. We assume a parametric form for CPS's preferences to evaluate the gains from APM. In particular, we impose that the preferences of CPS over the scores and diversity of the student body are represented by the following parametric utility function:

$$\xi(\bar{s}, x; \beta, \gamma) = \bar{s} + \sum_{t=1}^{4} \beta |x_t - 0.25|^{\gamma} \tag{20}$$

where $\bar{s}$ is the average score of admitted students, x_t is the percentage of tier t students. Motivated by CPS' desire to allocate the highest-scoring students, ξ is increasing in $\bar{s}$. To model the diversity preferences of CPS, we assume that CPS loses as the gap from equal representation in each tier increases. We do this through the functional form $\beta|x_t - 0.25|^{\gamma}$. The parameters β and γ index the slope and curvature of utility in losses from unequal representation and are the two free parameters of our framework. To probe robustness to this functional form assumption, in Appendix E.2 we consider two other parametric utility functions that estimate separate coefficients for underrepresented and overrepresented tiers and only consider losses from underrepresented tiers.

Estimation. We estimate β and γ to best rationalize the choice of 17.5% reserves for each tier as optimal. We believe this to be a reasonable approach, as the size of the reserves is an important issue that is decided only after much deliberation.[26] Moreover, CPS has used the size of the reserves as a policy tool, increasing them from 15% to 17.5% in 2012 and is currently deliberating another change that would further boost the representation of tier 1 and tier 2 students (Chicago Public Schools, 2022).

Given our functional form, the optimality of the chosen reserve sizes yields moment conditions that we use to estimate the parameters β and γ. Formally, we index reserve mechanisms by the reserve sizes of the four socioeconomic tiers $r = (r_1, r_2, r_3, r_4)$. We let $\bar{s}(r, y)$ and $x(r, y)$ denote the average scores and tier percentages that would be obtained in year y, with distribution F_y, under reserve policy r. The payoff of the policymaker under reserve policy r is given by $\Xi(r, \Lambda; \beta, \gamma)$, as per Equation 7:

$$\Xi(r, \Lambda; \beta, \gamma) = \mathbb{E}_{\Lambda}[\xi(\bar{s}(r, y), x(r, y); \beta, \gamma)] \tag{21}$$

[26] These points are emphasized in Dur, Pathak, and Sönmez (2020): "This change was made at the urging of a Blue Ribbon Commission (BRC, 2011), which examined the racial makeup of schools under the 60% reservation compared to the old Chicago's old system of racial quotas. They advocated for the increase in tier reservations on the basis it would be "improving the chances for students in neighborhoods with low performing schools, increasing diversity, and complementing the other variables."

where the expectation is taken over distributions of agents' characteristics F_y under the subjective probability measure Λ. Define the expected marginal benefit of increasing reserve i and decreasing reserve j as:

$$G_{ij}(r,\Lambda;\beta,\gamma)=\frac{\partial}{\partial r_i}\Xi(r,\Lambda;\beta,\gamma)-\frac{\partial}{\partial r_j}\Xi(r,\Lambda;\beta,\gamma) \tag{22}$$

Any (interior) optimal reserve policy r^* must equate the expected marginal benefit of increasing reserve i and decreasing reserve j at r^* to zero for all (i,j) pairs, *i.e.*, $G_{ij}(r^*,\Lambda;\beta,\gamma)=0$ for all $\{i,j\}\subset\{1,2,3,4\}$ such that $j>i$. These six first-order conditions yield six moments.

We take empirical analogs of the theoretical moments and estimate preference parameters by minimizing the sum of squared deviations of these moments from zero. We take CPS' pursued reserve policy as optimal, $\hat{r}^*=(0.175,0.175,0.175,0.175)$. We estimate the empirical joint distribution of students' scores and tiers in CPS in each year $\hat{F}_y$ for $y\in\{2013,2014,2015,2016,2017\}$ and estimate $\hat{\Lambda}$ as a distribution that places equal probability on each of these five measured distributions. We plug these sample estimates into the theoretical moment functions. This yields six empirical moment functions that depend only on the preference parameters, $G_{ij}(\hat{r}^*,\hat{\Lambda};\beta,\gamma)$. Motivated by the theoretical necessity of $G_{ij}(r^*,\Lambda;\beta,\gamma)=0$, we estimate the preference parameters by minimizing the sum of squared deviations of the empirical moments from zero:

$$(\beta^*,\gamma^*)\in\arg\min_{\beta,\gamma}\sum_{i=1}^{4}\sum_{j>i}G_{ij}(\hat{r}^*,\hat{\Lambda};\beta,\gamma)^2 \tag{23}$$

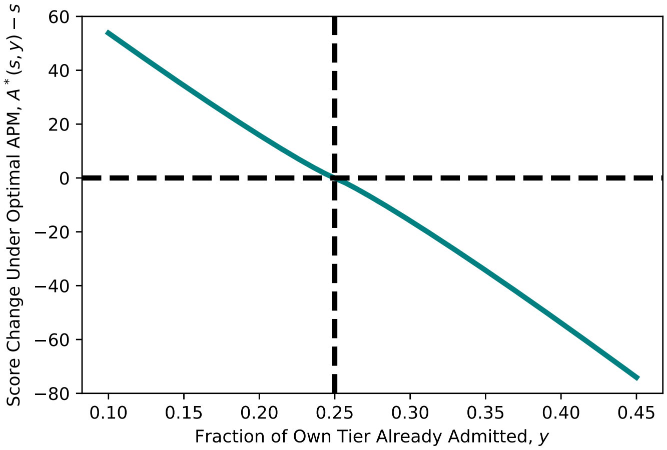
Performing this estimation yields estimated parameter values of $\beta^*=-209.5$ and $\gamma^*=2.11$.

5.3 The Estimated Gains from APM

We now use our estimated model to quantify the welfare gains from using APM. To do this, we compare the empirical payoff $\Xi(\phi,\hat{\Lambda},\beta^*,\gamma^*)$ under two mechanisms: the pursued quota policy, r^*, and the optimal APM from Theorem 1, A^*. In Figure 4, we illustrate how the estimated optimal APM changes students' scores to arrive at their ultimate priorities. In accordance with the preferences we have assumed, students receive a score boost when their tier is underrepresented and a score penalty when their tier is overrepresented. As we found $\gamma^*=2.11$, the estimated diversity preference is very close to quadratic. Thus, the optimal APM is very close to linear. From Theorems 1 and 2, we know that this APM achieves the first-best allocation in each year while the implemented quota policy does not. However, our theorems do not guarantee that the gains from APM are economically meaningful.

The empirical payoff under APM is 876.9, while it is 874.8 under the CPS reserve policy. Thus, the gains from APM are equivalent to increasing average scores by 2.1, holding diversity

Figure 4: The Estimated Optimal APM



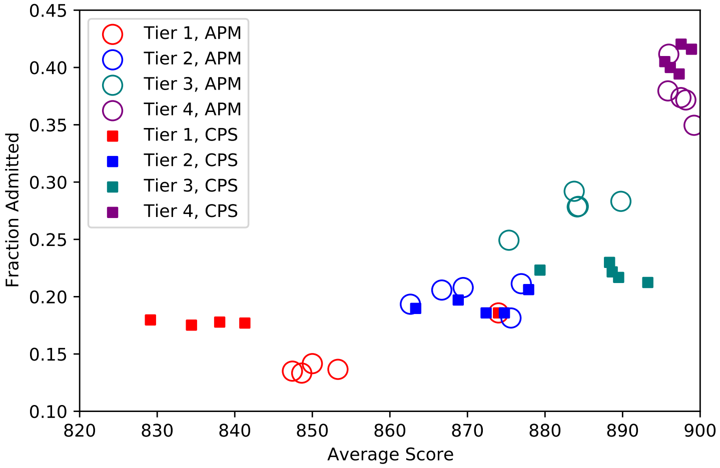
Note: This figure plots the change in a student's score when fraction y of students in their own tier has already been admitted under the estimated optimal APM, A^*. At $y = 0.25$ (the vertical dashed black line), the score in unchanged. For $y < 0.25$, students receive a score boost. For $y > 0.25$, students receive a score penalty. The range of the x-axis, $[0.1, 0.45]$, is chosen to cover the full range of fractions of admitted students under both the optimal and the CPS reserve policy from all tiers in all years of our sample (see Figure 5).

fixed. To benchmark the size of the gains, we require units in which they can be meaningfully expressed. To this end, we define the *loss from underrepresentation* as the payoff lost by CPS under its baseline policy from not admitting a fully balanced class, while holding fixed the average score of the class. This is equal to 5.6 points under our estimated parameters. Thus, the gains from APM are equal to 37.5% of the loss from underrepresentation incurred under the CPS policy.[27] To contextualize the magnitude of this improvement, we can compare the gains from switching to the optimal APM to the gains from the 2012 (the year before our sample) CPS reform, which increased the size of all reserves from 15% to 17.5%. Under the estimated preferences, the empirical payoff under the 15% reserve rule is 873.9, and so the gains from the reform are equivalent to increasing average scores by 0.9. Thus, the gains from switching to the optimal APM are 2.3 times larger than the gains from this recent reform.

These estimates suggest that the gains from APM are economically meaningful. These gains stem from the variation across years in the joint distribution of student scores and tiers. This can be seen in Table 1, which shows the variability in the scores of the marginally admitted students from tiers 1, 2 and 3. More systematically, we visualize the difference in outcomes under CPS' reserves and the optimal APM by plotting the average scores and

[27] This is equivalent to increasing the percentage of students from tier 1 from 0.179 to 0.21 and decreasing the percentage of students from tier 4 from 0.407 to 0.378. This corresponds to swapping 8.7 students from the most overrepresented group (tier 4) for the most underrepresented group (tier 1) each year.

Figure 5: Comparing Admissions under the Optimal APM and the CPS Policy



Note: Each point corresponds to one of the four tiers of students in one of the five years under either the optimal APM or the CPS policy. The x-axis corresponds to the average score of those admitted from that tier in that year under that policy. The y-axis corresponds to the fraction of admitted students from that tier in that tier under that policy.

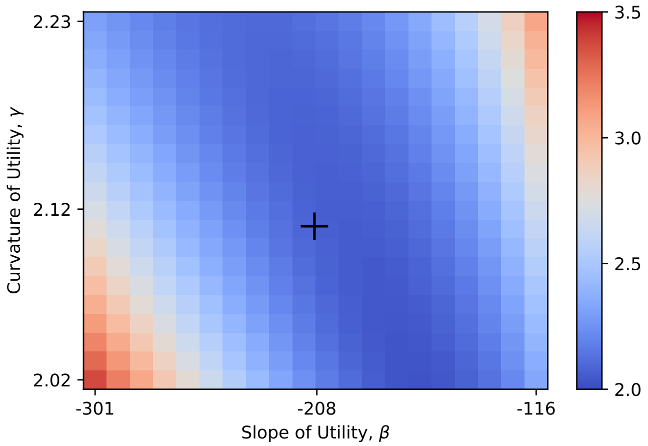
fraction admitted for each tier for each year under both mechanisms in Figure 5. There are two main differences between the allocations. First, the APM allocates systematically fewer tier 1 and tier 4 students and more tier 3 students. Second, the APM admits a greater fraction of students from each tier (especially tiers 1 and 3) in the years in which that tier scores well. This positive selection generates the welfare gains.

Robustness. We now explore the robustness of APM to the three core assumptions of our analysis: (i) that CPS has the correct beliefs about the distribution of distributions of students, (ii) that CPS has preferences that lie in the assumed parametric family, and (iii) that CPS separately optimizes the sizes of all four tiers.

Our baseline analysis took the beliefs of CPS to be the true empirical distribution of student distributions over years. To test robustness to this assumption, we take $\hat{\Lambda}$ as a Dirac distribution on the realized distribution for each of the five years of our data and re-estimate the preference parameters. In Figure 6, we plot the difference in welfare under the optimal APM and CPS reserve policies over the full range of these re-estimated parameters (*i.e.,* we take the minimum and maximum of the estimated parameters across years as the ranges for the axes). We find that the gains from APM range from 2.0 to 3.5, while our baseline estimate was 2.1. Thus, the point estimate of our welfare gains from APM appears to be conservative by this metric.

To gauge robustness to the functional form we have assumed for CPS' preferences, in Appendix E.2, we estimate two different parametric specifications of utility. First, we con-

Figure 6: Robustness of the Gains from APM



Note: This chart plots the difference in empirical payoffs from the optimal APM and CPS reserve policy under alternative parameter values, with the shaded colors corresponding to the numerical value of the gains from APM, ranging from 2.0 to 3.5. The black '+' indicates our baseline parameter values. The ranges for the axes are obtained by estimating β and γ separately for each year of our data and separately taking the minimum and maximum estimated values of each set of estimated parameters.

sider a utility function that includes a loss term only for underrepresented tiers (and does not penalize overrepresentation of any tier). Second, we allow for CPS to care differentially about underrepresentation and overrepresentation by considering a utility function with separate coefficients for underrepresented and overrepresented tiers. We find that under these specifications, the improvement from APM corresponds to 9.7% and 8.7% of the loss from underrepresentation, which is attenuated relative to our baseline, but remains considerable.

To study the robustness of our findings to the assumption that CPS separately optimizes the size of all four tiers, in Appendix E.1 we consider a setting where CPS sets a *single* reserve size for all tiers. As we now have only one moment condition, we vary γ over the interval [1,10], estimate $\beta^*(\gamma)$ as the exact solution to the moment condition, and compute the gains from APM as a function of γ. The *minimum* gain from APM over the estimated range is 1.98 points, which corresponds to 26.2% of the loss from underrepresentation under that parameterization. This is slightly smaller than our baseline estimate but still considerable.

Limitations. Finally, we state some limitations of our analysis. First, even though we argue that our functional form assumptions are reasonable and parsimonious in this setting, there are possibly many other parametric utility functions that might represent the preferences of CPS. Unfortunately, with the available data, richer methods of preference estimation that allow for higher dimensionality of preference parameters are severely limited. This is because admissions rules are only set once and we can therefore use only six first-order conditions. Richer data in which we observe choices of students from different applicant sets (either in practice or in hypothetical choice settings) would allow for a more detailed analysis. Second, one of the main aims of the tier system employed by CPS is to increase racial diversity in the prestigious exam schools. Indeed, the pursued tier system is a race-neutral alternative that replaced the previous race-based system following two Supreme Court Rulings in 2003 and 2007 (see Ellison and Pathak, 2021, for a summary). Because of this, CPS uses tiers based on socioeconomic status instead of race and so we estimate their preferences over tiers. Of course, if admission rules could depend on race, then one could perform a similar analysis in which APM simply prioritize based on race rather than tier.

6 Conclusion

Motivated by the use of priority and quota policies in resource allocation settings with diversity concerns, we consider an authority that has separable preferences over scores and diversity. We introduce Adaptive Priority Mechanisms (APM) and characterize an APM that is both optimal and can be specified solely in terms of the preferences of the authority. We study the priority and quota policies that are used in practice and show that they are optimal if and only if the authority is either risk-neutral or extremely risk-averse over diversity. Analyzing a setting with multiple authorities that dynamically admit agents, we show that the optimal APM is a dominant strategy. Thus, one could potentially advise authorities to follow an optimal APM with confidence (under our maintained assumptions on preferences) that they could do no better. Moreover, all authorities following the optimal APM implements the unique stable allocation.

Our analysis has potential implications for improving the design of real-world allocation mechanisms. First, we show that while both priorities and quotas can be better than one another (depending on the risk preferences of the authority), they are generally suboptimal. Second, we show how to improve upon these mechanisms using APM that harness the strengths of these policies: APM benefit both from the guarantee effect of quotas in ensuring certain levels of admissions from various groups and the positive selection effect of priorities in expanding affirmative action when it is least costly. Our quantitative analysis using CPS data suggests that the use of APM can yield considerable welfare gains over the *status quo*. We conclude that APM may have practical utility in delivering more desirable allocations of resources. Moreover, APM could be applied in many settings, including the allocation of seats at schools, places at universities, and medical resources to patients.

Appendices

A Omitted Proofs of Theorems

A.1 Proof of Theorem 1

Proof. We characterize the optimal allocation for each $\omega \in \Omega$ and show that the claimed adaptive priority mechanism implements the same allocation. Fix an $\omega \in \Omega$ and suppress the dependence of F_ω and f_ω thereon, and define the utility index of a score as $\tilde{s} = h(s)$ with induced densities over $\tilde{s}$ given by $\tilde{f}_m$ for all $m \in \mathcal{M}$. Let the measure of agents from any group $m \in \mathcal{M}$ that is allocated the resource be $x_m \in [0, \bar{x}_m]$ where $\bar{x}_m = \int_{h(0)}^{h(1)} \tilde{f}_m(\tilde{s})\mathrm{d}\tilde{s}$. Observe that, conditional on fixing the measures of agents from each group that are allocated the resource $x = \{x_m\}_{m\in\mathcal{M}}$, there is a unique optimal allocation (*i.e.,* ξ-maximal μ up to measure zero transformations). In particular, as g and h are continuous and strictly increasing, the optimal allocation conditional on x satisfies $\mu^*(\tilde{s}, m; x) = 1 \iff \tilde{s} \geq \underline{\tilde{s}}_m(x_m)$ for some thresholds $\{\underline{\tilde{s}}_m(x_m)\}_{m\in\mathcal{M}}$ that solve:

$$\int_{\underline{\tilde{s}}_m(x_m)}^{h(1)} \tilde{f}_m(\tilde{s})\mathrm{d}\tilde{s} = x_m \tag{24}$$

We can then express the problem of choosing the optimal $x = \{x_m\}_{m\in\mathcal{M}}$ as:

$$\max_{x_m\in[0,\bar{x}_m],\, \forall m\in\mathcal{M}} \sum_{m\in\mathcal{M}} \int_{\underline{\tilde{s}}_m(x_m)}^{h(1)} \tilde{s}\tilde{f}_m(\tilde{s})\mathrm{d}\tilde{s} + \sum_{m\in\mathcal{M}} u_m(x_m) \quad \text{s.t.} \sum_{m\in\mathcal{M}} x_m \leq q \tag{25}$$

where a solution exists by compactness of the constraint sets and continuity of the objective. We can derive necessary and sufficient conditions on the solution(s) to this problem by considering the Lagrangian:

$$\begin{aligned}\mathcal{L}(x, \lambda, \overline{\kappa}, \underline{\kappa}) = &\sum_{m\in\mathcal{M}} \int_{\underline{\tilde{s}}_m(x_m)}^{h(1)} \tilde{s}\tilde{f}_m(\tilde{s})\mathrm{d}\tilde{s} + \sum_{m\in\mathcal{M}} u_m(x_m) \\ &+ \lambda\left(q - \sum_{m\in\mathcal{M}} x_m\right) + \sum_{m\in\mathcal{M}} \overline{\kappa}_m(\bar{x}_m - x_m) + \sum_{m\in\mathcal{M}} \underline{\kappa}_m x_m\end{aligned} \tag{26}$$

The first-order necessary conditions to this program are given by:

$$\frac{\partial \mathcal{L}}{\partial x_m} = -\underline{\tilde{s}}_m'(x_m)\underline{\tilde{s}}_m(x_m)\tilde{f}_m(\underline{\tilde{s}}_m(x_m)) + u_m'(x_m) - \lambda - \overline{\kappa}_m + \underline{\kappa}_m = 0 \tag{27}$$

$$\lambda\frac{\partial \mathcal{L}}{\partial \lambda} = \lambda\left(q - \sum_{m\in\mathcal{M}} x_m\right) = 0 \tag{28}$$

$$\overline{\kappa}_m\frac{\partial \mathcal{L}}{\partial \overline{\kappa}_m} = \overline{\kappa}_m(\bar{x}_m - x_m) = 0 \tag{29}$$

$$\underline{\kappa}_m \frac{\partial \mathcal{L}}{\partial \underline{\kappa}_m} = \underline{\kappa}_m x_m = 0 \tag{30}$$

for all $m \in \mathcal{M}$. By implicitly differentiating Equation 24, we obtain that:

$$-\tilde{\underline{s}}'_m(x_m)\tilde{f}_m(\tilde{\underline{s}}_m(x_m)) = 1 \tag{31}$$

Thus, we can simplify Equation 27 to:

$$\frac{\partial \mathcal{L}}{\partial x_m} = \tilde{\underline{s}}_m(x_m) + u'_m(x_m) - \lambda - \overline{\kappa}_m + \underline{\kappa}_m = 0 \tag{32}$$

Observe that all constraints are linear. Thus, if the objective function is strictly concave, the first-order conditions are also sufficient. Observe by Equation 31 that $\tilde{\underline{s}}_m(x_m)$ is a strictly decreasing function of x_m, and all cross-partial derivatives are zero. Therefore, the first summation is strictly concave. Moreover u'_m is a decreasing function of x_m by virtue of the assumption that u_m is concave for all $m \in \mathcal{M}$. Then the second summation is concave. Thus, the objective function is strictly concave and the optimal allocation is unique.

Thus, to verify that our claimed adaptive priority mechanism is a first-best mechanism, it suffices to show that the allocation it implements satisfies Equations 27 to 30. The adaptive priority mechanism $A_m(y_m, s) = h^{-1}\left(h(s) + u'_m(y_m)\right)$ in the transformed score space yields transformed scores $h\left(A_m(y_m, s)\right) = \tilde{s} + u'_m(y_m)$. Define x_m as the admitted measure of agents from group m under this mechanism. Agents in group $m \in \mathcal{M}$ are allocated the resource if and only if $\tilde{s} + u'_m(x_m) \geq s^C$ for some threshold s^C that solves:

$$\sum_{m \in \mathcal{M}} \int_{\max\{\min\{s^C - u'_m(x_m), h(1)\}, h(0)\}}^{h(1)} \tilde{f}_m(\tilde{s}) \mathrm{d}\tilde{s} = q \tag{33}$$

We can therefore partition $\mathcal{M}$ into three sets that are uniquely defined: (i) interior $\mathcal{M}_I = \{m \in \mathcal{M} | s^C - u'_m(x_m) \in (h(0), h(1))\}$; (ii) no allocation $\mathcal{M}_0 = \{m \in \mathcal{M} | s^C - u'_m(x_m) \geq h(1)\}$; (iii) full allocation $\mathcal{M}_I = \{m \in \mathcal{M} | s^C - u'_m(x_m) \leq h(0)\}$. For all $m \in \mathcal{M}_0$, we implement $x_m = 0$. For all $m \in \mathcal{M}_I$, we implement $x_m = \bar{x}_m$. For all $m \in \mathcal{M}_I$, we implement $x_m \in (0, \bar{x}_m)$. For any $m \in \mathcal{M}_I$, the allocation threshold is $\tilde{\underline{s}}_m(x_m) = s^C - u'_m(x_m)$. For any $m \in \mathcal{M}_0$, the allocation threshold is $h(1)$. For any $m \in \mathcal{M}_I$, the allocation threshold is $h(0)$.

We now verify that this outcome satisfies the established necessary and sufficient conditions. For all $m \in \mathcal{M}_I$, by the complementary slackness conditions we have that $\underline{\kappa}_m = \overline{\kappa}_m = 0$. Substituting the above into Equation 27 for all $m \in \mathcal{M}_I$ we obtain that:

$$s^C - \lambda = 0 \tag{34}$$

which is satisfied for $\lambda = s^C$. As $q = \sum_{m \in \mathcal{M}} x_m$, the complementary slackness condition for λ is then satisfied. For all $m \in \mathcal{M}_0$, by complementary slackness we have that $\overline{\kappa}_m = 0$ and

Equation 27 is satisfied by:

$$\underline{\kappa}_m = \lambda - h(1) - u'_m(0) \tag{35}$$

For all $m \in \mathcal{M}_I$, by complementary slackness we have that $\underline{\kappa}_m = 0$ and Equation 27 is satisfied by:

$$\overline{\kappa}_m = h(0) + u'_m(\bar{x}_m) - \lambda \tag{36}$$

This completes the proof of first-best optimality of A^*. Moreover, as the optimal allocation is unique for all ω, any allocation that differs from the allocation implemented by the optimal APM at any ω would not be first-best optimal. Therefore, any first-best-optimal mechanism must implement essentially the same allocation as A^*. □

A.2 Proof of Theorem 2

Proof. First, we prove the if parts of the results. Part (i): When u_m is linear, u'_m is constant and the first-best optimal adaptive priority mechanism is a priority mechanism $P(s,m) = h^{-1}(h(s)+u'_m)$. Part (ii): When $\tilde{u}'_m(x_m) \geq k$ for $x_m \leq x^{\text{tar}}_m$ and $\tilde{u}'_m(x_m) = 0$ for $x_m > x^{\text{tar}}_m$ and $\sum_{m\in\mathcal{M}} x^{\text{tar}}_m < q$, observe that the optimal mechanism admits $x_m \geq x^{\text{tar}}_m$ for all $m \in \mathcal{M}$ in all states of the world, but conditional on $x_m \geq x^{\text{tar}}_m$ for all $m \in \mathcal{M}$ admits the highest-scoring set of agents. A quota $Q_m = x^{\text{tar}}_m$ and $Q_R = q - \sum_{m\in\mathcal{M}} x^{\text{tar}}_m$, with $D(R) = |\mathcal{M}|+1$ implements this allocation and is first-best optimal for any authority that is extremely risk-averse.

Second, we prove the only if parts of the results. Part (i): Assume the utility functions are not linear and let m denote a group where u'_m is not constant in $[0,q]$. We say that a state ω has full support if f_w has full support. A state ω has *full support in m and n* if $f_w(\cdot,m) > 0$ and $f_w(\cdot,n) > 0$ for some m and n and positive measures of only m and n. Let ω be a state that has full support in m and n. Moreover, assume both groups have a measure q of agents. We first establish that in any optimal allocation, agents from both groups are allocated the resource.

Claim 1. *If preferences are non-trivial, then the optimal allocation has $x_n, x_m > 0$.*

Proof. Toward a contradiction, suppose without loss of generality that $x_n = 0$. This implies that $x_m = q$. By the necessary first-order condition from Theorem 1 (combing Equations 27 and 30), we have that:

$$u'_m(q) + h(0) = u'_n(0) + h(1) + \underline{\kappa}_n \geq u'_n(0) + h(1) \tag{37}$$

where the inequality follows as $\underline{\kappa}_n \geq 0$. Thus, we have that:

$$u'_m(q) - u'_n(0) \geq h(1) - h(0) > u'_m(q) - u'_n(0) \tag{38}$$

where the first inequality follows by rearranging Equation 37 and the second follows by the definition of non-triviality of preferences. This is a contradiction, thus $x_n, x_m > 0$ in any optimal allocation. □

We now establish an equation relating x_n and x_m that will be useful in the steps to come.

Claim 2. *Let ω have full support in m and n, μ denote a cutoff matching with cutoffs s_m and s_n. Let x_m and x_n denote the measures of agents who are allocated the object at μ. μ is optimal if and only if $u'_m(x_m) + h(s_m) = u'_n(x_n) + h(s_n)$ and $x_n + x_m = q$.*

Proof. By the necessary and sufficient first-order conditions from Theorem 1, we again have that:

$$u'_m(x_m) + h(s_m) - \overline{\kappa}_m + \underline{\kappa}_m = u'_n(x_n) + h(s_n) - \overline{\kappa}_n + \underline{\kappa}_n \tag{39}$$

By Claim 1, we have $x_m, x_n > 0$. Thus, by the complementary slackness conditions (Equations 29 and 30), we have that $\overline{\kappa}_m = \underline{\kappa}_m = \overline{\kappa}_n = \underline{\kappa}_n = 0$. Thus, we obtain:

$$u'_m(x_m) + h(s_m) = u'_n(x_n) + h(s_n) \tag{40}$$

together with $x_n + x_m = q$, we have characterized the optimal allocation as claimed. □

Continue to let x_m and x_n denote the measures of group m and n agents at the optimal allocation under ω, and s_m and s_n denote the cutoff scores for admission. There are now two cases to consider: (i) $u'_m(x_m)$ and $u'_n(x_n)$ are locally constant. (ii) $u'_m(x_m)$ or $u'_n(x_n)$ are not locally constant. If we are in case (i), we will construct an ω' with a unique optimal allocation x'_m and x'_n where $u'_m(x'_m)$ or $u'_n(x'_n)$ is not locally constant, and then show jointly how we arrive at a contradiction in both cases (i) and (ii).

To this end, suppose that we are in case (i). Let x^*_m and x^*_n denote the measures that are closest to x_m and x_n such that $u'_m(x_m)$ and $u'_n(x_n)$ are not locally constant, *i.e.*,:

$$\begin{aligned} x^*_k = \arg\min_{x'_k} \Big\{ |x_k - x'_k| \Big| u'_k(x_k) = u'_k(x'_k) \text{ and for all } \varepsilon > 0 \\ \text{either } u'_k(x'_k - \varepsilon) > u'_k(x_k) \text{ or } u'_k(x'_k + \varepsilon) > u'_k(x_k) \Big\} \end{aligned} \tag{41}$$

As u'_k is continuous, this minimum is attained and x^*_k is well-defined. Without loss of generality, assume $|x_m - x^*_m| \leq |x_n - x^*_n|$ and define both $\hat{x}_m = x^*_m$ and $\hat{x}_n = q - x^*_m$. We now construct a state ω' such that $\hat{x}$ is optimal:

Claim 3. *Define ω' where $F_m(1) - F_m(s_m) = \hat{x}_m$, $F_n(1) - F_n(s_n) = \hat{x}_n$ and ω' has full support in m and n. The allocation that admits the highest-scoring $\hat{x}_m$ group m agents and the highest-scoring $\hat{x}_n$ group n agents is the unique optimal allocation.*

Proof. By Claim 2, as $\hat{x}_m + \hat{x}_n = q$ by construction, $\hat{x}$ is optimal if and only if Equation 40 holds. To this end, observe that if we admit $\hat{x}$, then the cutoff scores are the same as under x as $F_m(1) - F_m(s_m) = \hat{x}_m$ and $F_n(1) - F_n(s_n) = \hat{x}_n$, by construction. Thus, we have that:

$$u'_m(\hat{x}_m) + h(s_m) = u'_m(x_m) + h(s_m) = u'_n(x_n) + h(s_n) = u'_n(\hat{x}_n) + h(s_n) \tag{42}$$

where the first equality holds by construction as $\hat{x}_m = x^*_m$ and $u'_m(x^*_m) = u'_m(x_m)$, the second equality holds by optimality of x, and the third equality holds as $|x_m - x^*_m| \leq |x_n - x^*_n|$, which implies that $u'_n(\hat{x}_n) = u'_n(x^*_n)$. Thus, Equation 40 holds, and $\hat{x}$ is optimal, as claimed. □

Observe that this construction also applies trivially in case (ii) with $x^*_m = x_m$. Thus, using this construction, we can now study cases (i) and (ii) together. In state ω', to implement this optimal allocation, we must have that $P(s, m) < P(s_n, n)$ for all but a measure zero set of s such that $s < s_m$. We will now construct another state ω'' such that any priority mechanism with this property is suboptimal.

First, suppose that $x^*_m \leq x_m$ and fix some $\varepsilon \in (0, x^*_m)$. Define $\tilde{s}_m$ as solving the following equation:

$$u'_m(\hat{x}_m - \varepsilon) + h(\tilde{s}_m) = u'_n(\hat{x}_n + \varepsilon) + h(s_n) \tag{43}$$

We then have that:

$$\tilde{s}_m = h^{-1}\left(h(s_n) + u'_n(\hat{x}_n + \varepsilon) - u'_m(\hat{x}_m - \varepsilon)\right) < h^{-1}\left(h(s_n) + u'_n(\hat{x}_n) - u'_m(\hat{x}_m)\right) = s_m \tag{44}$$

where the first equality rearranges Equation 43 and the second inequality uses the facts that $u'_n(\hat{x}_n) \leq u'_n(\hat{x}_n + \varepsilon)$ and $u'_m(\hat{x}_m) < u'_m(\hat{x}_m - \varepsilon)$. We now construct a state ω'' such that $(\hat{x}_m - \varepsilon, \hat{x}_n + \varepsilon)$ is optimal.

Claim 4. *Define ω'' where $1 - F_m(\tilde{s}_m) = \hat{x}_m - \varepsilon$, $1 - F_n(s_n) = \hat{x}_n + \varepsilon$ with full support in m and n. The allocation that admits the highest-scoring $(\hat{x}_m - \varepsilon, \hat{x}_n + \varepsilon)$ agents is the unique optimal allocation.*

Proof. Following the same steps as Claim 3, and the fact that Equation 43 holds by construction, we have that the claim holds. □

Observe that to implement this optimal allocation a priority mechanism must set $P(s, m) \geq P(s_n, n)$ for all but zero measure $s > \tilde{s}_m$. However, since $\tilde{s}_m < s_m$, this contradicts the optimality condition for state ω' that $P(s, m) < P(s_n, n)$. This is because for all but measure zero $s \in (\tilde{s}_m, s_m)$, which we have established is non-empty, we have that:

$$P(s, m) \geq P(s_n, n) > P(s, m) \tag{45}$$

which is a contradiction. To complete the proof, we need only consider the case that $x^*_m > x_m$. In this case, we can apply essentially the same steps and the result follows. Concretely, instead increasing $\hat{x}_m$ by ε and following the same steps yields the required contradiction.

We have now constructed three states $\omega, \omega', \omega''$ such that no priority mechanism can be optimal in each state when the authority is not risk-neural, completing the proof.

Part (ii): Assume that a quota policy is optimal, we now show that the authority's preferences must be extremely risk-averse. For each group $m \in \mathcal{M}$, let $c_m \in [0, 1]$ and $c_m \neq c_n$ if $m \neq n$. Let ω be such that the scores of agents from each group m are uniformly distributed between $[c_m, c_m + \epsilon]$, where ϵ is chosen to be small so that there is no overlap of these supports and each group has measure q agents. Let m_ω denote the group with the highest c_m at ω. Now, compute the optimal allocation at ω and denote the measure of admitted agents from each group at the optimal allocation by $\{x^*_m(\omega)\}_{m \in \mathcal{M}}$. We first show that under any optimal quota policy, the level of the quotas must be set equal to the optimal allocation for all but the highest-scoring group:

Claim 5. *If a quota Q attains the optimal allocation, then for each $m \neq m_\omega$, $Q_m = x^*_m(\omega)$.*

Proof. If $Q_m > x^*_m(\omega)$, then we admit $x_m \geq Q_m > x^*_m(\omega)$, which is suboptimal as there is a unique optimal allocation by Theorem 1. If $Q_m < x^*_m(\omega)$ and $m \neq m_\omega$, then $x_m = Q_m$ as $c_{m_\omega} > c_m + \varepsilon$ and no agent from group m can claim a merit slot. This is suboptimal. Thus, $Q_m = x^*_m(\omega)$ for all $m \neq m_\omega$. □

Next, create ω' by changing the highest-scoring group, *i.e.*, $m_\omega \neq m_{\omega'}$. Let $x^*_{m_\omega}(\omega')$ denote the measure of admitted agents from group m_ω under ω'. Applying Claim 5, If Q attains the optimal allocation, then it must be that $Q_{m_\omega} = x^*_{m_\omega}(\omega')$. Define Q^*_m by $Q^*_m = x^*_m$ for all $m \in \mathcal{M} \setminus \{m_\omega\}$ and $Q^*_{m_\omega} = x^*_{m_\omega}(\omega')$.

Now, we have proved that if Q is an optimal policy, then $Q_m = Q^*_m$ for all $m \in \mathcal{M} \setminus \{m_\omega\}$ and $Q_{m_\omega} = Q^*_{m_\omega}$. We now establish that merit slots must be processed after any positive measure quota slots if the merit slots are of positive measure:

Claim 6. *If there is a quota policy that attains the first-best, Q, then $Q_m = Q^*_m$ and either $\sum_{m \in \mathcal{M}} Q^*_m = q$, i.e., there are no merit slots (merit slot processing does not matter), or $\sum_{m \in \mathcal{M}} Q^*_m < q$ and merit slots are processed after any positive measure quota slots.*

Proof. We have already proved $Q_m = Q^*_m$. If $\sum_{m \in \mathcal{M}} Q^*_m = q$, there are no merit slots and any processing order yields the same result. If $\sum_{m \in \mathcal{M}} Q^*_m < q$, for a contradiction, assume merit slots are processed before quota slots for group m and $Q^*_m > 0$. There are two cases, $m \neq m_\omega$ and $m = m_\omega$. We start with the first case. Note that there is a cutoff s_m for group m with $s_m < c_m + \epsilon$ and all agents from group m who score above s_m are allocated the resource. Create ω'' by taking measure $x_m/2$ of these agents who are allocated the resource and give them scores above c_{m_ω} (the highest-scoring group at ω). The scores of the remaining $x_m/2$ agents are distributed uniformly at $[c_m, c_m + \epsilon]$.

We now observe that the optimal allocations at ω and ω'' are the same. This is because increasing the scores of already admitted agents does not change the preferences of the authority of whom to admit. Moreover, the optimal allocation at ω'' cannot be attained if the quota slots for group m are processed after the merit slots. This follows as, if merit slots are processed before quota slots for group m, a strictly positive measure of them would go to group m agents at ω'' since now they have a measure of agents with the highest scores, which violates optimality.

This proves the claim for $m \neq m_\omega$. To prove the result for $m = m_\omega$, replicate the above steps with ω' where m_ω is not the highest-scoring group. □

We now use these claims to establish that if quotas are first-best optimal, then (u, h) must agree with $(\tilde{u}, \tilde{h})$ on optimal allocations.

Claim 7. *The quota first policy with $Q_m = x_m^{tar}$ maximizes the utility with respect to $\tilde{u}, \tilde{h}$.*

Proof. This is clear as for $\tilde{u}, \tilde{h}$, diversity utility dominates until x_m^{tar} and has no effect after. □

This proves the result since if there exists a first-best optimal quota policy, then it is rationalized by $(\tilde{u}, \tilde{h})$ with $x_m^{tar} = Q_m^*$. Hence, if there is a first-best quota mechanism, the authority is extremely risk-averse. □

A.3 Proof of Theorem 3

Proof. We first prove the following lemma.

Lemma 1. *Any stable matching is a cutoff matching.*

Proof. Assume that μ is a stable matching. Let $S_{m,c} = \inf_\theta\{s_c(\theta) : m(\theta) = m, \mu(\theta) = c\}$. Since μ satisfies within-group fairness, for all m and $s' > S_{m,c}$, if $m(\theta) = m$ and $s_c(\theta) = s'$, $\mu(\theta) \succeq_\theta c$. Moreover, from part (iv) of the definition of matching, this extends to the case where $s' = S_{m,c}$. Concretely, suppose that $\mu(\theta) \neq c$, $c \succ_\theta \mu(\theta)$ and $s_c(\theta) = S_{m,c}$. Consider a sequence of types $\{\theta_k\}_{k\in\mathbb{N}}$ with common group m and scores $\{s_c(\theta_k)\}_{k\in\mathbb{N}}$ such that $s_c(\theta_k) > S_{m,c}$ for all $k \in \mathbb{N}$ and $s_k(\theta) \to S_{m,c}$. Define the set $\Theta^E = \{\theta \in \Theta : c \succ \mu(\theta)\}$, which must be open by part (iv) of the definition of a matching. We have that $\theta_k \notin \Theta^E$ for all $k \in \mathbb{N}$ but $\lim_{k\to\infty} \theta_k \in \Theta^E$, which contradicts that Θ^E is open. Thus, if μ is stable, then it is also a cutoff matching. □

Therefore, to characterize stable matchings, it is enough to characterize cutoffs that induce a stable matching, which we call *stable cutoffs.*

Definition 7. *A vector S is a market-clearing cutoff if it satisfies the following:*

1. *$D_c(S) \leq q_c$ for all c.*
2. *$D_c(S) = q_c$ if $S_{m,c} > 0$ for some $m \in \mathcal{M}$.*

Since an authority can admit different measures of agents from different groups, there is a continuum of cutoffs that clear the market given S_{-c}, as long as $\{(0, \ldots, 0)\}$ is not the only market-clearing cutoff. Let $I(S_{-c})$ denote the set of market-clearing cutoffs. Let $I^*(S_{-c}) \subseteq I(S_{-c})$ denote the unique (by Lemma 2) cutoffs that implement the outcome under APM A^*_c when the authority faces the induced type measure over the set $\tilde{D}_c(S_{-c})$. Define the map $T_c : [0,1]^{|\mathcal{M}|\times|\mathcal{C}|} \to [0,1]^{|\mathcal{M}|}$ as $T_c(S) = I^*_c(S_{-c})$ with $T : [0,1]^{|\mathcal{M}|\times|\mathcal{C}|} \to [0,1]^{|\mathcal{M}|\times|\mathcal{C}|}$ given by $T = \{T_c\}_{c\in\mathcal{C}}$. We first show that the set of fixed points of T equals the set of stable cutoffs and that T is increasing.

Claim 8. *The set of fixed points of T equals the set of stable cutoffs.*

Proof. If S^*, with corresponding matching μ^* (by Lemma 1), is a fixed point of T, then each $c \neq c_0$ admits their most preferred measure q_c agents in $\tilde{D}_c(S^*_{-c})$ (by Theorem 1). Note that any $\hat{\Theta}$ that can block the matching must prefer c to their allocation at μ^* and therefore $\hat{\Theta} \subset \tilde{D}_c(S^*_{-c})$. Then there cannot be a $\hat{\Theta}$ that blocks μ^* at c since c already attains the first-best utility under $\tilde{D}_c(S^*_{-c})$ from the definition of $T_c(S)$ and Theorem 1. Conversely, if S^*, with corresponding matching μ^*, is a not fixed point of T, then there exists c such that $T_c(S^*) \neq S^*_c$. Let $\hat{\Theta}$ denote the set of agents who are not matched to c at μ^* but have scores greater than $T_c(S^*)$, and $\tilde{\Theta}$ denote the set of agents who are matched to c at μ^* but have scores lower than $T_c(S^*)$. From optimality of A^*_c (by Theorem 1), $\hat{\Theta}$ blocks μ^* at c with $\tilde{\Theta}$. □

Claim 9. *T is increasing.*

Proof. Fix an arbitrary $c \in \mathcal{C}$ and suppose that $S'_{-c} \geq S_{-c}$. Toward a contradiction suppose that there exists $m \in \mathcal{M}$ such that $t'_{c,m} = T_{c,m}(S') = I^*_c(S'_{-c}) < I^*_c(S_{-c}) = T_{c,m}(S) = t_{c,m}$, *i.e.,* the admissions threshold for group m at authority c goes down. Let f and f' be the induced joint densities of agents over scores at c and groups by the sets $\tilde{D}_c(S_{-c})$ and $\tilde{D}_c(S'_{-c})$, respectively. Let $\{x_{m,c}\}_{m\in\mathcal{M}}$ and $\{x'_{m,c}\}_{m\in\mathcal{M}}$ denote the measure agents who score above $t_{m,c}$ for their group (*i.e.,* admitted under A^*_c) under $\tilde{D}_c(S_{-c})$ and $\tilde{D}_c(S'_{-c})$, respectively. As $S'_{-c} \geq S_{-c}$, we have that $D^c(S_{-c}, S_c) \subseteq D^c(S'_{-c}, S_c)$ for all $S_c \in [0,1]^{|\mathcal{M}|}$. It follows that $f'(\theta_c) \geq f(\theta_c) > 0$ for all $\theta_c = (s_c, m_c) \in [0,1] \times \mathcal{M}$. As $t'_{c,m} < t_{c,m}$, f' has full support, and $f' \geq f$, we have that the measure of admitted group m agents under increases $x'_{c,m} > x_{c,m}$. But as $\sum_{k \in \mathcal{M}} x'_k = \sum_{k \in \mathcal{M}} x_k = q$, we know that there exists an $m' \in \mathcal{M}$ such that $x'_{c,m'} < x_{c,m'}$. It follows that $t'_{c,m'} > t_{c,m'}$, otherwise, if $t'_{c,m'} \leq t_{c,m'}$, then $x'_{c,m'} \geq x_{c,m'}$. But now we have shown the following:

$$\begin{aligned} h_c(t'_{c,m'}) + u'_{m',c}(x'_{c,m'}) &> h_c(t_{c,m'}) + u'_{m',c}(x_{c,m'}) \\ &\geq h_c(t_{c,m}) + u'_{m,c}(x_{c,m}) > h_c(t'_{c,m}) + u'_{m,c}(x'_{c,m}) \end{aligned} \tag{46}$$

where the first inequality follows by $t_{c,m'} < t'_{c,m'}$, $x_{c,m'} > x'_{c,m'}$, concavity of u_m and strictly increasing h_c. The second inequality follows by optimality. This is because the facts that $t_{c,m} > 0$ and $t'_{c,m'} < 1$ imply that $\overline{\kappa}_m = \underline{\kappa}_{m'} = 0$ and so Equation 32 implies that:

$$h_c(t_{c,m'}) + u'_{m',c}(x_{c,m'}) - \overline{\kappa}_{m'} = h_c(t_{c,m}) + u'_{m,c}(x_{c,m}) + \underline{\kappa}_m \tag{47}$$

with $\overline{\kappa}_{m'}, \underline{\kappa}_m \geq 0$. The final inequality follows as $t'_{c,m} < t_{c,m}$ and $x'_{c,m} > x_{c,m}$. But this contradicts the optimality condition for APM (Theorem 1), which implies that $T_c \neq I^*_c$, which is a contradiction. Hence, for all c and $m \in \mathcal{M}$, $T_{c,m}$ is an increasing function. □

As $T : [0,1]^{|\mathcal{M}| \times |\mathcal{C}|} \to [0,1]^{|\mathcal{M}| \times |\mathcal{C}|}$ is monotone and $[0,1]^{|\mathcal{M}| \times |\mathcal{C}|}$ is a lattice under the elementwise order $\geq$, Tarski's fixed point theorem implies that the set of stable matching cutoffs is a non-empty lattice.

Finally, we use the fact that the set of stable cutoffs is a complete lattice to argue that there is a unique cutoff consistent with stability.

Claim 10. *The stable matching cutoffs are unique.*

Proof. Assume that there are multiple stable cutoffs. As the set of stable cutoffs is a lattice, there exists a largest (S^+) and smallest (S^-) stable cutoffs, where $S^+ \geq S^-$, with strict inequality for some $m \in \mathcal{M}$, $c \in \mathcal{C}$ as $S^+ \neq S^-$. But then, as there is full support of agent types and authority c fills the capacity under stable cutoffs S^+, it must exceed its capacity under S^-, which is a contradiction. Hence, we have shown that there exists a unique stable matching cutoff. □

The combination of Lemma 1 and Claim 10 completes the proof. □

A.4 Proof of Theorem 4

Proof. If ϕ is equivalent to A^*_c, Claim 8 implies that ϕ is consistent with stability.

We prove consistency with stability implies that ϕ is equivalent to A^*_c by the contrapositive. To this end, suppose that ϕ is not equivalent to A^*_c. It follows that there exists a full-support density $\{\tilde{f}(s_c, m)\}_{s_c \in [0,1], m \in \mathcal{M}}$ such that ϕ yields a different allocation than A^*_c

under $\tilde{f}$. The rest of the proof constructs a full-support measure F with unique stable matching μ_F such that $\tilde{f}$ is the induced density of scores and groups of the agents who demand authority c at μ_F. Given such an F, we will have that ϕ cannot be consistent with stability as it yields a different allocation than A^*_c, which itself yields $\mu_F(c)$, the set of students c is matched to in the unique stable matching.

We first define some notation. Given a density f, for any set of types $\check{\Theta} \subseteq \Theta$, we define the marginal density of agents with score $s_c \in [0,1]$ at authority c in group $m \in \mathcal{M}$ as:

$$f_{\mathrm{marg}(\check{\Theta})}(s_c, m) = \int_{\check{\Theta}} \mathbb{I}[s_c(\theta) = s_c, m(\theta) = m] \, \mathrm{d}F(\theta) \tag{48}$$

To construct such an F, we proceed in three steps. First, take a full-support density f^0 that satisfies the following two conditions: i) Define $\hat{S}_c \in [0,1]^{|\mathcal{M}|}$ as the cutoff vector that obtains by applying A^*_c to $\tilde{f}$.[28] We assume that f^0 is such that authority c's cutoff vector that is consistent with the unique stable matching, μ_{F_0}, coincides with $\hat{S}_c$; ii) for all $m \in \mathcal{M}$ and $s_c < \hat{S}_{m,c}$, $f^0_{\mathrm{marg}(\Theta)}(s,m) < \tilde{f}(s,m)$; and iii) all authorities have strictly positive cutoffs for all groups at the unique stable matching.

Second, transform f^0 into a new density f^1 that differs from f^0 on the set of types that is matched with c under μ_{F_0}, which we call Θ_c.[29] We define the scaling factor $\iota^1(s_c, m)$ as:

$$\iota^1(s_c, m) = \frac{\tilde{f}(s_c, m)}{f^0_{\mathrm{marg}(\Theta_c)}(s_c, m)} \tag{49}$$

Moreover, we define:

$$f^1(\theta) = \begin{cases} f^0(\theta)\iota^1(s_c(\theta), m(\theta)) & \text{if} \quad \theta \in \Theta_c, \\ f^0(\theta) & \text{otherwise.} \end{cases} \tag{50}$$

This changes the scores of the types who are allocated to c under μ_{F^0} but does not change their total measure, their composition, or their scores at any other authority. Thus, the unique stable matching under f^1, μ_{F^1}, coincides with μ_{F^0}. Moreover, by assumption i) of step 1, we have that $f^1_{\mathrm{marg}(\Theta_c)}(s_c, m) = \tilde{f}(s_c, m)$ for all m and $s_c \geq \hat{S}_{m,c}$.

Third, transform f^1 into a new density f^2 that differs on the set of unmatched agents under f^0 (and also therefore f^1 by step 2), $\tilde{\Theta}$, and define the set of types who strictly prefer c to their assignment under μ_{F^0} (and also therefore μ_{F^1} by step 2), $\hat{\Theta}_c$.[30] We define a new

[28] Which exists as any monotone APM admits a cutoff structure (Lemma 1) and the optimal APM is monotone (Theorem 1).

[29] Formally, $\Theta_c = \{\theta : \theta \in D_c(\mu_{F^0}), s_c(\theta) \geq \hat{S}_{m(\theta),c}\}$.

[30] Formally, $\tilde{\Theta} = \{\theta : \theta \in D_c(\mu_{F^1}), s_c(\theta) < \hat{S}_{m(\theta),c}, s_{c'}(\theta) < S^{\mu_{F^1}}_{m(\theta),c'}$ for all $c' \neq c\}$, where $S^{\mu_{F^1}}_{m,c'}$ denotes the group m cutoff at school c' at the stable matching μ_{F^1}, which is strictly positive by assumption iii) of step 2. Moreover, $\hat{\Theta}_c = \{\theta : \theta \in D_c(\mu_{F^1}), s_c(\theta) < \hat{S}_{m(\theta),c}\}$.

scaling factor $\iota^2(s_c, m)$ as:

$$\iota^2(s_c, m) = \frac{\tilde{f}(s_c, m) - f_{\text{marg}(\hat{\Theta}_c)}(s_c, m)}{f_{\text{marg}(\tilde{\Theta})}(s_c, m)} \tag{51}$$

which is strictly positive by assumption ii) of step 1. We then define f^2 as:

$$f^2(\theta) = \begin{cases} f^1(\theta)(1 + \iota^2(s_c(\theta), m(\theta))) & \text{if} \quad \theta \in \tilde{\Theta}, \\ f^1(\theta) & \text{otherwise.} \end{cases} \tag{52}$$

By construction, $f^2_{\text{marg}(\hat{\Theta}_c)}(s_c, m) = \tilde{f}(s_c, m)$ for all m and $s_c < \hat{S}_{m,c}$. Moreover, $\mu_{F^2} = \mu_{F^1} = \mu_{F^0}$ as all $\theta \in \tilde{\Theta}$ remain unmatched.

We have now constructed a full-support density f^2 with unique stable matching μ_{F^2} (by Theorem 3) such that the density over $D_c(\mu_{F^2})$ coincides with $\tilde{f}$. Moreover, by Claim 8, A^*_c selects $\mu_{F^2}(c)$ from $D_c(\mu_{F^2})$. As ϕ selects a different allocation from $D_c(\mu_{F^2})$ (as it has density of types $\tilde{f}$), it is inconsistent with stability. □

A.5 Proof of Theorem 5

Proof. We prove that APM A^*_c implements a dominant strategy by backward induction. Consider the terminal time $t = |\mathcal{C}| - 1$. Some measure of agents λ applies to the authority. Regardless of the measure λ, by Theorem 1 we have that the APM A^*_c is first-best optimal (to see this more concretely, simply index λ by an arbitrary $\omega \in \Omega$ and apply Theorem 1). Thus, A^*_c is dominant. Moreover, from Theorem 1, any strategy that differs from A^*_c on a strictly positive measure set cannot be optimal. Thus any dominant strategy implements essentially the same allocation as A^*_c. Consider now any time $t < |\mathcal{C}| - 1$, precisely the same argument applies and A^*_c is (essentially uniquely) dominant. □

References

Abdulkadiroğlu, A., Y.-K. Che, and Y. Yasuda (2015): "Expanding 'choice' in school choice," *American Economic Journal: Microeconomics*, 7(1), 1–42.

Abdulkadiroğlu, A. (2005): "College Admission with Affirmative Action," *International Journal of Game Theory*, 33, 535–549.

Abdulkadiroğlu, A., and T. Sönmez (2003): "School Choice: A Mechanism Design Approach," *American Economic Review*, 93, 729–747.

Angrist, J. D., P. A. Pathak, and R. A. Zárate (2019): "Choice and consequence: Assessing mismatch at chicago exam schools," Discussion paper, NBER.

Athey, S., C. Avery, and P. Zemsky (2000): "Mentoring and diversity," *American Economic Review*, 90(4), 765–786.

Aygun, O., and I. Bó (2021): "College admission with multidimensional privileges: The Brazilian affirmative action case," *American Economic Journal: Microeconomics*, 13(3), 1–28.

Aygün, O., and B. Turhan (2020): "Dynamic reserves in matching markets," *Journal of Economic Theory*, 188, 105069.

Azevedo, E. M., and J. D. Leshno (2016): "A supply and demand framework for two-sided matching markets," *Journal of Political Economy*, 124(5), 1235–1268.

Balinski, M., and T. Sönmez (1999): "A Tale of Two Mechanisms: Student Placement," *Journal of Economic Theory*, 84, 73–94.

Boston Herald (2022): "New Boston point system dings students for good grades," *Boston Herald*, January 18, 2022.

Çelebi, O. (2022): "Diversity Preferences, Affirmative Action and Choice Rules," *Working Paper*.

Çelebi, O., and J. P. Flynn (2022): "Priority design in centralized matching markets," *The Review of Economic Studies*, 89(3), 1245–1277.

Chan, J., and E. Eyster (2003): "Does banning affirmative action lower college student quality?," *American Economic Review*, 93(3), 858–872.

Chicago Public Schools (2022): "Selective Enrollment Admissions Process: Community Feedback," *Chicago Public Schools*, March 9, 2022.

Doğan, B. (2016): "Responsive affirmative action in school choice," *Journal of Economic Theory*, 165, 69–105.

Dur, U., S. D. Kominers, P. A. Pathak, and T. Sönmez (2018): "Reserve design: Unintended consequences and the demise of Boston's walk zones," *Journal of Political Economy*, 126(6), 2457–2479.

Dur, U., P. A. Pathak, and T. Sönmez (2020): "Explicit vs. statistical targeting in affirmative action: theory and evidence from Chicago's exam schools," *Journal of Economic Theory*, p. 104996.

Echenique, F., and M. B. Yenmez (2015): "How to control controlled school choice," *American Economic Review*, 105(8), 2679–94.

Ehlers, L., I. E. Hafalir, M. B. Yenmez, and M. A. Yildirim (2014): "School Choice with Controlled Choice Constraints: Hard Bounds versus Soft Bounds," *Journal of Economic Theory*, 153, 648–683.

Ellison, G., and P. A. Pathak (2021): "The Efficiency of Race-Neutral Alternatives to Race-Based Affirmative Action: Evidence from Chicago's Exam Schools," *American Economic Review*, 111(3), 943–75.

Epple, D., R. Romano, and H. Sieg (2008): "Diversity and affirmative action in higher education," *Journal of Public Economic Theory*, 10(4), 475–501.

Erdil, A., and T. Kumano (2019): "Efficiency and stability under substitutable priorities with ties," *Journal of Economic Theory*, 184, 104950.

Goto, M., F. Kojima, R. Kurata, A. Tamura, and M. Yokoo (2017): "Designing matching mechanisms under general distributional constraints," *American Economic Journal: Microeconomics*, 9(2), 226–62.

Hafalir, I. E., M. B. Yenmez, and M. A. Yildirim (2013): "Effective Affirmative Action in School Choice," *Theoretical Economics*, 8, 325–363.

Imamura, K. (2020): "Meritocracy versus diversity," Discussion paper, Boston College.

Kamada, Y., and F. Kojima (2017): "Stability concepts in matching under distributional constraints," *Journal of Economic Theory*, 168, 107–142.

——— (2018): "Stability and strategy-proofness for matching with constraints: A necessary and sufficient condition," *Theoretical Economics*, 13(2), 761–793.

Kojima, F. (2012): "School Choice: Impossibilities for Affirmative Action," *Games and Economic Behavior*, 75, 685–693.

Kominers, S. D., and T. Sönmez (2016): "Matching with slot-specific priorities: Theory," *Theoretical Economics*, 11(2), 683–710.

Pathak, P. A., A. Rees-Jones, and T. Sönmez (2022): "Immigration lottery design: engineered and coincidental consequences of H-1B reforms," *Review of Economics and Statistics*, pp. 1–43.

Pathak, P. A., T. Sönmez, M. U. Unver, and M. B. Yenmez (2021): "Leaving No Ethical Value Behind: Triage Protocol Design for Pandemic Rationing," Discussion paper.

Sönmez, T., and M. B. Yenmez (2022a): "Affirmative action in India via vertical, horizontal, and overlapping reservations," *Econometrica*, 90(3), 1143–1176.

——— (2022b): "Constitutional Implementation of Affirmative Action Policies in India," *arXiv preprint arXiv:2203.01483*.

Temnyalov, E. (2023): "An information theory of efficient differential treatment," *American Economic Journal: Microeconomics*, 15(1), 323–358.

Tuoi Tre News (2017): "Vietnamese college hopefuls miss dream school despite acing national exam," *Tuoi Tre News*, August 4, 2017.

Weitzman, M. L. (1974): "Prices vs. quantities," *The Review of Economic Studies*, 41(4), 477–491.

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

Online Appendix

for "Adaptive Priority Mechanisms" by Çelebi and Flynn

B Omitted Proofs of Propositions

B.1 Proof of Proposition 1

Proof. Part (i): In state ω the payoff from admitting the highest-scoring minority students of measure $x(\omega)$ is:

$$q\omega + (1+\gamma-\omega)x(\omega) - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)x(\omega)^2 \tag{53}$$

Thus, the $x(\omega)$ that solves the FOC is given by:

$$x(\omega) = \frac{\kappa(1+\gamma-\omega)}{1+\kappa\gamma\beta} \tag{54}$$

Under our maintained assumptions, we have that:

$$x(\omega) = \frac{\kappa(1+\gamma-\omega)}{1+\kappa\gamma\beta} \leq \frac{1+\gamma-\underline{\omega}}{\frac{1}{\kappa}+\gamma\beta} + \kappa(\overline{\omega}-\underline{\omega}) < \min\{\kappa, q\} \tag{55}$$

and:

$$x(\omega) = \frac{\kappa(1+\gamma-\omega)}{1+\kappa\gamma\beta} \geq \frac{1+\gamma-\overline{\omega}}{\frac{1}{\kappa}+\gamma\beta} > \kappa(1-\underline{\omega}) \geq 0 \tag{56}$$

Thus, this level of minority admissions is feasible. Substituting, we have that:

$$V^* = q\mathbb{E}[\omega] + \frac{1}{2}\frac{\mathbb{E}[\kappa(1+\gamma-\omega)^2]}{1+\kappa\gamma\beta} \tag{57}$$

Consider now the APM $A(y) = \gamma(1-\beta y)$. Agents are allocated the resource if their modified scores exceed ω, with a uniform lottery over students with score exactly ω. Thus, in state ω, this policy admits measure $y(\omega)$ minorities that solve the fixed point equation:

$$y(\omega) = \min\left\{\kappa\int_0^1 \mathbb{I}[s+A(y(\omega))\geq\omega]\mathrm{d}s, q\right\} = \min\{\kappa\left(1-\max\{\omega-A(y(\omega)),0\}\right), q\} \tag{58}$$

Denote the RHS of this fixed point equation by the function $\mathrm{RHS}(y,\omega)$, which is continuous and decreasing in y. Moreover, $\mathrm{RHS}(0,\omega) = \min\{\kappa(1-\max\{\omega-\gamma,0\}),q\} > 0$ and $\mathrm{RHS}(\min\{\kappa,q\},\omega) < \min\{\kappa,q\}$. The second of these inequalities is true because the condition $\underline{\omega} > \gamma(1-\beta\min\{q,\kappa\})$ follows from our assumption that $\min\{\kappa,q\} > \frac{1+\gamma-\underline{\omega}}{\frac{1}{\kappa}+\gamma\beta} + \kappa(\overline{\omega}-\underline{\omega})$. Thus, there exists a unique $y(\omega)$ implemented by the APM. Moreover, let $y_A(\omega)$ denote the unique solution to the equation 58, which gives the measure of admitted minority students under APM A at state ω.

$$\begin{aligned} y_A(\omega) &= \kappa(1-(\omega-\gamma(1-\beta y_A(\omega)))) \\ &= \kappa(1-\omega+\gamma)-\kappa\gamma\beta y_A(\omega) \\ &= \frac{\kappa(1-\omega+\gamma)}{1+\kappa\gamma\beta} \end{aligned} \tag{59}$$

Thus, A implements the optimal level of minority admissions characterized in equation 54 and $V_A = V^*$.

Part (ii): First, if we admit all minority students over some threshold $\hat{s}$, the total score of admitted minority students is $\kappa\int_{\hat{s}}^{1} s\mathrm{d}s$. Moreover, when we admit measure x minority students where $x \leq \min\{\kappa, q\}$, this admissions threshold is defined by $x = \kappa\int_{\hat{s}}^{1}\mathrm{d}s = \kappa(1-\hat{s})$. Thus, we have that $\hat{s} = 1-\frac{x}{\kappa}$. Finally, the residual measure $q-x$ admitted majority students all score ω. Thus, the total score is given by $\bar{s} = q\omega+(1-\omega)x-\frac{1}{2\kappa}x^2$ for $0 \leq x \leq \min\{\kappa, q\}$. As both quota and priority policies always admit the highest-scoring minority students, the authority's utility is given by:

$$\mathcal{U} = q\mathbb{E}[\omega]+\mathbb{E}[(1+\gamma-\omega)x]-\frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\mathbb{E}[x^2] \tag{60}$$

We now derive the admitted measure of minority students. In the absence of a priority or quota policy, $\alpha = 0$ or $Q = 0$, we have that $x = \kappa(1-\omega)$ measure minority students is admitted. Thus, under a quota policy Q, measure $x = \max\{Q, \kappa(1-\omega)\}$ minority students are admitted. Under a priority policy, the measure of admitted minority students is $x = \kappa\int_{\omega-\alpha}^{1}\mathrm{d}x = \kappa(1+\alpha-\omega)$. In each case x is capped by $\min\{\kappa, q\}$ and floored by 0.

The expected utility function over quotas is given by one of four cases. First, $Q > \min\{\kappa, q\}$ and:

$$\mathcal{U}_Q(Q) = q\mathbb{E}[\omega]+(1+\gamma-\mathbb{E}[\omega])\min\{\kappa, q\}-\frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\min\{\kappa, q\}^2 \tag{61}$$

Second, $Q \in [\kappa(1-\underline{\omega}), \min\{\kappa, q\})$ and:[31]

$$\mathcal{U}_Q(Q) = q\mathbb{E}[\omega]+(1+\gamma-\mathbb{E}[\omega])Q-\frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)Q^2 \tag{62}$$

Third, $Q \in (\kappa(1-\overline{\omega}), \kappa(1-\underline{\omega}))$ and:

$$\begin{aligned} \mathcal{U}_Q(Q) = q\mathbb{E}[\omega] &+ \int_{1-\frac{Q}{\kappa}}^{\overline{\omega}}\left((1+\gamma-\omega)Q-\frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)Q^2\right)\mathrm{d}\Lambda(\omega) \\ &+ \int_{\underline{\omega}}^{1-\frac{Q}{\kappa}}\left((1+\gamma-\omega)\kappa(1-\omega)-\frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)(\kappa(1-\omega))^2\right)\mathrm{d}\Lambda(\omega) \end{aligned} \tag{63}$$

[31] By our maintained assumptions we have that this interval has non-empty interior.

Finally, $Q \leq \kappa(1-\overline{\omega})$ and:

$$\mathcal{U}_Q(Q) = q\mathbb{E}[\omega] + \mathbb{E}\left[(1+\gamma-\omega)\kappa(1-\omega)\right] - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\mathbb{E}\left[(\kappa(1-\omega))^2\right] \tag{64}$$

We claim that the optimum lies in the second case. See that in case two the strict maximum is attained at $Q^* = \frac{1+\gamma-\mathbb{E}[\omega]}{\frac{1}{\kappa}+\gamma\beta} \in (\kappa(1-\underline{\omega}), \min\{\kappa, q\})$, by our assumptions that $\min\{\kappa, q\} > \frac{1+\gamma-\omega}{\frac{1}{\kappa}+\gamma\beta} + \kappa(\overline{\omega}-\underline{\omega})$ and $\kappa(1-\underline{\omega}) < \frac{1+\gamma-\overline{\omega}}{\frac{1}{\kappa}+\gamma\beta}$. Moreover, in case three, the first derivative of the payoff is given by:

$$\mathcal{U}_Q'(Q) = \int_{1-\frac{Q}{\kappa}}^{\overline{\omega}} \left((1+\gamma-\omega) - \left(\frac{1}{\kappa}+\gamma\beta\right)Q\right)\mathrm{d}\Lambda(\omega) \tag{65}$$

Thus, checking that the sign of this is positive amounts to verifying that for all $Q \in (\kappa(1-\overline{\omega}), \kappa(1-\underline{\omega}))$, we have that:

$$Q < \frac{1+\gamma-\mathbb{E}[\omega|\omega \geq 1-\frac{Q}{\kappa}]}{\frac{1}{\kappa}+\gamma\beta} \tag{66}$$

As the RHS is an increasing function of Q, it suffices to show that:

$$\kappa(1-\underline{\omega}) < \frac{1+\gamma-\overline{\omega}}{\frac{1}{\kappa}+\gamma\beta} \tag{67}$$

which we have assumed. Moreover, the expected utility in the first case equals $\mathcal{U}_Q(\kappa(1-\underline{\omega}))$, thus is lower than the optimum of the second case. The expected utility in the fourth case equals $\mathcal{U}_Q(\kappa(1-\overline{\omega}))$, thus is lower than the optimum of the third case. We therefore have that:

$$V_Q = q\mathbb{E}[\omega] + (1+\gamma-\mathbb{E}[\omega])Q^* - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)Q^{*2} \tag{68}$$

We now turn to characterizing the value of priorities. There are three cases to consider. First, when $\kappa(1+\alpha-\overline{\omega}) \geq \min\{\kappa, q\}$ we have that $x = \min\{\kappa, q\}$ and:

$$\mathcal{U}_P(\alpha) = q\mathbb{E}[\omega] + (1+\gamma-\mathbb{E}[\omega])\min\{\kappa, q\} - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\min\{\kappa, q\}^2 \tag{69}$$

Second, when $\kappa(1+\alpha-\underline{\omega}) \geq \min\{\kappa, q\} \geq \kappa(1+\alpha-\overline{\omega})$ we have that:

$$\begin{aligned}\mathcal{U}_P(\alpha) = q\mathbb{E}[\omega] + &\int_{\underline{\omega}}^{1+\alpha-\min\{\frac{q}{\kappa},1\}} \left((1+\gamma-\omega)\min\{\kappa, q\} - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\min\{\kappa, q\}^2\right)\mathrm{d}\Lambda(\omega)\\ &+ \int_{1+\alpha-\min\{\frac{q}{\kappa},1\}}^{\overline{\omega}} \left((1+\gamma-\omega)\kappa(1+\alpha-\omega) - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\left[\kappa(1+\alpha-\omega)\right]^2\right)\mathrm{d}\Lambda(\omega)\end{aligned} \tag{70}$$

Finally, when $\min\{\kappa, q\} \geq \kappa(1+\alpha-\underline{\omega})$, we have that:

$$\mathcal{U}_P(\alpha) = q\mathbb{E}[\omega] + \mathbb{E}[(1+\gamma-\omega)\kappa(1+\alpha-\omega)] - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\mathbb{E}[(\kappa(1+\alpha-\omega))^2] \tag{71}$$

We claim that the optimum under our assumptions lies only in the third case. First, we argue that there is a unique local maximum in the third case. Second, we show the value in the second case is decreasing in α. By continuity, the unique optimum then lies in the third case.

First, it is helpful to write $\bar{x}(\alpha) = \kappa(1 + \alpha - \mathbb{E}[\omega])$ and $\varepsilon = \kappa(\mathbb{E}[\omega] - \omega)$. The value in the third case can then be re-expressed as:

$$\begin{aligned}\mathcal{U}_P(\alpha) &= q\mathbb{E}[\omega] + \mathbb{E}[(1+\gamma-\omega)(\bar{x}(\alpha)+\varepsilon)] - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\mathbb{E}[(\bar{x}(\alpha)+\varepsilon)^2] \\ &= q\mathbb{E}[\omega] + (1+\gamma-\mathbb{E}[\omega])\bar{x}(\alpha) - \mathbb{E}[\omega\varepsilon] - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\bar{x}(\alpha)^2 - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\mathbb{E}[\varepsilon^2]\end{aligned} \tag{72}$$

Finally, we have that $\mathbb{E}[\varepsilon^2] = \kappa^2 \mathrm{Var}[\omega]$ and $\mathbb{E}[\omega\varepsilon] = \mathrm{Cov}[\omega, \varepsilon] = -\kappa\mathrm{Var}[\omega]$. Thus:

$$\mathcal{U}_P(\alpha) = q\mathbb{E}[\omega] + (1+\gamma-\mathbb{E}[\omega])\bar{x}(\alpha) - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)\bar{x}(\alpha)^2 + \frac{\kappa}{2}(1-\kappa\gamma\beta)\mathrm{Var}[\omega] \tag{73}$$

We then see that the optimal α^* in this range sets $\bar{x}(\alpha^*) = Q^* < \min\{\kappa, q\}$. It remains only to check that this optimal α^* indeed lies within this case, or equivalently that $\kappa(1+\alpha^*-\underline{\omega}) \leq \min\{\kappa, q\}$. To this end, see that $\kappa(1+\alpha^*-\mathbb{E}[\omega]) = Q^*$, and:

$$\begin{aligned}\kappa(1+\alpha^*-\underline{\omega}) &= Q^* + \kappa(\mathbb{E}[\omega]-\underline{\omega}) \leq Q^* + \kappa(\overline{\omega}-\underline{\omega}) \\ &\leq \frac{1+\gamma-\underline{\omega}}{\frac{1}{\kappa}+\gamma\beta} + \kappa(\overline{\omega}-\underline{\omega}) < \min\{\kappa, q\}\end{aligned} \tag{74}$$

where the final inequality follows by our assumption that $\min\{\kappa, q\} > \frac{1+\gamma-\underline{\omega}}{\frac{1}{\kappa}+\gamma\beta} + \kappa(\overline{\omega}-\underline{\omega})$.

Second, in the second case we have that the first derivative of the payoff in α is given by:

$$\begin{aligned}\mathcal{U}_P'(\alpha) &= \int_{1+\alpha-\min\{\frac{q}{\kappa},1\}}^{\overline{\omega}} \frac{d}{d\alpha}\left((1+\gamma-\omega)\kappa(1+\alpha-\omega) - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)[\kappa(1+\alpha-\omega)]^2\right)\mathrm{d}\Lambda(\omega) \\ &= \kappa\int_{1+\alpha-\min\{\frac{q}{\kappa},1\}}^{\overline{\omega}}\left((1+\gamma-\omega) - \left(\frac{1}{\kappa}+\gamma\beta\right)(\bar{x}(\alpha)+\varepsilon(\omega))\right)\mathrm{d}\Lambda(\omega)\end{aligned} \tag{75}$$

Checking that the sign of this is negative for all α such that $\kappa(1+\alpha-\underline{\omega}) \geq \min\{\kappa, q\} \geq \kappa(1+\alpha-\overline{\omega})$ then amounts to checking that:

$$\bar{x}(\alpha) > \frac{1+\gamma-\mathbb{E}[\omega|\omega \geq 1+\alpha-\min\{\frac{q}{\kappa},1\}]}{\frac{1}{\kappa}+\gamma\beta} - \mathbb{E}\left[\varepsilon(\omega)|\omega \geq 1+\alpha-\min\{\frac{q}{\kappa},1\}\right] \tag{76}$$

for all $\bar{x}(\alpha) \in [\min\{\kappa, q\} - \kappa(\mathbb{E}[\omega]-\underline{\omega}), \min\{\kappa, q\} - \kappa(\mathbb{E}[\omega]-\overline{\omega})]$. So it suffices to check that the minimal possible value of the LHS exceeds the maximal possible value of the RHS. A sufficient condition for this is that:

$$\min\{\kappa, q\} - \kappa(\mathbb{E}[\omega] - \underline{\omega}) > \frac{1+\gamma-\underline{\omega}}{\frac{1}{\kappa}+\gamma\beta} - \kappa(\mathbb{E}[\omega] - \overline{\omega}) \tag{77}$$

Which holds as we assumed that $\min\{\kappa, q\} > \frac{1+\gamma-\omega}{\frac{1}{\kappa}+\gamma\beta} + \kappa(\overline{\omega} - \underline{\omega})$. Substituting the optimal priority policy $\bar{x}(\alpha) = Q^*$ in equation 71, we obtain

$$V_P = q\mathbb{E}[\omega] + (1+\gamma-\mathbb{E}[\omega])Q^* - \frac{1}{2}\left(\frac{1}{\kappa}+\gamma\beta\right)Q^{*2} + \frac{\kappa}{2}(1-\kappa\gamma\beta)\operatorname{Var}[\omega] \tag{78}$$

We have now established that:

$$\Delta = V_P - V_Q = \frac{\kappa}{2}(1-\kappa\gamma\beta)\operatorname{Var}[\omega] \tag{79}$$

Part (iii): We have V^*, V_Q, V_P. Thus, we can compute the loss from restricting to quota policies:

$$\mathcal{L}_Q = \frac{1}{2}\frac{\kappa \operatorname{Var}[\omega]}{1+\kappa\gamma\beta} \tag{80}$$

To find the loss from restricting to priority policies, we compute:

$$\mathcal{L}_P = \mathcal{L}_Q - \Delta = \frac{1}{2}(\kappa\gamma\beta)^2\frac{\kappa \operatorname{Var}[\omega]}{1+\kappa\gamma\beta} \tag{81}$$

Enveloping over these losses yields the claimed formula. □

B.2 Proof of Proposition 2

Proof. Adapting Definition 5 to single object setting, we say that a matching μ admits a cutoff structure if there exists $S(\omega) = \{S_m(\omega)\}_{m\in\mathcal{M}}$ such that $\mu(s, m; \omega) = 1$ if and only if $s \geq S_m(\omega)$. A mechanism admits a cutoff structure if it admits a cutoff structure at every ω. We will first prove that any monotone APM admits a cutoff structure.

Lemma 2. *A monotone APM admits a cutoff structure.*

Proof. For a contradiction, assume it does not. Then there exists ω and matching μ implemented by the monotone APM such that for some $m \in \mathcal{M}$, $s > s'$, $\mu(s, m; \omega) = 0$ but $\mu(s', m; \omega) = 1$. Let x_m denote the measure of group m agents allocated the resource at μ. Since A is a monotone APM and $s > s'$, we have that $A(x_m, s) > A(x_m, s')$, which contradicts that A implements μ. □

We now use Lemma 2 to show that a monotone APM implements a unique allocation. Assume for a contradiction that $A_m(y_m, s)$ is monotone and implements two different allocations, μ and μ'. Let x_m and x'_m denote the measure of type m students assigned the resource at μ and μ'. First, we prove that if μ and μ' admit the same measure of students from each group, *i.e.*, $x_m = x'_m$ for all m, then the average score of admitted students are the same.

Let s_m and s'_m denote the score of the lowest-scoring type m students assigned the resource at μ and μ'.

Claim 11. *If $x_m = x'_m$ for all $m \in \mathcal{M}$, then $\bar{s}_h(\mu, \omega) = \bar{s}_h(\mu', \omega)$*

Proof. Fix an m. Without loss of generality, let $s_m \geq s'_m$. First, since APM has cutoff structure and $x_m = x'_m$, we have that

$$\int_\Theta \mathbb{I}\{s(\theta) \in [s'_m, s_m], m(\theta) = m\} \mathrm{d}F_\omega(s, m) = 0 \tag{82}$$

Note that this holds regardless of $m' \in \mathcal{M}$ and whether $s_m \geq s'_m$ or $s'_m \geq s_m$. Therefore,

$$\begin{aligned}
\bar{s}_h(\mu, \omega) &= \int_\Theta \mu(s, m) h(s) \mathrm{d}F_\omega(s, m) \\
&= \sum_{m \in \mathcal{M}} \int_\Theta \mathbb{I}\{s(\theta) \geq s_m, m(\theta) = m\} h(s(\theta)) \mathrm{d}F_\omega(s, m) \\
&= \sum_{m \in \mathcal{M}} \int_\Theta \mathbb{I}\{s(\theta) \geq s'_m, m(\theta) = m\} h(s(\theta)) \mathrm{d}F_\omega(s, m) \\
&= \int_\Theta \mu(s, m) h(s) \mathrm{d}F_\omega(s, m) \\
&= \bar{s}_h(\mu', \omega)
\end{aligned} \tag{83}$$

where line equation holds from Equation 82 and all others are by definition. This finishes the proof of the claim. □

Therefore, if μ and μ' do not yield identical measures, then there are m and n such that $x_m > x'_m$ and $x'_n > x_n$. Since $x_m > x'_m$, it follows that $s'_m > s_m$. Likewise $x'_n > x_n$ implies that $s_n > s'_n$. Note that these imply: (i) $\mu'(s'_n, n) = 1$ while $\mu'(s'_m, n) = 0$ and (ii) $\mu(s_m, m) = 1$ while $\mu(s'_n, n) = 0$. Thus, the following inequalities hold:

$$A_n(s'_n, x'_n) > A_m(s'_m, x'_m) \geq A_m(s_m, x_m) > A_n(s'_n, x_n) \geq A_n(s'_n, x'_n) \tag{84}$$

where the first inequality follows from (i), the second inequality follows from the fact that $x'_m < x_m$ and A is monotone, the third inequality follows from (ii) and the fourth inequality follows from the fact that $x_n < x'_n$ and A is monotone. This equation yields $A_n(s'_n, x'_n) > A_n(s'_n, x'_n)$, which is a contradiction. Therefore, all allocations implemented by A yield the same x. Thus, from Lemma 2, if a monotone APM A implements μ and μ', both allocations admit the highest-scoring measure x_m agents from group m and can differ (at most) on a measure 0 set, proving that all allocations implemented by A are essentially the same. □

B.3 Proof of Proposition 3

Proof. Let x_m^* denote the measure of group m agents in the optimal allocation, with $x^* = \{x_m^*\}_{m \in \mathcal{M}}$. A priority policy $P(s,m) = h^{-1}(h(s) + u_m'(x_m^*)) = A_m(x_m^*, s)$ implements the same allocation as the optimal adaptive priority mechanism and by Theorem 1, is optimal. A quota mechanism with (Q, D) where $Q_m = x_m^*$ implements x^* for all D, and is therefore optimal. □

B.4 Proof of Proposition 4

Proof. We first prove the following claim.

Claim 12. *μ_{Σ^*} is (almost surely) a deterministic allocation that corresponds to a cutoff matching μ^*.*

Proof. Since there is a continuum of agents, under any Σ^*, with probability 1, any authority c faces a given set of agents who apply Θ_c^{A,Σ^*} with induced measure $\lambda_c^{\Sigma^*}$. As c uses APM A_c^*, with probability 1, any agent θ is admitted to an authority if and only if $s_c(\theta) \geq S_{m,c}^{\Sigma^*}$, where $S_{m,c}^{\Sigma^*}$ denotes the cutoffs when APM A_c^* is applied to agent measure $\lambda_c^{\Sigma^*}$. Since the agents have strict preferences, in any equilibrium, each agent applies to the $\succeq_\theta$ −maximal authority in $\{c : s_c(\theta) \geq S_{m,c}^{\Sigma^*}\}$, and is admitted, which establishes that μ_{Σ^*} is (almost surely) deterministic allocation that corresponds to a cutoff matching with cutoffs $S_{m,c}^{\Sigma^*}$. □

We now establish that μ_{Σ^*} is the unique stable matching of the economy.

Claim 13. *μ^* is the unique stable matching of this economy.*

Proof. For a contradiction, assume μ_{Σ^*} is not stable. Let S denote the unique cutoffs associated with μ_{Σ^*}. Since μ_{Σ^*} is not stable, by Claim 8, S is not a fixed point of T. Let $t_c = T_c(S)$. Since S is not a fixed point of T, there exists $m \in \mathcal{M}$ and $c \in \mathcal{C}$ such that $t_{m,c} \neq S_{m,c}$. Moreover, let $\{x_{m,c}^t\}_{m \in \mathcal{M}}$ and $\{x_{m,c}^s\}_{m \in \mathcal{M}}$ denote the measure of agents in $\tilde{D}_c(S_{-c})$ who are above the admission thresholds for authority c under t_c and S_c. As in Claim 9, note that if there exists m, c such that $t_{m,c} > S_{m,c}$, then from full support, we have that $x_{m,c}^s > x_{m,c}^t$. Since the authority fills its capacity in both cases, there must exists m' such that $x_{m',c}^t > x_{m',c}^s$ which is only possible if $S_{m,c} > t_{m,c}$. By an identical argument, if there is m, c such that $t_{m,c} < S_{m,c}$, then there exists m' $S_{m',c} < t_{m',c}$. Therefore, whenever $t_{m,c} \neq S_{m,c}$, there exists c and m, m' such that $t_{m,c} > S_{m,c}$ and $S_{m',c} > t_{m',c}$. But now we have shown the following:

$$h_c(S_{c,m'}) + u_{m'}(x_{c,m'}^s) > h_c(t_{c,m'}) + u_{m'}(x_{c,m'}^t) \geq h_c(t_{c,m}) + u_m(x_{c,m}^t) > h_c(S_{c,m}) + u_m(x_{c,m}^s)$$

where the first inequality follows by $t_{c,m'} < S_{c,m'}$, $x^s_{c,m'} < x^t_{c,m'}$, and concavity of u_m. The second inequality follows by optimality. This is because the facts that $t_{c,m} > 0$ and $t'_{c,m'} < 1$ imply that the Lagrange multipliers in the proof of Theorem 1 $\overline{\kappa}_m = \underline{\kappa}_{m'} = 0$. The final inequality follows since $t_{m,c} > S_{m,c}$ and $x^t_{m,c} < x^s_{m,c}$. However, this is a contradiction since $h_c(S_{c,m'}) + u_{m'}(x^s_{c,m'}) > h_c(S_{c,m}) + u_m(x^s_{c,m})$ implies that there exists $\varepsilon > 0$, an agent θ with score $s_c(\theta) = S_{c,m'} - \epsilon$ and type $m(\theta) = m$ has higher score under A^* than the agent θ' with score $s_c(\theta') = S_{c,m}$ and type $m(\theta') = m$. Since θ' is admitted to c, θ would be if it applied to c. Moreover, from full support, there is such θ whose top choice is c and the strategy of this agent is not a best response, which is a contradiction. □

The combination of Claims 12 and 13 completes the proof. □

C Additional Results for the Example (Section 2)

C.1 Formal Equivalence Between Prices *vs.* Quantities and Priorities *vs.* Quotas

The structure of the comparative advantage of priorities over quotas from Section 2 hints at a more formal relationship between our analysis of affirmative action policies and Weitzman's analysis of price and quantity regulation. In Weitzman's model, there is a single firm producing a quantity of a single good $x \in \mathbb{R}$ with production costs $C(x, \zeta)$ and benefits $B(x, \zeta')$:

$$\begin{aligned} C(x,\zeta) &= a_0(\zeta) + (C' + a_1(\zeta))(x - \hat{x}) + \frac{C''}{2}(x - \hat{x})^2 \\ B(x,\zeta') &= b_0(\zeta') + (B' + b_1(\zeta'))(x - \hat{x}) + \frac{B''}{2}(x - \hat{x})^2 \end{aligned} \tag{85}$$

where $B', C', C'' > 0$, $B'' < 0$, and ζ and ζ' are random variables. The regulator can either set a price that the firm must charge (after which the firm chooses its optimal production quantity) or mandate the production of a given quantity. The comparative advantage of prices over quantities Δ^{Weitzman} is then defined as the difference between expected benefits net of costs under the optimal price regime minus the corresponding net benefits under the optimal quantity regime. This comparative advantage is given by:

$$\Delta^{\text{Weitzman}} = \frac{C''^{-1}}{2}\left(1 + C''^{-1}B''\right)\operatorname{Var}[a_1(\zeta)] \tag{86}$$

The intuition for this formula is that when benefits are more curved than costs $|B''| > C''$, reducing variability in production is more valuable than the gain of having producers minimize costs. Thus, quantities are preferred. On the other hand, when costs are more curved than benefits, prices are preferred as there is greater production when producers have the lowest marginal costs of production.

These trade-offs are, in a certain sense, formally analogous to those that we have highlighted between priorities and quotas. In particular, under the mapping $C''^{-1} \mapsto \kappa$, $B'' \mapsto -\gamma\beta$, $\mathrm{Var}[\omega] \mapsto \mathrm{Var}[a_1(\zeta)]$, we have that $\Delta^{\text{Weitzman}} = \Delta$. The intuition for this is that C''^{-1} in the Weitzman framework determines how sensitive production is to changes in marginal cost, while κ in our framework determines how sensitive the admitted measure of minority students is to the relative scores. Moreover, B'' corresponds to curvature in the benefits of production while $\gamma\beta$ corresponds to curvature in the benefits of admitting more minority students. Finally, $\mathrm{Var}[a_1(\zeta)]$ corresponds to the authority's uncertainty in the level of marginal costs of production while $\mathrm{Var}[\omega]$ corresponds to the authority's uncertainty regarding the marginal cost of admitting more minority students in terms of lost total score. Thus, the positive selection effect whereby priorities admit more minority students in the states of the world where they score highest is directly analogous to the effect that price regulation gives rise to the greatest production in states where the firm's marginal cost is lowest. Moreover, the guarantee effect whereby quotas prevent variation in the measure of admitted minority students across states of the world is analogous to the ability of quantity regulation to stabilize the level of production. Importantly, our results therefore allow one to apply established price-theoretic intuition for the benefits of price *vs* quantity choice to matching markets without an explicit price mechanism.

C.2 Beyond Affirmative Action: Medical Resource Allocation

The lessons of this paper apply not only to affirmative action in academic admissions, but also more broadly to settings in which centralized authorities must allocate resources to various groups. One prominent such context is the allocation of medical resources during the Covid-19 pandemic. An important issue faced by hospitals is how to prioritize health workers (doctors, nurses and other staff) in the receipt of scarce medical resources: hospitals wish to both treat patients according to clinical need and ensure the health of the frontline workers needed to fight the pandemic. To map this setting to our example, suppose that the score s is an index of clinical need for a scarce medical resource available in amount q, the measure of frontline health workers is κ, and ω indexes the level of clinical need in the patients currently (or soon to be) treated by the hospital, which is unknown. The risk aversion of the authority $\gamma\beta$ corresponds to both a fear of not treating sufficiently many frontline workers and excluding too many clinically needy members of the general population.

In practice, both priority systems and quota policies have been used, as detailed extensively by Pathak, Sönmez, Unver, and Yenmez (2021).[32] The primary concern that has been voiced is that if a priority system is used, some groups (or characteristics) may be completely shut out of allocation of the scarce resource and that this is unethical, so quotas should be preferred. Our framework can be used to understand this argument: if there is an unusually high draw of ω, a priority system would lead to the allocation of very few resources to frontline workers, and vice-versa. Our Proposition 1 implies that if the authority is very averse to such outcomes ($\gamma\beta$ is high), quotas will be preferred and for exactly the reasons suggested. However, we also highlight a fundamental benefit of priority systems in inducing positive selection in allocation: when ω is high, it is beneficial that fewer resources go to the less sick medical workers and more to the relatively sicker general population. More generally, we argue that an adaptive priority mechanism that awards frontline workers a score subsidy that depends on the number of more clinically needy frontline workers could further improve outcomes.

An important additional consideration in this context arises if the hospital or authority must select a regime (priorities or quotas) before it understands the clinical need of its frontline workers κ, after which it can decide exactly how to prioritize these workers or set quotas, but before ultimate demand for medical resources ω is known. It follows from Proposition 1 that the comparative advantage of priorities over quotas is:

$$\mathbb{E}[\Delta] = \frac{1}{2}\left(\mathbb{E}[\kappa] - (\mathrm{Var}[\kappa] + \mathbb{E}^2[\kappa])\gamma\beta\right)\mathrm{Var}[\omega] \tag{87}$$

Thus, an increase in uncertainty $\mathrm{Var}[\kappa]$ regarding the need of frontline workers leads to a greater preference for quotas. This highlights a further advantage of quotas in settings where a clinical framework must be adopted in the face of uncertainty regarding the clinical needs of frontline workers, as was the case at the onset of the Covid-19 pandemic.

C.3 Optimal Precedence Orders

Thus far we have modelled quotas by first allocating minority students to quota slots and then allocating all remaining students according to the underlying score. However, we could have instead allocated $q - Q$ places to all agents according to the underlying score and then allocated the remaining Q places to minority students. The order in which quotas are processed is called the *precedence order* in the matching literature and their importance for driving outcomes has been the subject of a growing literature (see *e.g.,* Dur, Kominers, Pathak, and Sönmez, 2018; Dur, Pathak, and Sönmez, 2020; Pathak, Rees-Jones, and Sönmez, 2022). Our framework can be used to understand which precedence order is optimal,

[32] Some other papers that study the allocation of scarce medical resources are Akbarpour, Budish, Dworczak, and Kominers (2021), Grigoryan (2021) and Dur, Thayer, and Phan (2022).

a question that has not yet been addressed.

In this example, the same factors that determine whether one should prefer priorities or quotas determine whether one should prefer processing quotas second or first. By virtue of uniformity of scores, it can be shown in the relevant parameter range that a priority subsidy of α is equivalent to a quota policy of $\kappa\alpha$ when quotas are processed second. Thus, the comparative advantage of priorities over quotas is exactly equal to the comparative advantage of processing quotas second over first. The intuition is analogous: processing quotas second allows for positive selection while processing quotas first fixes the number of admitted minority students. Thus, on the one hand, when the authority is more risk-averse, they should process quota slots first to reduce the variability in the admitted measure of minority students. On the other hand, when they are less risk-averse, they should process quotas second to take advantage of the positive selection effect such policies induce.

Corollary 1. *The optimal quota-second policy achieves the same value as the optimal priority policy; quota-second policies are preferred to quota-first policies if and only if $\frac{1}{\kappa} \geq \gamma\beta$.*

Proof. We show that a quota-second policy Q is equivalent to a priority subsidy of $\alpha(Q) = \frac{Q}{\kappa}$. A quota-second policy admits the highest-scoring $x = \kappa(1-\omega)+Q$ minority students, floored by zero and capped by $\min\{\kappa, q\}$. A priority policy $\alpha(Q) = \frac{Q}{\kappa}$ admits the highest-scoring $x = \kappa(1 + \alpha(Q) - \omega) = \kappa(1 - \omega) + Q$ minority students, floored by zero and capped by $\min\{\kappa, q\}$. Thus, state-by-state, quota-second policy Q and priority subsidy $\alpha(Q) = \frac{Q}{\kappa}$ yield the same allocation. The claims then follow from Proposition 1. □

We emphasize that this equivalence is a result of the uniform distribution of scores and merely illustrates the similarity between priority policies and processing quotas second. This result does not hold in the more general model we study in the remainder of the paper. Indeed, in Theorem 2, we show that for any quota policy to be optimal in the presence of uncertainty, it must process quotas first.

D Efficient Mechanisms with Multiple Authorities

We have so far characterized the decentralized outcome, but two natural questions remain. First, is the decentralized outcome, which corresponds to the stable allocation, efficient? Second, if not, what kind of centralized solution can remedy any inefficiency? We show that the decentralized outcome is generally inefficient and that a modified, centralized APM mechanism restores efficiency.

D.1 Inefficiency of the Decentralized Outcome

The notion of efficiency that we will consider is utilitarian efficiency over authorities. A mechanism in the multi-authority setting is a function $\phi : \Omega \to \mathcal{U}$, where $\mathcal{U}$ is the set of matchings (which, by definition, encodes feasibility requirement imposed in the single authority setting). We define the total authority value Ξ_T of a mechanism ϕ under distribution $\Lambda \in \Delta(\Omega)$ as the total expected utility of the allocations induced by that mechanism:

$$\Xi_T(\phi, \Lambda) = \sum_{c \in \bar{\mathcal{C}}} \Xi_c(\phi, \Lambda) \tag{88}$$

A mechanism is efficient if it maximizes total authority value for all possible distributions:

Definition 8 (Efficiency)**.** *A mechanism ϕ^* is efficient if:*

$$\Xi_T(\phi^*, \Lambda) = \sup_{\phi} \Xi_T(\phi, \Lambda) \tag{89}$$

for all $\Lambda \in \Delta(\Omega)$.

For the remainder of the paper, so that scores are directly comparable across authorities and allocations are interior, we impose the following assumption:

Assumption 2. *Scores and preferences are such that $s_c(\theta) = s_{c'}(\theta)$, $h_c = h$ and $g_c = Id$, where Id is the identity function, for all $c, c' \in \bar{\mathcal{C}}$ and $\theta \in \Theta$. Moreover, $\lim_{x \to^+ 0} u'_{m,c}(x) = \infty$ and $u_{m,c}$ is strictly concave for all $m \in \mathcal{M}$ and $c \in \bar{\mathcal{C}}$.*

Assumption 2 makes scores a common numeraire across authorities and is akin to the standard quasi-linearity assumption in mechanism design. For example, it may be suitable in settings where the score is derived from a common index of academic attainment, such as in Chicago Public Schools. This assumption does not impose that all authorities have common marginal rates of substitution between scores and diversity, as they are allowed unrestricted heterogeneity in diversity preferences. We add the Inada condition for analytical tractability. We argue that it is also reasonable to assume that failing to admit any individuals from a given group is intolerable for authorities.

With the efficiency benchmark defined, we now demonstrate that the decentralized equilibrium outcome can fail to be efficient. We prove this result via an explicit example with two authorities, c and c' of capacity $\frac{1}{4}$, and two groups of agents, m and m' of measure $\frac{1}{2}$. All agents in group m prefer c' to c and all agents in group m' prefer c to c'. Authority c values admitting group m agents more on the margin than group m' agents, and authority c' values admitting group m' agents more on the margin than group m agents. Using the optimal APMs, both authorities admit more agents of the group whose admissions they value relatively less than the efficient benchmark. The intuition for this is that both authorities "steal" the high-scoring agents of the group whom they relatively less value from the other authority, an externality that they do not internalize.

Proposition 5 (Equilibrium Inefficiency)**.** *All authorities using the privately optimal APMs* $\{A_c^*\}_{c\in\mathcal{C}}$ *is not necessarily efficient.*

Proof. We prove the result by explicitly constructing an economy in which the optimal APMs lead to inefficiency. There are two authorities, c and c', both with capacity 1/2 and two groups of agents, m and m'. Both agent groups have a measure of 1 and their scores are uniformly distributed in $[1/2, 1]$. Authorities' utility functions are given by

$$\xi_c\left(\bar{s}_h, x\right) \equiv \bar{s}_h + \frac{1}{4}\sqrt{x_m} + \frac{1}{8}\sqrt{x_{m'}} \tag{90}$$

$$\xi_{c'}\left(\bar{s}_h, x\right) \equiv \bar{s}_h + \frac{1}{4}\sqrt{x_{m'}} + \frac{1}{8}\sqrt{x_m} \tag{91}$$

with $h(x) \equiv x$ while all agents of type m prefer authority c' to c while all agents of type m' prefer authority c to c'.[33]

We will now derive the stable outcome of this economy, which is (up to measure zero transformation) the unique outcome implemented when the authorities use the optimal APM. Let x_m^c denote the measure of type m agents at authority c. First, note that higher-scoring agents from the same group go to the more preferred authority. To see why this is true, note that if $m(\theta) = m(\theta') = m$, $s(\theta) > s(\theta')$ and $\mu(\theta) = c$ while $\mu(\theta') = c'$, c and θ would violate within group fairness since θ has higher priority at c than θ' regardless of the allocation. As a result, in any stable allocation μ, the highest-scoring $x_{m'}^c$ type m' agents are assigned to c and the next highest-scoring $x_{m'}^{c'}$ agents are assigned to authority c', while rest of the type m' agents are not assigned to any authority. The allocation for type m agents is analogous. Moreover, since $q = 1/2$ for both authorities, $x_{m'}^{c'}$=1/2-$x_{m'}^c$ and x_m^c=1/2-$x_m^{c'}$ and the allocation is completely determined by the measures $x_{m'}^c$ and $x_m^{c'}$.

[33] This assumption on the preferences and the distribution of scores violate our full support assumption, but adding an arbitrarily small full support density to all types makes arbitrarily small changes in the utility under the stable matching and optimal allocation but complicates the calculation, so we omit it for expositional clarity.

Next, note that at μ, the adaptive priority of the lowest-scoring type m and m' agents must be equal at both authorities. To see why this is true, take authority c without loss of generality. Let $s^c_{m'} = 1 - x^c_{m'}$ and $s^c_m = 1 - x^c_{m'} - x^c_m$ denote the scores of the lowest-scoring type m and m' agents and A_m denote the optimal APM. For a contradiction, assume $A_m(x^c_m, s^c_m) > A_{m'}(x^c_{m'}, s^c_{m'})$. Since agents of type m' with scores lower than s^c_m are unassigned at μ, for small enough ϵ, a type m agent with score $s^c_m - \epsilon$ and authority c blocks the matching. Similarly, assume $A_m(x^c_m, s^c_m) < A_{m'}(x^c_{m'}, s^c_{m'})$. Since agents of type m' with scores lower than $s^{c'}_m$ are assigned to authority c or unmatched at μ, a type m' agent with score $s^c_{m'} - \epsilon$ and authority c blocks the matching μ. Thus, the following equations must be satisfied:

$$A_m(x^c_m, s^c_m) = A_{m'}(x^c_{m'}, s^c_{m'}) \text{ and } A_m(x^{c'}_m, s^{c'}_m) = A_{m'}(x^{c'}_{m'}, s^{c'}_{m'}) \tag{92}$$

As the optimal APM in this setting is given by:

$$A^*_{\hat{m},\hat{c}}(y_{\hat{m}}, s) \equiv s + u'_{\hat{m},\hat{c}}(y_{\hat{m}}) \tag{93}$$

for all $\hat{m} \in \{m, m'\}$ and $\hat{c} \in \{c, c'\}$, we have that:

$$1 - x^c_{m'} + \frac{1}{8}\frac{1}{\sqrt{x^c_{m'}}} = 1 - x^c_{m'} - x^c_m + \frac{1}{4}\frac{1}{\sqrt{1/2 - x^c_{m'}}} \tag{94}$$

and:

$$1 - x^{c'}_m + \frac{1}{8}\frac{1}{\sqrt{x^{c'}_m}} = 1 - x^{c'}_m - x^{c'}_{m'} + \frac{1}{4}\frac{1}{\sqrt{1/2 - x^{c'}_m}} \tag{95}$$

These equations are identical up to relabelling and so $x^c_{m'} = x^{c'}_{m'} = x^*$ for some x^*. Thus, we need to find the solution to the following single equation to characterize the allocation:

$$1 - x^* + \frac{1}{8}\frac{1}{\sqrt{x^*}} = \frac{1}{2} + \frac{1}{4}\frac{1}{\sqrt{1/2 - x^*}} \tag{96}$$

Observe that this equation can be rewritten as the fixed point equation:

$$x^* = \frac{1}{2} + \frac{1}{8}\frac{1}{\sqrt{x^*}} - \frac{1}{4}\frac{1}{\sqrt{1/2 - x^*}} \tag{97}$$

We observe that the RHS satisfies the following properties: (i) $\lim_{x^* \to 0} \text{RHS}(x^*) = \infty$, (ii) $\lim_{x^* \to \frac{1}{2}} \text{RHS}(x^*) = -\infty$, and (iii) $\text{RHS}'(x^*) < 0$ for all $x^* \in (0, \frac{1}{2})$. Thus, there exists a unique solution. Moreover, we can guess-and-verify that this solution is $x^* = \frac{1}{4}$.

In summary, if both authorities use the optimal APM, then the outcome is

$$\mu(\theta) = \begin{cases} c & \text{if } m(\theta) = m, s(\theta) \in [1/2, 3/4) \text{ or } m(\theta) = m', s(\theta) \in [3/4, 1] \\ c' & \text{if } m(\theta) = m', s(\theta) \in [1/2, 3/4) \text{ or } m(\theta) = m, s(\theta) \in [3/4, 1] \\ \theta & \text{otherwise} \end{cases} \tag{98}$$

In this outcome, both authorities have an average score of 3/4 and admit measure 1/4 agents from both groups, giving them a utility of 15/16. Thus, total utilitarian welfare is 15/8 under the decentralized outcome.

We now show that this does not attain the efficient benchmark. A necessary condition for the (utilitarian) efficient outcome is that for c:

$$\frac{1}{4}\frac{1}{\sqrt{x_m^c}} = \frac{1}{8}\frac{1}{\sqrt{1/2 - x_m^c}} \tag{99}$$

and for c':

$$\frac{1}{4}\frac{1}{\sqrt{x_{m'}^{c'}}} = \frac{1}{8}\frac{1}{\sqrt{1/2 - x_{m'}^{c'}}} \tag{100}$$

This implies that $x_m^c = x_{m'}^{c'} = 4/10$ and $x_{m'}^c = x_m^{c'} = 1/10$. In this case, the same set of agents is admitted overall, so the score contribution to utility remains 3/4 on average across the authorities. Total utilitarian welfare is now:

$$3/2 + 1/2 \times \sqrt{4/10} + 1/4 \times \sqrt{1/10} \approx 1.895 > 1.875 = 15/16 \tag{101}$$

Completing the proof. □

D.2 An Efficient Centralized Mechanism

The inefficiency of each authority using a decentralized APM stems from the implicit incompleteness of markets: if we added the ability for authorities to pay each other for agents in the equilibrium allocation, then they would have willingness-to-pay to do so. A centralized mechanism can remedy this issue by ensuring the cross-sectional allocation of agents to authorities is optimal.

We propose the following augmentation of an APM to solve this problem, an *adaptive priority mechanism with quotas* (APM-Q). The idea behind this hybrid mechanism is to use aggregate, market-level priorities with authority-specific quotas. To this end, an APM-Q comprises an aggregate non-separable APM $\tilde{A} = \{\tilde{A}_m\}_{m\in\mathcal{M}}$ with $\tilde{A}_m : \mathbb{R}^{|\mathcal{M}|} \times [0,1] \to \mathbb{R}$ and a profile of quota functions $Q = \{Q_{m,c}\}_{m,c\in\mathcal{M}}$ with with $Q_{m,c} : \mathbb{R}^{|\mathcal{M}|} \to \mathbb{R}_+$. Intuitively, the aggregate APM pins down the aggregate measures of allocations of each group to *any* authority $\{x_m\}_{m\in\mathcal{M}}$, where $x_m = \sum_{c\in\bar{\mathcal{C}}} x_{m,c}$. The non-separability of this APM simply means that the measures of all groups matter for the adaptive priority of any agent. Given the aggregate measure of allocation for group m, the quota function for authority c assigns $Q_{m,c}(\{x_m\}_{m\in\mathcal{M}})$ agents of type m to authority c.

Definition 9 (Adaptive Priority Mechanism with Quotas)**.** *An adaptive priority mechanism with quotas* $(\tilde{A}, Q)$ *comprises a non-separable APM* $\tilde{A}$ *and a quota function* Q*. An APM-Q implements allocation* μ *if the following are satisfied:*

1. *Aggregate allocations are in order or priorities: $\mu(\theta) \in \bar{\mathcal{C}}$ if and only if for all θ' with $\mu(\theta') = c_0$, we have that:*

$$\tilde{A}_{m(\theta)}(\{x_m(\mu)\}_{m\in\mathcal{M}}, s(\theta)) > \tilde{A}_{m(\theta')}(\{x_m(\mu)\}_{m\in\mathcal{M}}, s(\theta')) \tag{102}$$

2. *Authority-level allocations are given by the corresponding quota functions:*

$$x_{m,c}(\mu) = Q_{m,c}(\{x_m(\mu)\}_{m\in\mathcal{M}}) \tag{103}$$

3. *The resources are fully allocated:*

$$\sum_{m\in\mathcal{M}} x_{m,c}(\mu) = q_c \tag{104}$$

By appropriate choice of the APM and quota functions, we can derive an APM-Q that is efficient. To this end, define the optimally-allocated aggregate utility from diversity:

$$\begin{aligned} \tilde{u}(\{x_m\}_{m\in\mathcal{M}}) &= \max_{\{x_{m,c}\}_{c\in\mathcal{C}}} \sum_{c\in\mathcal{C}} \sum_{m\in\mathcal{M}} u_{m,c}(x_{m,c}) \\ \text{s.t.} & \sum_{c\in\mathcal{C}} x_{m,c} \le x_m, \sum_{m\in\mathcal{M}} x_{m,c} \le q_c, \forall m \in \mathcal{M}, c \in \mathcal{C} \end{aligned} \tag{105}$$

Moreover, define the marginal value of aggregate group m admissions $\tilde{u}^{(m)}(y) = \frac{\partial}{\partial y_m}\tilde{u}(y)$ and the marginal value of authority capacity $\tilde{u}_{q_c}(y) = \frac{\partial}{\partial q_c}\tilde{u}(y)$. Using these marginal values, we can design an efficient APM-Q that combines market-level APMs with authority-level quotas:

Theorem 6 (Efficient APM-Q)**.** *Every allocation induced by the following APM-Q $(\tilde{A}, Q)$ is efficient:*

1. *The non-separable APM is given by $\tilde{A}_m(y, s) = h^{-1}\left(h(s) + \tilde{u}^{(m)}(y)\right)$*
2. *The quota functions are given by $Q_{m,c}(y) = \left(u'_{m,c}\right)^{-1}\left(\tilde{u}^{(m)}(y) + \tilde{u}_{q_c}(y)\right)$*

Proof. First, we define a fictitious *composite authority* with utility function defined over vectors of total scores $\bar{s}_h = \{\bar{s}^c_h\}_{c\in\mathcal{C}}$, and aggregate allocation to each group $x = \{x_m\}_{m\in\mathcal{M}}$. To do this, we define:

$$\begin{aligned} \tilde{u}(\{x_m\}_{m\in\mathcal{M}}) &= \max_{\{x_{m,c}\}_{c\in\mathcal{C}}} \sum_{c\in\mathcal{C}} \sum_{m\in\mathcal{M}} u_{m,c}(x_{m,c}) \\ \text{s.t.} & \sum_{c\in\mathcal{C}} x_{m,c} \le x_m, \sum_{m\in\mathcal{M}} x_{m,c} \le q_c, \forall m \in \mathcal{M}, c \in \mathcal{C} \end{aligned} \tag{106}$$

and $\tilde{\bar{s}}_h = \sum_{c\in\mathcal{C}} \bar{s}^c_h$. We write the utility function of this composite authority as

$$\tilde{\xi}\left(\tilde{\bar{s}}_h, x\right) = \tilde{\bar{s}}_h + \tilde{u}(x) \tag{107}$$

We first establish that $\tilde{u}$ satisfies the properties necessary to invoke Proposition 6, which establishes the optimality of the claimed APM for the fictitious authority.

Claim 14. *The function $\tilde{u}$ is concave and partially differentiable in each argument.*

Proof. First, we establish concavity. That is, for all $\lambda \in [0,1]$ and $x, x' \in \mathbb{R}_+^{|\mathcal{M}|}$, we have that $\tilde{u}(\lambda x' + (1-\lambda)x) \geq \lambda \tilde{u}(x') + (1-\lambda)\tilde{u}(x)$. Let $\{x_{m,c}^*\}_{m\in\mathcal{M},c\in\mathcal{C}}$ and $\{x_{m,c}^{*'}\}_{m\in\mathcal{M},c\in\mathcal{C}}$ correspond to optimal values under x and x'. Under $\tilde{x} = \lambda x' + (1-\lambda)x$, we have that $\lambda x_{m,c}^{*'} + (1-\lambda)x_{m,c}^*$ is feasible for all $m \in \mathcal{M}$ and $c \in \mathcal{C}$. Thus, we have that:

$$\begin{aligned}\tilde{u}(\tilde{x}) &\geq \sum_{m\in\mathcal{M}}\sum_{c\in\mathcal{C}} u_{m,c}(\lambda x_{m,c}^{*'} + (1-\lambda)x_{m,c}^*) \\ &\geq \sum_{m\in\mathcal{M}}\sum_{c\in\mathcal{C}} \lambda u_{m,c}(x_{m,c}^{*'}) + (1-\lambda)u_{m,c}(x_{m,c}^*) \\ &= \lambda\tilde{u}(x') + (1-\lambda)\tilde{u}(x)\end{aligned} \tag{108}$$

where the second inequality is by concavity of $u_{m,c}$ for all $m \in \mathcal{M}, c \in \mathcal{C}$.

Second, we establish partial differentiability in each argument. That is, for all $x \in \mathbb{R}_{++}^{|\mathcal{M}|}$, $\frac{\partial}{\partial x_m}\tilde{u}(x) = \tilde{u}^{(m)}(x)$ exists. This follows by Corollary 5 in Milgrom and Segal (2002). Concretely, the domain of optimization can be taken to be a compact and convex subset of a normed vector space – a sufficiently large cube in $\mathbb{R}_+^{|\mathcal{M}|\times|\mathcal{C}|}$ equipped with the Euclidean norm, for example. The objective function does not depend on x, and constraints are linear in x (and therefore both continuous and continuously differentiable). Moreover, as $x \gg 0$, there exists a $\{x_{m,c}\}$ that satisfies all constraints with strict inequality.

□

It follows that the objective function of the composite authority satisfies Assumption 3, and so Proposition 6 implies that the non-separable APM $\tilde{A}_m(y,s) = h^{-1}\left(h(s) + \tilde{u}^{(m)}(y)\right)$ uniquely implements the first-best optimal allocation for the composite authority.

It remains to establish that the quota functions implement the optimal allocation $\{x_{m,c}\}$ conditional on $\{x_m\}$. Let λ_m be the Lagrange multiplier on the x_m constraint, γ_c be the Lagrange multiplier on the q_c constraint and $\underline{\kappa}_{m,c}$ be the Lagrange multiplier on the positivity constraint. Under our maintained Inada condition, we have that $\underline{\kappa}_{m,c} = 0$. Moreover, by Corollary 5 in Milgrom and Segal (2002), we have that $\tilde{u}^{(m)}(x) = \lambda_m$, $\tilde{u}_{q_c}(x) = \gamma_c$, and $u'_{m,c}(x_{m,c}^*) = \lambda_m + \gamma_c - \underline{\kappa}_{m,c}$. Hence, we obtain that:

$$x_{m,c}^* = \left(u_{m,c}^{'}\right)^{-1}\left(\tilde{u}^{(m)}(x) + \tilde{u}_{q_c}(x)\right) \tag{109}$$

Thus, the following profile of quota functions implements the optimal cross-sectional allocation:

$$Q_{m,c}(x) = \left(u_{m,c}^{'}\right)^{-1}\left(\tilde{u}^{(m)}(x) + \tilde{u}_{q_c}(x)\right) \tag{110}$$

Completing the proof. □

The proof of this result constructs a fictitious aggregate authority in our single object setting. The claimed APM is optimal for this aggregate authority by a non-separable adaptation of Theorem 1. The substantial step in this proof establishes that $\tilde{u}$ is increasing, concave, and differentiable by employing the restrictions provided by Assumption 2. Then, given the allocation induced by this APM, we construct the quota function to optimally allocate the level of aggregate admissions induced by the APM.

Intuitively, this mechanism remedies inefficiency by "completing markets." There is a common "market price" for each group given by $\mathcal{P}_m = \tilde{u}^{(m)}(x)$ and an authority-level "shadow price of admissions" $\mathcal{P}_c = \tilde{u}_{q_c}(x)$. Authorities are allocated agents so that the marginal benefit of additional agents equals the sum of the market price and shadow price of admissions $u'_{m,c}(x_{m,c}) = \mathcal{P}_m + \mathcal{P}_c$. Hence, through the completion of markets, a centralized planner can allocate agents efficiently and internalize the externalities that prevented efficiency under the decentralized outcome. Notice that this market involves relatively few prices as it involves only $|\mathcal{M}|+|\mathcal{C}|$ shadow prices rather than the full set of $|\mathcal{M}|\times|\mathcal{C}|$ marginal values.

E Additional Quantitative Results

In this Appendix, we describe both the methodology and results of the two robustness exercises that are not discussed in full detail in the main text. First, we estimate the gains from APM when we assume that CPS sets one tier size for all tiers rather than separately optimizing the sizes of the four tiers. Second, we estimate the gains from APM under alternative utility functions that differentially penalize underrepresentation and overrepresentation.

E.1 Estimation with Homogeneous Reserves

As we have motivated, in this section we estimate an alternative model, where CPS chooses a single reserve size, r, instead of separate reserve sizes for all tiers. Formally, we replace the vector of reserve sizes of the four socioeconomic tiers, $r = (r_1, r_2, r_3, r_4)$ by $r = (r, r, r, r)$. In this setting, we define the marginal benefit of increasing reserve size as

$$G(r, \Lambda; \beta, \gamma) = \frac{\partial}{\partial r}\Xi(r, \Lambda; \beta, \gamma) \tag{111}$$

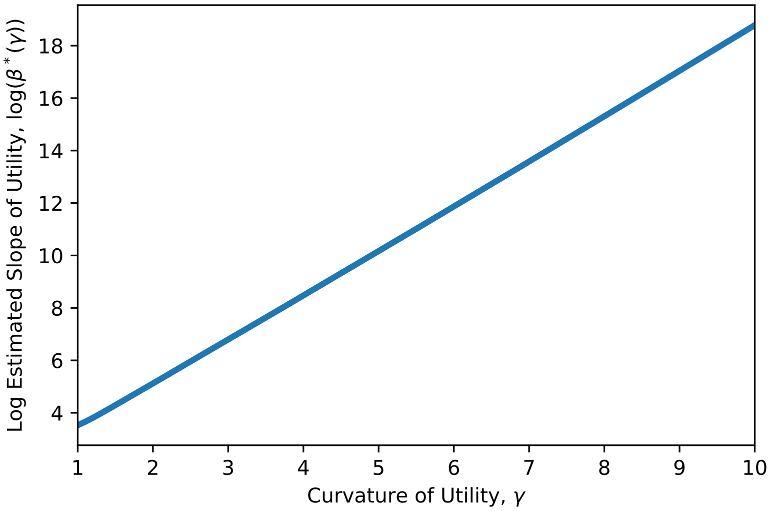
As in the general model, any (interior) reserve policy r^* must satisfy $G(\hat{r}^*, \hat{\Lambda}; \beta, \gamma) = 0$. This first-order condition yields one moment, and so we can estimate one parameter. To this end, we fix γ, and for each $\gamma \in [1, 10]$, and we estimate $\beta^*(\gamma)$ as the exact solution to the following empirical moment condition:

$$G(\hat{r}^*, \hat{\Lambda}; \beta^*(\gamma), \gamma) = 0 \tag{112}$$

Figure 7 plots the logarithm of the estimated $\beta^*(\gamma)$. The estimated $\beta^*(\gamma)$ is increasing in γ. As the loss term $|x_t - 0.25|$ is in $(0, 1)$, $\beta^*(\gamma)$ is increasing and convex in γ, where $\beta^*(1) = 34$ and $\beta^*(10) = 1.436 \times 10^8$. In Figure 8, we plot the gains as a function of γ, which shows that even though the estimated value for β moves quite a lot, the empirical gains range from 2 to 4 points. This also shows that the estimated gain from APM of 2.1 under our benchmark specification is close to the lower bound of the estimated gains under the alternative specification with homogeneous reserves.

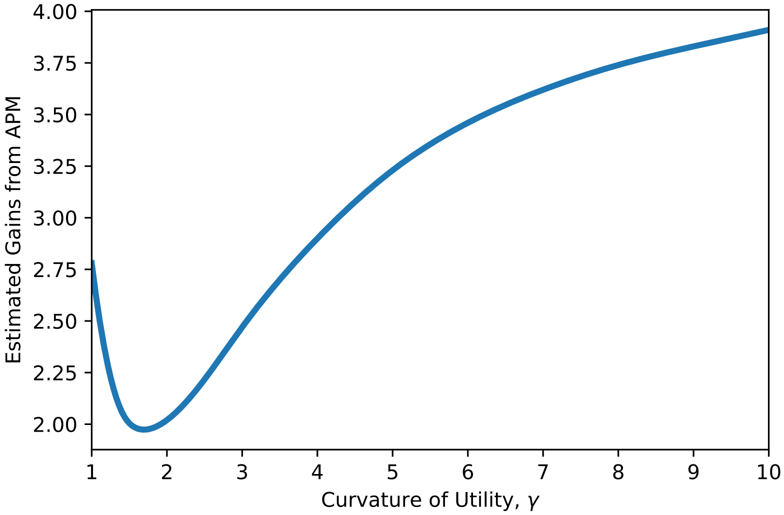
Finally, we benchmark these gains as a fraction of loss from underrepresentation under the CPS policy, where the loss of underrepresentation is calculated under the estimated parameter values. In Figure 9, we plot the gains under APM as a percentage of diversity loss under the CPS policy. These range from 26% to 300%. Our baseline percentage gain estimate of 37.5% is again close to the lower bound that we estimate under the alternative specification with homogeneous reserves.

Figure 7: Estimated Slope of Utility Under Homogeneous Reserves



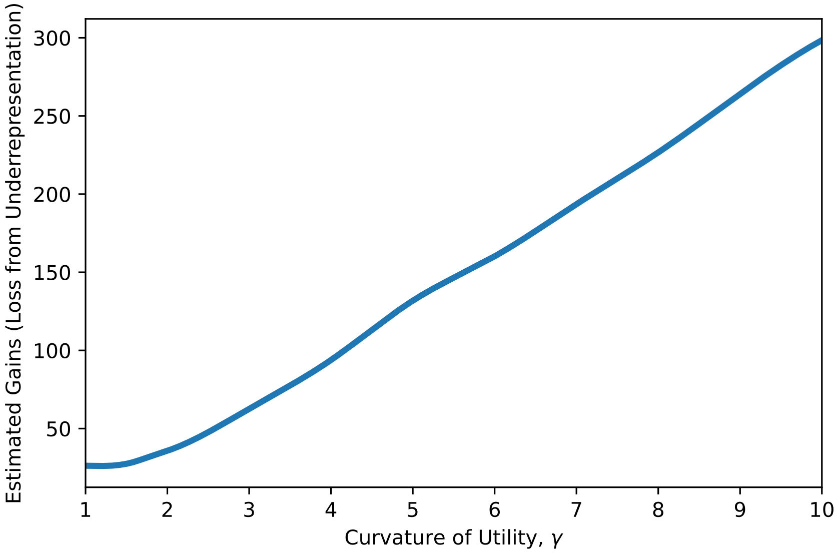
Note: This graph plots the estimated logarithm of the slope of utility $\log \beta^*(\gamma)$ in the homogeneous reserve case as we vary the curvature of utility $\gamma \in [1, 10]$.

Figure 8: Payoff Gains from APM Under Homogeneous Reserves



Note: This graph plots the estimated difference in payoffs in the homogeneous reserves case between the optimal APM and the CPS policy as we vary the curvature of utility $\gamma \in [1, 10]$.

Figure 9: The Gains from APM as a Fraction of the Loss From Underrepresentation Under Homogeneous Reserves



Note: This graph plots the estimated difference between the payoffs under the optimal APM under homogeneous reserves as a fraction of loss from underrepresentation as we vary the curvature of utility $\gamma \in [1, 10]$.

E.2 Gains from APM Under Different Utility Functions

In this section, as we have motivated, we estimate alternative objective functions to investigate the robustness of our findings.

First, we analyze a setting that includes a loss term only for underrepresented tiers (and does not penalize overrepresentation of any tier). To this end, we replace the term $|0.25 - x_t|$ with $\min\{0, (0.25 - x_t)\}$ and perform the same estimation with the following parametric utility function:

$$\xi(\bar{s}, x; \beta, \gamma) = \bar{s} + \beta \sum_{t=1}^{4} (\min\{0, (0.25 - x_t)\})^{\gamma} \tag{113}$$

The estimated parameter values are $\beta^* = -52058$ and $\gamma^* = 3.87467$. We compute the difference between the empirical payoffs under APM and the CPS reserve policy to be 0.262, which is significantly lower than our estimate of 2.1. However, the reason for this is that the diversity domain is estimated to be less important under this specification, and the diversity loss under the CPS policy is 2.71. Thus, improvements from APM correspond to 9.6% of the loss from underrepresentation, which is attenuated relative to our baseline specification, but remains non-negligible.

Second, we allow CPS to care differentially about underrepresentation and overrepresentation by considering a utility function with separate coefficients for underrepresented and overrepresented tiers. To this end, we define the following loss function:

$$f(x_t, \beta_l, \beta_h, \gamma) = \begin{cases} \beta_l(0.25 - x_t)^\gamma & \text{if } x_t \leq 0.25 \\ \beta_h(x_t - 0.25)^\gamma & \text{if } x_t > 0.25 \end{cases} \tag{114}$$

where β_l indexes the loss from underrepresentation of a tier, while β_h indexes the loss from overrepresentation. We then perform the same estimation with the following parametric utility function:

$$\xi(\bar{s}, x; \beta, \gamma) = \bar{s} + \sum_{t=1}^{4} f(x_t, \beta_l, \beta_h, \gamma) \tag{115}$$

This yields the following estimated values: $\beta_l^* = -1362270$, $\beta_h^* = -12278$, $\gamma^* = 5.28021$. We compute the difference between the empirical payoffs under APM and the CPS reserve policy to be 0.195 and the loss from underrepresentation under the CPS policy to be 2.24. Thus, we conclude that improvement from APM corresponds to 8.7% of loss from underrepresentation under the CPS policy, which is similar to what we obtain under the specification in which there is no loss from overrepresentation.

References

Akbarpour, M., E. B. Budish, P. Dworczak, and S. D. Kominers (2021): "An economic framework for vaccine prioritization," *Available at SSRN 3846931.*

Dur, U., S. D. Kominers, P. A. Pathak, and T. Sönmez (2018): "Reserve design: Unintended consequences and the demise of Boston's walk zones," *Journal of Political Economy*, 126(6), 2457–2479.

Dur, U., P. A. Pathak, and T. Sönmez (2020): "Explicit vs. statistical targeting in affirmative action: theory and evidence from Chicago's exam schools," *Journal of Economic Theory*, p. 104996.

Dur, U., M. Thayer, and W. Phan (2022): "Allocating Medical Resources During a Pandemic," *Working Paper.*

Grigoryan, A. (2021): "Effective, fair and equitable pandemic rationing," *Available at SSRN 3646539.*

Milgrom, P., and I. Segal (2002): "Envelope theorems for arbitrary choice sets," *Econometrica*, 70(2), 583–601.

Pathak, P. A., A. Rees-Jones, and T. Sönmez (2022): "Immigration lottery design:

engineered and coincidental consequences of H-1B reforms," *Review of Economics and Statistics*, pp. 1–43.

Pathak, P. A., T. Sönmez, M. U. Unver, and M. B. Yenmez (2021): "Leaving No Ethical Value Behind: Triage Protocol Design for Pandemic Rationing," Discussion paper.

Supplementary Material

for "Adaptive Priority Mechanisms" by Çelebi and Flynn

F Extension to More General Authority Preferences

In this Appendix, we relax Assumption 1 to allow for (i) non-separable diversity preferences, (ii) non-separable score and diversity preferences, (iii) non-differentiable preferences, and (iv) non-concave diversity preferences. We show how these changes in assumptions lead to certain modified APM mechanisms becoming first-best optimal.

F.1 Non-Separable Diversity Preferences

First, we relax Assumption 1 and instead suppose that the authority's preferences satisfy the following assumption:

Assumption 3. *The authority's utility function can be represented as:*

$$\xi\left(\bar{s}_h, x\right) \equiv g\left(\bar{s}_h + u(x)\right) \tag{116}$$

for some continuous, strictly increasing function $g : \mathbb{R} \to \mathbb{R}$ and a concave, partially differentiable u in each argument.

In this environment, we define a *non-separable APM* $\tilde{A} = \{\tilde{A}_m\}_{m \in \mathcal{M}}$ where $\tilde{A}_m : \mathbb{R}^{|\mathcal{M}|} \times [0,1] \to \mathbb{R}$. This implements allocation μ in state ω as per Definition 2 (under the modification of point 1 in Definition 2 to allow A_m to depend on x rather than just x_m).

We generalize Theorem 1 to show that the following non-separable APM uniquely implements the first-best optimal allocation:

Proposition 6. *The non-separable APM $\tilde{A}^*_m(y,s) \equiv h^{-1}(h(s) + u^{(m)}(y))$ and uniquely implements the first-best optimal allocation.*[34]

Proof. Follow every step in the proof of Theorem 1 with $\sum_{m \in \mathcal{M}} u_m(x_m)$ replaced by $u(x)$ and $u'_m(x_m)$ replaced by $u^{(m)}(x)$. □

Thus, allowing for non-separable diversity preferences does not substantially change the analysis of adaptive priority mechanisms. One must simply adapt the APM to be non-separable to allow cross-group diversity concerns to shape the marginal benefits of admitting agents from various groups. The main difference is that this a non-separable APM does not necessarily allow the greedy implementation of Algorithm 1. This is because, in the presence

[34] Where we define $u^{(m)}(y) = \frac{\partial}{\partial y_m} u(y)$.

of cross-group adaptive priorities, it is no longer enough to rank agents within their own group. A small adaptation to this algorithm that dynamically admits agents, starting from the highest-scoring agents in each group, would naturally implement the unique first-best optimal allocation.

F.2 Non-Separable Score and Diversity Preferences

Second, we relax Assumption 1 and instead suppose that the authority's preferences are represented by:

Assumption 4. *The authority's Bernoulli utility function can be represented as:*

$$\xi\left(\bar{s}_h, x\right) \tag{117}$$

where ξ is monotone, differentiable, and concave.

We define a state-dependent APM $\hat{A} = \{\hat{A}_m\}_{m\in\mathcal{M}}$ where $\hat{A}_m : \mathbb{R}^{|\mathcal{M}|} \times [0,1] \times \Omega \to \mathbb{R}$. This implements allocation μ in state ω as per Definition 2 (where point 1 in Definition 2 is modified to allow A_m to depend on both x and ω).

In this more general setting, we now find a state-dependent APM that implements the optimal allocation.

Proposition 7. *The following state-dependent APM implements a first-best optimal allocation:*

$$A_m(y, s, \omega) \equiv h^{-1}\left(h(s) + \frac{\xi_{x_m}\left(\bar{s}_h(y,\omega), y\right)}{\xi_{\bar{s}_h}\left(\bar{s}_h(y,\omega), y\right)}\right) \tag{118}$$

where $\bar{s}_h(y,\omega)$ is the score index in state ω when the highest-scoring $y = \{y_m\}_{m\in\mathcal{M}}$ agents of each attribute are allocated.

Proof. Follow every step in Theorem 1 with $\sum_{m\in\mathcal{M}} \int_{\tilde{\underline{s}}_m(x_m)}^{h(1)} \tilde{s}\tilde{f}_m(\tilde{s})\mathrm{d}\tilde{s} + \sum_{m\in\mathcal{M}} u_m(x_m)$ replaced with $\xi\left(\bar{s}_h(y,\omega), x\right)$ where $\bar{s}_h(y,\omega) = \sum_{m\in\mathcal{M}} \int_{\tilde{\underline{s}}_m(x_m)}^{h(1)} \tilde{s}\tilde{f}_{m,\omega}(\tilde{s})\mathrm{d}\tilde{s}$. □

There are two substantial differences in this optimal policy from our baseline APM. First, the policy depends on the joint distribution of agents in the population. Thus, specifying it *ex ante* is likely to be extremely challenging in any practical setting. This is necessary because the marginal rate of substitution between diversity and scores depends on the level of scores, which depends on the distribution of agents. Second, without assumptions on the shape of the distribution of agents, there is no guarantee that this policy is monotone and thus no guarantee that it implements a unique policy.

Thus, while Proposition 6 showed that cross-group separability is largely inessential for our main conclusions, separability between score and diversity preferences is key to the power of APM.

F.3 Non-Differentiable Preferences

In this section, we retain the majority of Assumption 1, where we instead suppose that the authority's diversity preferences $\{u_m\}_{m\in\mathcal{M}}$ are potentially non-differentiable at finitely many points.

As u_m is concave, the left and right derivatives of u_m, u_m^- and u_m^+, exist. The definition of our first-best APM is not applicable to this case since u_m' might not exist. Therefore, we define the following generalized optimal APM $A_m^*(y_m, s) \equiv h^{-1}(h(s)+u_m^-(y_m))$, which simply replaces u_m' with u_m^- in the definition. By concavity of u_m, u_m^- is monotone decreasing. Thus, this generalized optimal APM (as it is a montone APM) implements a unique allocation by Proposition 2. Moreover, the unique allocation that it implements is an optimal allocation:

Proposition 8. *Let μ^* denote the allocation implemented by the generalized optimal APM. μ^* is an optimal allocation.*

Proof. We first prove a claim. An allocation in this setting is a cutoff allocation if there exists cutoffs $\{s_m\}_{m\in\mathcal{M}}$ such that an agent θ is assigned the resource if and only if $s(\theta) \geq s_m$ and $m(\theta) = m$.

Claim 15. *There exists a unique optimal allocation μ' in the sense that all other allocations that attain the optimal payoff differ from μ' on at most a measure zero set of types. Moreover, there exists an optimal allocation that is a cutoff allocation.*

Proof. In the setting of Theorem 1, observe that $\underline{\tilde{s}}_m(x_m)$ is strictly decreasing in x_m. This, together with the concavity of u implies that the objective is strictly concave and constraints are linear. Therefore an optimal allocation exists and is unique up to measure zero transformations. Given this allocation μ' (with measures x_m), an optimal cutoff allocation is obtained by the cutoff scores s_m^* that satisfy

$$s_m' = \sup\left\{s_m \in [0,1] : \int_{s_m}^1 \tilde{f}_m(\tilde{s})\mathrm{d}\tilde{s} = x_m\right\} \tag{119}$$

□

Using this claim, toward a contradiction, assume there exists another allocation μ', which gives the authority a strictly higher utility. Moreover, take μ' to be an optimal cutoff allocation (which must exist by the claim). As μ' differs from μ^* and both are cutoff allocations, we have that there exist two groups $m, n \in \mathcal{M}$ such that: (i) $s_m' > s_m^*$ and $x_m' < x_m^*$ and (ii) $s_n' < s_n^*$ and $x_n' > x_m^*$. We have that:

$$A_m^*(x, s) \geq A_m^*(x_m^*, s) > A_m^*(x_m^*, s_m^*) \geq A_n^*(x_n^*, \hat{s}) \geq A_n^*(\hat{x}, \hat{s}) \tag{120}$$

for all $s \in (s^*_m, s'_m)$, $\hat{s} \in (s'_n, s^*_n)$, $x \leq x^*_m$, $\hat{x} \geq x^*_n$. The first inequality follows by concavity of u_m, the second follows by the fact that h is strictly increasing, the third follows by the definition of APM and the fact that $\mu^*(s^*_m, m) = 1$ and $\mu^*(\hat{s}, n) = 0$, and the fourth follows from concavity of u_n. Thus, we have that, for all $s \in (s^*_m, s'_m)$, $\hat{s} \in (s'_n, s^*_n)$, $x \leq x^*_m$, $\hat{x} \geq x^*_n$:

$$u^-_m(x) + h(s) > u^-_n(\hat{x}) + h(\hat{s}) \tag{121}$$

Thus, the total marginal utility obtained by replacing any positive measure type m students with scores $s \in (s^*_m, s'_m)$ with an identical measure of type n students with scores $\hat{s} \in (s'_n, s^*_n)$ is positive. But this contradicts the optimality of μ'. Thus, if $\tilde{\mu}$ is optimal, then $\tilde{\mu} = \mu^*$ (up to a measure zero set). □

F.4 Non-Concave Preferences

In this section, we relax the assumption that the u_m are concave.

Proposition 9. *If μ is an optimal allocation, then μ is implemented by A^*.*

Proof. Without concavity, the optimal allocation characterized in the proof of Theorem 1 is no longer unique. However, the Lagrangian conditions we have derived are still necessary for any optimal allocation $x = \{x_m\}_{m \in \mathcal{M}}$. Thus, any optimal allocation is implemented by A^*. □

This result shows that any optimal allocation is implemented by the optimal APM. However, when $\{u_m\}_{m \in \mathcal{M}}$ are not concave, A^* is not necessarily monotone. Therefore, A^* does not necessarily implement a unique allocation. Indeed, it is possible that A^* implements suboptimal allocations, as it will implement any locally optimal allocation. Therefore, a mechanism defined by an arbitrary selection from the allocations implemented by A^* would not be first-best optimal. However, A^* may still help decision-making in this setting as it implements any optimal allocation.

G Extension of the Main Results to Discrete Economies

In this Appendix, we extend the results in the main text to discrete economies and thereby establish that our analysis generalizes from the continuum framework. Concretely, we show that appropriate analogs of Theorems 1, 2, 5 and Proposition 3 carry over to discrete economies. As discrete economies do not necessarily admit a unique stable matching, Theorems 3 and 4 do not generalize. As Theorem 6 relies on convex optimization techniques, it also does not generalize as written to a discrete setting.

G.1 Primitives

An authority has q resources to allocate. At each state ω, the economy the authority faces corresponds to agents $\Theta^\omega = \{\theta_1, \ldots, \theta_{N(\omega)}\}$ where $q \leq |N(\omega)|$. As in the continuum case, $\theta \in [0,1] \times \mathcal{M}$ denotes the type of an agent who has score s and belongs to group m. We denote the score and group of any type θ by $s(\theta)$ and $m(\theta)$, respectively. For simplicity, we assume that no two agents have the same score at any ω, formally, if $\{\theta, \theta'\} \subseteq \Theta^\omega$, then $s(\theta) \neq s(\theta')$.

An allocation $\mu : \Theta \to \{0,1\}$ specifies for any type $\theta \in \Theta$ whether they are assigned to the resource. The set of possible allocations is $\mathcal{U}$ and Ω is the set of all possible economies. An allocation is feasible if it allocates no more than measure q of the resource. A mechanism is a function $\phi : \Omega \to \mathcal{U}$ that returns a feasible allocation for any possible Θ^ω.

The authority believes ω has distribution $\Lambda \in \Delta(\Omega)$. $x(\mu, \omega) = \{x_m(\mu, \omega)\}_{m \in \mathcal{M}}$ denotes the number of agents of each group allocated the resource at matching μ, while $\bar{s}_h(\mu, \omega) = \sum_{\theta \in \Theta^\omega} \mu(\theta) h(s(\theta))$ denotes the utility the authority derives from scores at μ. The preferences of the authority are given by $\xi : \mathbb{R}^{|\mathcal{M}|+1} \to \mathbb{R}$:

$$\xi\left(\bar{s}_h, x\right) \equiv \bar{s}_h + \sum_{m \in \mathcal{M}} u_m(x_m) \tag{122}$$

where h is continuous and strictly increasing and u_m is concave for all $m \in \mathcal{M}$.

G.2 Optimal Mechanisms in Discrete Economies

We adapt our definition of the Adaptive Priority Mechanisms to the discrete setting. An *adaptive priority policy* $A = \{A_m\}_{m \in \mathcal{M}}$, where $A_m : \mathbb{R} \times [0,1] \to \mathbb{R}$. The adaptive priority policy assigns priority $A_m(y_m, s)$ to an agent with score s in group m when y_m of agents of the same group is allocated the object. Given an adaptive priority policy, an APM implements allocations in the following way:

Definition 10 (Adaptive Priority Mechanism)**.** *An adaptive priority mechanism, induced by an adaptive priority A, implements an allocation μ in state ω if the following are satisfied:*

1. *Allocations are in order of priorities: $\mu(\theta) = 1$ if and only if*

   (i) for all θ' with $m(\theta') \neq m(\theta)$ and $\mu(\theta') = 0$,

$$A_{m(\theta)}(x_{m(\theta)}(\mu, \omega), s(\theta)) \geq A_{m(\theta')}(x_{m(\theta')}(\mu, \omega) + 1, s(\theta')) \tag{123}$$

   (ii) for all θ' with $m(\theta') = m(\theta)$ and $\mu(\theta') = 0$, $s(\theta) > s(\theta')$

2. The resource is fully allocated:

$$\sum_{m \in \mathcal{M}} x_m(\mu, \omega) = q \tag{124}$$

Definition 10 makes two modifications relative to the continuum model. First, the measures of agents from each group are replaced by the number of agents from each group. Second, when $m(\theta) \neq m(\theta')$, the adaptive priority of θ' is now evaluated in the case where an extra agent from $m(\theta')$ is assigned the resource.[35] Unlike the continuum case, it is possible for a monotone APM to implement two different allocations, since it can assign the same priority to two different agents, which could happen only for a zero-measure set of agents in the continuum model.

Define $A^*_m(y_m, s) \equiv h(s) + u_m(y_m) - u_m(y_m - 1)$, which will turn out to be the optimal APM. We first show that A^* is monotone, and all allocations that A^* implements give the authority the same utility.

Lemma 3. *A^* is monotone. Moreover, if A^* implements μ and $\mu' \neq \mu$ in state ω, then $\xi(\mu, \omega) = \xi(\mu', \omega)$.*

Proof. Monotonicity is immediate from the definition of A^* and concavity of u_m. Assume that A^* implements two different allocations, μ and μ' at ω. Let x_l and x'_l denote the number of group $l \in \mathcal{M}$ agents assigned the resource at μ and μ'. Since A^* is monotone and $\mu \neq \mu'$, there are m and n such that $x_m > x'_m$ and $x'_n > x_n$. Let $\tilde{\theta}_l$ and $\tilde{\theta}'_l$ denote the lowest-scoring type l agent assigned the resource at μ and μ', respectively. Similarly, let $\hat{\theta}_l$ and $\hat{\theta}'_l$ denote the highest-scoring type l agents who is not assigned the resource at μ and μ', respectively. Let $\tilde{\mu}$ denote the matching given by: $\tilde{\mu}(\theta) = \mu(\theta)$ if $\theta \notin \{\tilde{\theta}_m, \hat{\theta}'_n\}$, $\tilde{\mu}(\tilde{\theta}_m) = 0$ while $\mu(\hat{\theta}_n) = 1$. $\tilde{\mu}$ starts with μ, takes the resource away from the lowest-scoring group m agent who has it, $\tilde{\theta}_m$, and allocates it to the highest-scoring group n agent who does not have it, $\hat{\theta}_n$. Note that since A^* is monotone, from $x_m > x'_m$ and $x'_n > x_n$, under μ', $\hat{\theta}_n$ is already allocated the resource while $\tilde{\theta}_m$ is not.

Claim 16. *$\tilde{\mu}$ is implemented under A^* in state ω and $\xi(\mu, \omega) = \xi(\tilde{\mu}, \omega)$.*

Proof. Since A^* implements μ and $\mu(\hat{\theta}_n) = 0$, we have that $A^*_m(s(\tilde{\theta}_m), x_m) \geq A^*_n(s(\hat{\theta}_n), x_n + 1)$. Conversely, since A^* also implements μ' and $\mu'(\hat{\theta}'_m) = 0$, we have that $A^*_n(s(\tilde{\theta}'_n), x'_n) \geq A^*_m(s(\hat{\theta}'_m), x'_m + 1)$. Moreover, since $x_m > x'_m$ and $x'_n > x_n$, we have that $s(\hat{\theta}'_m) \geq s(\tilde{\theta}_m)$ and $s(\hat{\theta}_n) \geq s(\tilde{\theta}'_n)$. From this, it follows that:

$$\begin{aligned} A^*_n(s(\hat{\theta}_n), x_n + 1) &\geq A^*_n(s(\tilde{\theta}'_n), x_n + 1) \geq A^*_n(s(\tilde{\theta}'_n), x'_n) \\ &\geq A^*_m(s(\hat{\theta}'_m), x'_m + 1) \geq A^*_m(s(\hat{\theta}'_m), x_m) \geq A^*_m(s(\tilde{\theta}_m), x_m) \end{aligned} \tag{125}$$

[35] This was not the case in the continuum model since all types of agents have measure 0 and therefore replacing θ with θ' has no effect the evaluation of diversity.

where the first inequality holds as $s(\hat{\theta}_n) \geq s(\tilde{\theta}'_n)$, the second inequality holds as $x'_n > x_n$ (which implies $x'_n \geq x_n + 1$) and A^*_n is decreasing in its second argument, the third inequality holds as A^* also implements μ' (as stated above), the fourth inequality holds as $x'_m < x_m$ (which implies $x'_m + 1 \leq x_m$) and A^*_n is decreasing in its second argument, and the fifth inequality holds as $s(\hat{\theta}'_m) \geq s(\tilde{\theta}_m)$. Thus, $A^*_m(s(\tilde{\theta}_m), x_m) \leq A^*_n(s(\hat{\theta}_n), x_n + 1)$. This shows that $A^*_m(s(\tilde{\theta}_m), x_m) = A^*_n(s(\hat{\theta}_n), x_n + 1)$, which implies that $\tilde{\mu}$ is implemented under A^* and $\xi(\mu, \omega) = \xi(\tilde{\mu}, \omega)$. □

Note that Claim 16 shows that starting from a matching μ which is implemented by A^*, taking away the object from a particular agent who does not have it in μ' and allocating it to a particular agent who has it in μ', we arrive at another matching $\tilde{\mu}$ that is implemented under A^* and gives the authority the same payoff. Therefore, starting from any μ that is implemented by A^* and repeating this construction (by replacing μ at step i with $\tilde{\mu}$ at step $i-1$) where at each step we take the resource from an agent who is not allocated the resource at μ' and assign it to an agent who is, in finitely many steps we arrive at μ'. Since the payoff stays the same at each step, μ' gives the authority the same payoff as μ. □

Theorem 7. *If μ is implemented by A^*, then μ is an optimal matching.*

Proof. First, note that an optimal matching exists since the economy (and therefore the set of matchings) is finite. We first show the following lemma.

Lemma 4. *If μ is not implemented by A^*, then there exists μ' that gives the authority a strictly higher payoff.*

Proof. If μ is not implemented by A^*, then there exists θ and θ' such that $\mu(\theta) = 0$, $\mu(\theta') = 1$ and either $m(\theta) = m(\theta')$ and $s(\theta) > s(\theta')$ or $m(\theta) \neq m(\theta')$ and

$$\begin{aligned} &h(s(\theta)) + u_{m(\theta)}(x_{m(\theta)}(\mu) + 1) - u_{m(\theta)}(x_{m(\theta)}(\mu)) > \\ &h(s(\theta')) + u_{m(\theta')}(x_{m(\theta')}(\mu)) - u_{m(\theta')}(x_{m(\theta')}(\mu) - 1) \end{aligned} \tag{126}$$

However, in both cases, a μ' that allocates the resource to θ instead of θ' (while not changing any other agent's matching) strictly improves the utility of the authority. □

Lemma 4 proves that the optimal matching cannot be a matching that is not implemented by A^*. Since the optimal matching exists, then it is implemented by A^*. From Lemma 3, all matchings implemented by A^* give the authority the same payoff, proving the result. □

Note that Lemma 3 and Theorem 7 imply that any mechanism that is defined by an arbitrary singleton selection from the set of matchings that A^* implements would achieve the optimal matching under any ω and therefore would be first-best optimal.

G.3 Priorities *vs.* Quotas in Discrete Economies

Now, we define Priority and Quota Mechanisms in the discrete model and extend our (sub)optimality results to discrete economies.

A *priority policy* $P : \Theta \to [0, 1]$ awards an agent of type $\theta \in \Theta$ a priority $P(\theta)$.

Definition 11 (Priority Mechanisms). *A priority mechanism, induced by a priority policy P, allocates the resource in order of priorities until measure q has been allocated, with ties broken uniformly and at random.*

A *quota policy* is given by (Q, D), where $Q = \{Q_m\}_{m\in\mathcal{M}}$ and $D : \mathcal{M} \cup \{R\} \to \{1, 2, \ldots, |\mathcal{M}|+1\}$ is a bijection. The vector Q reserves Q_m objects for agents in group m, with residual capacity $Q_R = q - \sum_{m\in\mathcal{M}} Q_m$ open to agents of all types. The bijection D (often called the precedence order) determines the order in which the groups are processed.

Definition 12 (Quota Mechanisms). *A quota mechanism, induced by a quota policy (Q, D), proceeds by allocating $Q_{D^{-1}(k)}$ objects to agents from group $D^{-1}(k)$ (if there are sufficient agents from this group) to the resource in ascending order of k, and in descending order of score within each k. If there are insufficiently many agents of any group to fill the quota, the residual capacity is allocated to a final round in which all agents are eligible.*

We also extend the definitions of risk-neutrality and high risk aversion to the discrete setting. Authority preferences are *non-trivial* if for all $m, n \in \mathcal{M}$:

$$h(1) + (u_n(1) - u_n(0)) > h(0) + (u_m(q) - u_m(q - 1)) \tag{127}$$

The authority is *risk-neutral* if for all $m \in \mathcal{M}$, $u_m(x) = c_m x$ for some $c_m \geq 0$ and all $x \in \{0, 1, \ldots, q\}$. Define $\tilde{u}$ and $\tilde{h}$ as follows: there exists x_m^{tar} such that $\tilde{u}_m(x_m+1)-\tilde{u}_m(x_m) = 0$ for all $x_m \geq x_m^{\text{tar}}$ and $\tilde{u}_m(x_m+1)-\tilde{u}_m(x_m) \geq h(1)-h(0)$ for $x_m < x_m^{\text{tar}}$ and where $\sum_{m\in\mathcal{M}} x_m^{\text{tar}} \leq q$. Let $\tilde{\xi}$ denote the preferences of the authority under $\tilde{u}$ and $\tilde{h}$. The authority with preferences ξ is *extremely risk-averse* if the set of optimal allocations under ξ and $\tilde{\xi}$ coincide for all ω.

Theorem 8. *The following statements are true:*

1. *If there is no uncertainty, then there exist first-best priority and quota mechanisms.*
2. *Suppose that the authority has non-trivial preferences. There exists a first-best priority mechanism if and only if the authority is risk-neutral. This mechanism is given by $P(s, m) = s + u_m(1) - u_m(0)$.*
3. *Suppose that the authority has non-trivial preferences. There exists a first-best quota mechanism if and only if the authority is extremely risk-averse. This mechanism is given by $Q_m = x_m^{tar}$ and $D(R) = |\mathcal{M}| + 1$.*

Proof. Part (1):

Claim 17. *Let μ denote an optimal allocation at ω. Then μ is a cutoff matching.*

Proof. If μ is not a cutoff matching, then there exists (s, m) and (s', m) where $\mu(s, m) = 1$, $\mu(s', m) = 0$ and $s' > s$. Define μ' by setting: $\mu'(s, m) = 0$, $\mu(s', m) = 1$ and $\mu(\tilde{s}, \tilde{m}) = \mu(\tilde{s}, \tilde{m})$ for all $(\tilde{s}, \tilde{m})$ such that $(\tilde{s}, \tilde{m}) \notin \{(s, m), (s', m)\}$. Observe that, $\xi(\mu', \omega) - \xi(\mu, \omega) = s' - s > 0$. Therefore, μ is not an optimal allocation, which is a contradiction. □

Let μ denote an optimal allocation under ω, $\{\hat{s}_m(\mu, \omega)\}_{m \in \mathcal{M}}$ denote the cutoff scores at μ and s^* denote an arbitrary number. Any priority policy that assigns $P(\hat{s}_m(\omega), m) = s^*$ for all $m \in \mathcal{M}$ and is strictly increasing in the first argument allocates the resource to any agent who has a higher score than the cutoff for their group and implements the optimal allocation.

Let x_m denote the number of group m agents who are allocated the resource at an optimal allocation under ω. Then a quota policy that sets $Q_m = x_m$ allocates the resource to any agent who has a higher score than the cutoff for their group and implements the optimal allocation.

Part (2): The if part of the result follows from observing the priority policy $P(s, m) = s + u_m(1) - u_m(0)$ is equivalent to the optimal APM A^* under risk neutrality since $u_m(1) - u_m(0) = u_m(y_m+1) - u_m(y_m)$ for all m, y_m. Thus, by Theorem 7, $P(s, m) = s + u_m(1) - u_m(0)$ is first-best optimal.

To prove the only if part, assume risk neutrality does not hold and let m denote a group such that u_m does not satisfy risk neutrality. For a contradiction, assume that P is an optimal priority policy. First, we observe that $P(s, m)$ must be strictly increasing in s for all m. To see why, assume $P(s, m) = P(s', m)$ where $s > s'$ and just consider an ω where there are $q - 1$ group m agents with scores strictly higher than s, and no other agents. Clearly, the optimal allocation would be to allocate the resource to all agents but (s', m), while P allocates the resource to (s', m) with at least probability 1/2.

Second, let m denote a group such that u_m does not satisfy risk neutrality. Take another arbitrary group n. We have the following:

Claim 18. *Either (i) there exists $t < q$, s_m, s_n such that*

$$u_m(t+1) - u_m(t) + h(s_m) = u_n(q-t) - u_n(q-t-1) + h(s_n) \tag{128}$$

or (ii) there exists $t < q$ such that

$$u_m(t+1) - u_m(t) + h(1) < u_n(q-t) - u_n(q-t-1) + h(0) \tag{129}$$

$$u_m(t) - u_m(t-1) + h(0) > u_n(q-t+1) - u_n(q-t) + h(1) \tag{130}$$

Proof. From non-triviality, we know that $u_m(1) - u_m(0) + h(1) > u_n(q) - u_n(q-1) + h(0)$ and $u_n(1) - u_n(0) + h(1) > u_m(q) - u_m(q-1) + h(0)$. The result then follows from the fact that h is continuous and strictly increasing and u_m and u_n are concave. □

We first prove the result under case (ii). Fix two agents with scores $s_m \in (0,1)$, who belong to group m and $s_n \in (0,1)$, who belong to group n. Assume that there are $t-1$ group m agents and $q-t$ group n agents with higher scores than $\max\{s_n, s_m\}$, so a total of t group m agents and $q-t+1$ group n agents. Note that in this case, only one agent will not be allocated the resource in the optimal allocation, and that would be either (s_m, m) or (s_n, n). From equation 130, (s_m, m) is more preferred than (s_n, n) and therefore it must be that $P(s_n, n) < P(s_m, m)$, as otherwise P would not be optimal. Next, assume that there are t group m agents and $q-t-1$ group n agents with higher scores than $\max\{s_n, s_m\}$. From equation 129, (s_n, n) is more preferred than (s_m, m) and therefore it must be that $P(s_m, m) < P(s_n, n)$, which is a contradiction.

We now prove the result under case (i).

Claim 19. *In case (i), any optimal priority policy P must satisfy $P(s_m + \epsilon, m) > P(s_n, n)$ for all $\epsilon > 0$ and $P(s_m - \epsilon, m) < P(s_n, n)$ for all $\epsilon > 0$*

Proof. From Equation 128, we see that when there are t group m agents and $q-t-1$ group n agents with higher scores, $(s_m + \epsilon, m)$ is strictly preferred to (s_n, n), which is strictly preferred to $(s_m - \epsilon, m)$. □

Since u_m is not linear, there exists an l such that $u_m(l+1) - u_m(l) < u_m(l) - u_m(l-1)$. There are two possibilities: $l \leq t$ or $l > t$. First, suppose that $l \leq t$. We have that:

$$u_m(l) - u_m(l-1) + h(s_m) > u_m(l+1) - u_m(l) + h(s_m) \geq u_n(q-l) - u_n(q-l+1) + h(s_n) \tag{131}$$

where the first inequality follows from $u_m(l+1) - u_m(l) < u_m(l) - u_m(l-1)$ and the second inequality follows as $u_m(t+1) - u_m(t) + h(s_m) = u_n(q-t) - u_n(q-t-1) + h(s_n)$, u_m and u_n are concave, and $l \leq t$. Thus, for sufficiently small $\epsilon > 0$, we have that:

$$u_m(l) - u_m(l-1) + h(s_m - \epsilon) > u_n(q-l) - u_n(q-l+1) + h(s_n) \tag{132}$$

Given this inequality, we see that when there are $l-1$ group m agents and $q-l$ group n agents with higher scores, $(s_m - \epsilon, m)$ is strictly preferred to (s_n, n). Thus, to implement the optimal allocation, it must be that $P(s_m - \epsilon, m) \geq P(s_n, n)$, which is a contradiction to Claim 19.

Second, suppose that $l > t$. We know that:

$$u_m(t+1) - u_m(t) + h(s_m) = u_n(q-t) - u_n(q-t-1) + h(s_n) \tag{133}$$

As $l > t$, from concavity of u_m and u_n,

$$u_m(l) - u_m(l-1) + h(s_m) \leq u_n(q-l+1) - u_n(q-l) + h(s_n) \tag{134}$$

From concavity of u_n and u_m:

$$u_m(l+1) - u_m(l) + h(s_m) < u_n(q-l) - u_n(q-l-1) + h(s_n) \tag{135}$$

Thus, for sufficiently small $\epsilon > 0$, we have that:

$$u_m(l+1) - u_m(l) + h(s_m+\epsilon) < u_n(q-l) - u_n(q-l-1) + h(s_n) \tag{136}$$

Given this inequality, we see that when there are l group m agents and $q-l-1$ group n agents with higher scores, (s_n, n) is strictly preferred to $(s_m+\epsilon, m)$. Thus, to implement the optimal allocation, it must be that $P(s_m+\epsilon, m) \leq P(s_n, n)$, which is a contradiction to Claim 19.

Part (3): To prove the if part, fix an ω and let μ^* denote the optimal allocation under ω. Let x^*_m denote the number of group m agents allocated the resource at μ^* and $x_m(\omega)$ denote the total number of group m agents under ω.

Claim 20. *If the authority is extremely risk-averse, then $x^*_m \geq \min\{x_m(\omega), x^{tar}_m\}$*

Proof. Assume for a contradiction this is not the case. Then $x^*_m < x_m(\omega)$ and $x^*_m < x^{tar}_m$. Since $\sum_{m\in\mathcal{M}} x^{tar}_m \leq q$ and $x^*_m < x^{tar}_m$, there exists $n \in \mathcal{M}$ such that $x^*_n > x^{tar}_n$. Let s_n denote the score of the lowest-scoring group n agent who is allocated the resource, and let s_m denote the score of any group m agent who is not allocated the resource, which exists as $x^*_m < x_m(\omega)$. Since the authority is extremely risk-averse, we have the following:

$$h(s_m) + u_m(x^*_m+1) - u_m(x^*_m) > h(s_n) - u_n(x^*_n) + u_m(x^*_n - 1) \tag{137}$$

However, this contradicts the optimality of μ^* and proves the claim. □

Claim 21. *If the authority is extremely risk-averse, $x^*_m > x^{tar}_m$ and $x^*_n > x^{tar}_n$, $\mu^*(s,m) = 0$ and $\mu^*(s',n) = 1$, then $s' > s$.*

Proof. Assume for a contradiction that $s > s'$.[36] The difference in the utility of the authority when allocating the resource to (s,m) rather than (s',n) is given by

$$h(s) + u_m(x^*_m+1) - u_m(x^*_m) - (h(s') - u_n(x^*_n) + u_m(x^*_n-1)) = h(s) - h(s') > 0 \tag{138}$$

which is a contradiction to optimality of μ^*. □

The previous two claims show that under any ω, the optimal allocation admits (i) the highest-scoring x^{tar}_m agents from each group (provided that they exist) and (ii) highest-scoring

[36] Remember that $s' = s$ was ruled out by assumption.

agents who are not in (i), until the capacity is exhausted. Clearly, the quota policy $Q_m = x_m^{\text{tar}}$ and $D(R) = |\mathcal{M}| + 1$ implements this outcome at every ω.

To prove the only if part, assume that $\{Q_m\}_{m \in \mathcal{M}}$ is part of an optimal quota policy.

Claim 22. *For and each $m, n \in \mathcal{M}$ and any t, l such that $t \leq Q_m$, $Q_m > 0$ and $l \geq Q_n$, we have that:*

$$u_m(t) - u_m(t-1) + h(0) \geq u_n(l+1) - u_n(l) + h(1) \tag{139}$$

Proof. Assume that at ω, there are t group m agents, one of which one has score 0 and $l+1$ group n agents with scores higher than $1 - \epsilon_1$ and q agents from other groups who have scores higher than $1 - \epsilon_2$, where $\epsilon_1 > \epsilon_2 > 0$. As $t \leq Q_m$ and $Q_n < l+1$, t group m agents and $Q_n < l+1$ group n agents are admitted under Q. Since Q is optimal for all ϵ_1, we must have that:

$$u_m(t) - u_m(t-1) + h(0) \geq u_n(l+1) - u_n(l) + h(1 - \epsilon_1) \tag{140}$$

The statement then follows from continuity of h by taking the limit $\epsilon_1 \to 0$. □

Claim 23. *Merit slots are processed last at the optimal quota policy.*

Proof. For a contradiction, assume there is a merit slot that is processed before a quota slot. Let l denote the last merit slot that precedes a quota slot. Let m denote a group that has a quota slot after l. We consider a state in which: (i) there are q group n agents with scores $\hat{s} - \epsilon_i$, where $\epsilon_i > 0$ for all $i \in \{1, \ldots, q\}$ (let $\underline{\hat{s}}$ denote the score of the highest-scoring agent from this group), (ii) there are Q_m group m agents with scores $\hat{s} + \epsilon_j$ for $j \in \{1, \ldots, Q_m\}$ (let $\overline{\hat{s}}$ denote the score of lowest-scoring agent from this group) and one with score $\hat{s}/2$, and (iii) q agents from other groups with scores in $(\underline{\hat{s}}, \overline{\hat{s}})$. A group m agent with score $\hat{s} + \epsilon_k$ for some k is matched to l, thus $(\hat{s}/2, m)$ is matched to a later quota slot, while some agents with type $(\hat{s} - \epsilon_j, n)$ are rejected for some j. Let $\hat{s} - \epsilon_{j'}$ be the score of the highest-scoring such agent. From the optimality of the quota policy we have that

$$u_m(Q_m + 1) - u_m(Q_m) + h(\hat{s}/2) \geq u_n(Q_n + 1) - u_n(Q_n) + h(\hat{s} - \epsilon_{j'}) \tag{141}$$

Let s^* be the score of the lowest-scoring group n agent (*i.e.*, $s^* = \min_{i \in \{1,\ldots,q\}} \hat{s} - \epsilon_i$). Next, consider the modified version of the above state, all group n agents are the same, but all of the other Q_m group m agents as well as q agents from other groups now have scores in $(s^* - \hat{\epsilon}, s^*)$ and the group m agent who had a score of $\hat{s}/2$ now has a score of $\hat{s}/2 + \hat{\epsilon}$ for $\hat{\epsilon} > 0$. Note that now the group n agent with score $\hat{s} - \epsilon_{j'}$ is allocated the slot l or an earlier slot, while the agent $(\hat{s}/2 + \hat{\epsilon}, m)$ is not allocated to any slot. Thus

$$u_m(Q_m + 1) - u_m(Q_m) + h(\hat{s}/2 + \hat{\epsilon}) \leq u_n(Q_n + 1) - u_n(Q_n) + h(\hat{s} - \epsilon_{j'}) \tag{142}$$

which, since h is strictly increasing, implies that $u_m(Q_m+1) - u_m(Q_m) + h(\hat{s}/2) < u_n(Q_n+1) - u_n(Q_n) + h(\hat{s} - \epsilon_{j'})$. This contradicts Equation 141, proving the claim. □

Given the previous two claims, the following claim proves the result.

Claim 24. *If merit slots are processed last, then for all $l \geq Q_m$ and $j \geq Q_n$*

$$u_m(l+1) - u_m(l) = u_n(j+1) - u_n(j) \tag{143}$$

Proof. Assume for a contradiction this does not hold. Without loss of generality, assume $u_m(l+1) - u_m(l) > u_n(j+1) - u_n(j)$ and define δ as

$$\delta = (u_m(l+1) - u_m(l)) - (u_n(j+1) - u_n(j)) \tag{144}$$

Consider a state with $q-1$ agents with scores higher than s^*, of which exactly Q_m are group m agents and Q_n are group n agents. Moreover, there is one more group m agent with score $s' < s^*$ (denote this agent by θ_m) and one more group n agent with score $s'' \in (s', s^*)$ where $h(s'') - h(s') < \delta$ (denote this agent by θ_n). Note that all agents apart from θ_m and θ_n are allocated the resource before the final merit slot. Moreover, since θ_n has a higher score, she obtains the final merit slot. However, this is a contradiction to the optimality of Q as $h(s'') - h(s') < \delta$ and allocating that resource to θ_m gives the authority higher utility. This proves the claim. □

Taken together, claims 22 and 24 prove that a fictitious authority that is extremely risk-averse with $x_m^{\text{tar}} = Q_m$ agrees with the authority on the optimal allocation, for all ω. To see this, observe that claim 22 implies that diversity preferences dominate any concern for scores when a group is allocated less than Q_m. Moreover, conditional on being allocated at least Q_m, it is as if there is no residual diversity preference, by claim 24. This proves the only if part of (3), which finishes the proof of the result. □

G.4 Dominance of APM in Discrete Economies

In this section, we extend our discrete model to the multiple authority case and show that the dominance of the optimal APM in the decentralized admissions setting studied in Section 4.3 can be extended to this setting. Let Θ_0 denote the set of agents. $\mathcal{C} = \{c_0, c_1, \ldots, c_{|\mathcal{C}|-1}\}$ denote the set of authorities. q_c denotes the capacity of authority c and $q_{c_0} \geq |\Theta_0|$. $\theta = (s, m, \succ) \in [0,1]^{|\mathcal{C}|} \times \mathcal{M} \times \mathcal{R} = \Theta$, where $\mathcal{R}$ is set of all complete, transitive, and strict preference relations over $\mathcal{C}$ such that c_0 is less preferred than all $c \in \mathcal{C}$. For each type θ, $s_c(\theta)$ denotes the score of θ at authority c and $m(\theta)$ denotes the group of θ.

A matching in this environment is a function $\mu : \mathcal{C} \cup \Theta \to 2^{\Theta} \cup \mathcal{C}$ where $\mu(\theta) \in \mathcal{C}$ is the authority any type θ is assigned and $\mu(c) \subseteq \Theta$ is the set of agents assigned to authority c, which satisfies $|\mu(c)| \leq q_c$ for all c. $x_c(\mu) = \{x_{m,c}(\mu)\}_{m\in\mathcal{M}}$ denotes the number of agents of each group assigned to school c at μ while $\bar{s}_{h_c}(\mu) = \sum_{\theta\in\mu(c)} h(s(\theta))$ denotes the score utility the authority derives from μ. The preferences of the authority are given by:

$$\xi_c(\bar{s}_{h_c}, x_c) = \bar{s}_{h_c} + \sum_{m\in\mathcal{M}} u_{m,c}(x_{m,c}) \tag{145}$$

where h_c is continuous and strictly increasing and $u_{m,c} : \mathbb{R} \to \mathbb{R}$ is concave for all $m \in \mathcal{M}$ and $c \in \bar{\mathcal{C}}$.

Agents apply to the authorities sequentially, who decide which agents to admit. We index the stage of the game by $t \in \mathcal{T} = \{1, \ldots, |\mathcal{C}| - 1\}$. Each stage corresponds to an authority $I(t)$, where $I : \mathcal{T} \to \mathcal{T}$. At each stage t, any unmatched agents choose whether apply to authority $I(t)$. Given the set of applicants, authority $I(t)$ chooses to admit a subset of these agents. Given this, histories are indexed by the path of the remaining of agents who have not yet matched, $h^{t-1} = (\Theta_0, \Theta_1, \ldots, \Theta_{t-1}) \in \mathcal{H}^{t-1}$. Given each history h^{t-1} and set of applicants $\Theta_c^A \subseteq \Theta$, a strategy for an authority returns a set of agents $\Theta_c^G \subseteq \Theta$ whom they will admit such that $\Theta_c^G \subseteq \Theta_c^A$ and $|\Theta_c^G| \leq q_c$ for each time at which they could move $t \in \mathcal{T}$, $a_{c,t} : \mathcal{H}^{t-1} \times \mathcal{P}(\Theta) \to \mathcal{P}(\Theta)$, where $\mathcal{P}(\Theta)$ is the power set over Θ. A strategy for an agent returns a choice of whether to apply to authorities at each history and time for all agent types $\theta \in \Theta$, $\sigma_{\theta,t} : \mathcal{H}^{t-1} \to [0, 1]$. We moreover say that a strategy $a_{\tilde{c},t}$ for an authority $\tilde{c}$ at time t is *dominant* if it maximizes authority utility regardless of $\{\{a_{c,t}\}_{c\in\mathcal{C}/\{\tilde{c}\}}, \{\sigma_{\theta,t}\}_{\theta\in\Theta}\}_{t\in\mathcal{T}}$ and I.

Theorem 9. *The APM A_c^* is a dominant strategy for all authorities.*

Proof. We prove that APM A_c^* implements a dominant strategy for all authorities in all stages by backward induction. Consider the terminal time $t = |\mathcal{C}| - 1$. Some set of agents $\hat{\Theta} \subseteq \Theta$ applies to the authority. Regardless of $\hat{\Theta}$, by Theorem 7 we have that the set of agents chosen under any selection from APM A_c^* is first-best optimal. Thus, A_c^* is dominant. Consider now any time $t < |\mathcal{C}| - 1$, precisely the same argument applies and A_c^* is dominant. □

G.5 Discrete Model: Example under Imperfect Information

We develop a simple example to show how the qualitative trade-offs between priorities and quotas we have identified are those present in discrete matching markets. There are 4 students, $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and one authority c with capacity two. Students θ_3 and θ_4 belong to an underrepresented minority. The scores of minority students are distributed independently and uniformly on $[0, 1]$, so that $s_3, s_4 \sim U[0, 1]$. For simplicity, we assume there is no uncertainty over the scores of other students: $s_1 = s_2 = 1$.[37] We further specify that the authority has the following utility function:[38]

$$W(\beta, \mu) = \beta \mathbb{I}\{\mu(c) \cap \{\theta_3, \theta_4\} \neq \emptyset\} + \sum_{i:\mu(\theta_i)=c} s_i \tag{146}$$

This function embodies the main trade-off we have studied: the trade-off between scores and diversity. The first term indicates that whenever the authority admits at least one minority student, the utility of authority increases by β, which denotes the strength of affirmative action or diversity preferences. The second term simply indicates that the authority cares about scores and wants to admit the highest-scoring students they can. An alternative interpretation in this context, where allocating to agents with low scores is perceived as unfair, is that the authority wants to ensure outcomes that are fair in this sense.

The authority implements a stable matching and has two different policies at their disposal to influence the outcome of the matching mechanism. The first is a priority subsidy, denoted by $\alpha \in [0, 1]$. A subsidy α simply increases the scores of minority students by α and moves the distribution of scores of minority students to $U[\alpha, 1 + \alpha]$. The second is a minority quota $Q \in \{0, 1, 2\}$ that reserves Q seats for minority students.

We start by characterizing the first-best where the authority can choose the matching they most prefer in each state of the world. Intuitively, if the score of the highest-scoring minority student is sufficiently high, then the designer prefers to admit her and one of the non-minority students. Otherwise, it is optimal for the designer to admit the two highest-scoring students, that is the two non-minority students. The first-best matching μ^* is therefore given by:

$$\mu^*(c) = \begin{cases} \{\theta_1, \theta_2\} & \text{, if} \quad \max_{i \in \{3,4\}} s_i < 1 - \beta, \\ \{\theta_1, \theta_k\} & \text{, if} \quad \max_{i \in \{3,4\}} s_i > 1 - \beta \text{ and } s_k = \max_{i \in \{3,4\}} s_i. \end{cases} \tag{147}$$

See that if the authority had perfect information and knew s_3 and s_4 that both a priority subsidy and a quota can implement the first-best.[39] In particular, if $\max_{i \in \{3,4\}} s_i > 1 - \beta$, then both a quota $Q = 1$ that reserves one seat for minority students and a priority subsidy for minority students of $\alpha \in (1 - \min_{i \in \{3,4\}} s_i, 1 - \max_{i \in \{3,4\}} s_i)$ implement the first-best. When $\max_{i \in \{3,4\}} s_i \leq 1 - \beta$, then both a quota of $Q = 0$ and a subsidy $\alpha = 0$ implement the first-best. Thus, with perfect information, both policies yield the first-best and there is no trade-off for the authority.

[37] The qualitative result here does not change as long as non-minority students draw their scores from a distribution that FOSD $U[0, 1]$.

[38] μ denotes a matching, where $\mu(c) \subset \Theta$ and $|\mu(c)| = 2$. We assume μ is stable, which uniquely determines the allocations.

[39] Formally speaking, this is only true outside of the knife-edge case where $s_3 = s_4$, which is probability zero. In this case, there is no subsidy that can implement the first-best.

We now consider the second best where an authority is constrained to implement a priority or a quota before the realization of uncertainty. Note that implementing the first-best is impossible with both quotas and priorities as neither can be adapted to the underlying realized scores of the minority students. We now solve for the optimal quota and priority designs and compare their values. In order to characterize the optimal reserve policy, one first notes that reserving both seats for minority students is always strictly dominated by reserving only one. Thus the designer only needs to compare a policy with a quota of one against a policy with no quotas. With no quota, no minority student is admitted and the utility of the designer $W_{nr}(\beta) = 2$ with probability one. On the other hand, if a quota of one is used, only the highest-scoring minority student is admitted. Thus, the expected utility of the designer is:

$$\mathbb{E}[W(\beta)] = \beta + 1 + \mathbb{E}\left[\max\{s_3, s_4\}\right] = \frac{5}{3} + \beta \tag{148}$$

The optimal quota policy is therefore to reserve one seat for minority students if $\beta > 1/3$ and reserve no seats otherwise. Moreover, the utility of the designer under the optimal quota policy is:

$$V_Q(\beta) = \begin{cases} 2 & \text{, if} \quad \beta \leq \frac{1}{3}, \\ \dfrac{5}{3} + \beta & \text{, if} \quad \beta > \frac{1}{3}. \end{cases} \tag{149}$$

We now compute the optimal priority design and the authority's value thereof. To this end, we first calculate the utility of the designer under subsidy α. We start by calculating the matching conditional on the realized scores. There are three cases to consider. If both minority students have scores above $1-\alpha$, then both of them are admitted to the authority. If both minority students score below $1-\alpha$, then neither of them is admitted. Lastly, if one of them scores above $1-\alpha$ while the other scores below $1-\alpha$, then only one minority student is admitted. The following equation gives the utility of the designer as a function of β and α:

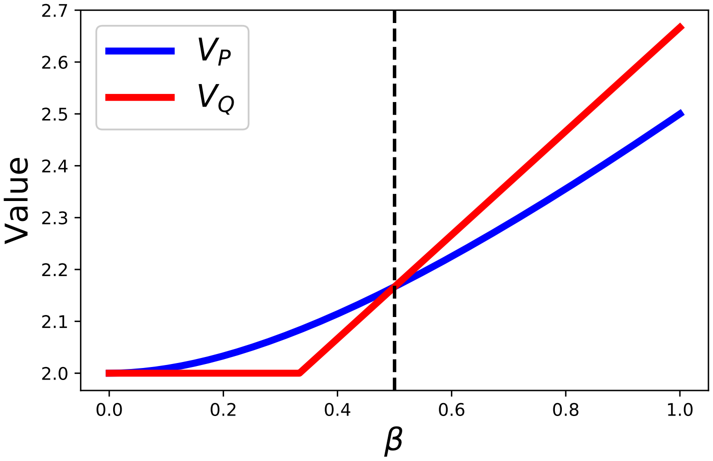
$$\begin{aligned} \mathbb{E}[W(\beta,\alpha)] = & \int_{1-\alpha}^{1}\int_{1-\alpha}^{1}(s_3 + s_4 + \beta)ds_3 ds_4 + 2\int_{1-\alpha}^{1}\int_{0}^{1-\alpha}(1 + s_4 + \beta)ds_3 ds_4 \\ & + \int_{0}^{1-\alpha}\int_{0}^{1-\alpha} 2 ds_3 ds_4 \\ = & -(1+\beta)\alpha^2 + 2\beta\alpha + 2 \end{aligned} \tag{150}$$

A quick calculation shows that the optimal subsidy is always interior and $\alpha^* = \frac{\beta}{1+\beta}$. Plugging the optimal subsidy policy into the authority's payoff function, we obtain:

$$V_P(\beta) = 1 + \beta + \frac{1}{1+\beta} \tag{151}$$

We now compare the value of the optimal quota and priority designs as the strength of

Figure 10: Comparative Statics for the Preference Between Priorities and Quotas



Note: Values of the optimal priority policy, V_P, and optimal quota policy, V_Q, as a function of the strength of the diversity preference β. The dashed black line corresponds to $\beta = 1/2$ and is the point at which both policies yield the same value.

the affirmative action motive changes. Comparing $V_P(\beta)$ and $V_Q(\beta)$ shows that the optimal policy depends on the strength of affirmative action preferences of the authority. Figure 10 plots these two values as a function of the affirmative action motive with the dotted line giving the value $\beta = \frac{1}{2}$ at which the two value functions cross. Importantly, we see that a quota policy is optimal whenever $\beta > 1/2$ and a priority subsidy policy is optimal whenever $\beta < 1/2$.

This example highlights the main differences between priorities and quotas under uncertainty and suggests when we might expect to prefer one over the other. When the preference for diversity is low, the authority only wants to admit a minority student if her score is high enough. In this case, a subsidy is a better policy as its outcome can depend on the relative scores of the students. In particular, it only admits minority students if they obtain sufficiently high scores while a quota admits minority students equally across states of the world. Consequently, priority designs generate a desirable positive selection of minority students which tends to improve scores. However, the drawback of a subsidy policy is that it applies to all students and can therefore cause either the admission of a second minority student with a lower average score or fail to admit any minority students. On the other hand, if the preference for diversity is sufficiently high, then the authority wants to admit one minority student for sure, regardless of her score. In this case, the subsidy policy is undesirable as even under the optimal subsidy, there are many realizations where neither or both minority students are admitted, while the reserve policy ensures that one minority student is admitted in all states of the world.

H Implementation, Precedence Orders, and an Illustration from H1-B Visa Allocation

In this appendix, we show that (with no uncertainty) priority and quota policies can implement the same set of allocations. We apply this insight to study the effect of precedence orders in US H1-B visa allocation.

H.1 Equivalence of Priorities and Quotas for Implementation

In Proposition 3, we showed that if there is no uncertainty, both priorities and quotas can achieve the optimal allocation. We say that a priority policy P is monotone if $P(s, m)$ is strictly increasing in s. Note that since the authority prefers higher-scoring agents to lower-scoring ones, monotone policies perform better non-monotone policies. We will now show that, in the setting of Section 3, for a given ω (which we suppress for the rest of this section), these quota and monotone priority policies are equivalent in the sense that any allocation that is achieved by one can also be achieved by the other.

Proposition 10. *μ is implemented by a quota policy if and only if it is also implemented by a monotone priority policy.*

Proof. Assume that μ is implemented by a quota policy. Then μ is a cutoff allocation since the resource is allocated in descending order of score. Let s_m denote the lowest-scoring agent from group m who is allocated the object at μ for $m \in \mathcal{M}$. Let $\bar{s} = \max_{m\in\mathcal{M}} s_m$. Define the priority policy as $P(s, m) = s + (s_m - \bar{s})$. Note that if $\mu(s, m) = 1$ and $\mu(s', m') = 0$, then $s + s_m - \bar{s} > s' + s_{m'} - \bar{s}$ and therefore $P(s, m) > P(s', m')$. As P allocates the resource to measure q highest-scoring agents under P and measure q of agents who are allocated the resource under the quota policy has higher priorities than those who are not, P implements the same allocation as the quota policy.

Conversely, assume that μ is implemented by a monotone priority policy. Let x_m denote the measure of agents from group m allocated the object at μ for $m \in \mathcal{M}$. Let Q denote a quota policy where $Q_m = x_m$. Under any processing order, Q implements the same allocation and allocates the resource to the highest-scoring measure x_m agents from group m, for all m. This is the allocation under $P(s, m)$ since P is a monotone priority policy and allocates the resource to the highest-scoring measure x_m agents from group m, which proves the result. □

In the next section, we use this result to provide a diagnostic test for evaluating quota policies with different precedence orders by the strength of the equivalent priority policy.

H.2 Application of a Diagnostic Test for the Effect of Precedence Orders to US H1-B Visa Allocation

In this Appendix, we argue that in light of Proposition 10, priorities can be used as a diagnostic test for the effect of precedence orders in the context of US H1-B Visa allocation, which has had historical issues in implementation arising from the choice of precedence order. The American H1-B visa program enables American companies to temporarily employ educated foreign workers in high-skill occupations.[40] The statutory law enacted by the U.S. Congress mandates the total number of visas to be granted and The U.S. Customs and Immigration Service (USCIS) implements this mandate. The visa allocation is governed by the H-1B Visa Reform Act of 2004 that established an annual system in which 65,000 visas were made available for all eligible applicants and an additional 20,000 visas were reserved for applicants with advanced degrees. Until 2009, USCIS used the arrival time of the application to determine priorities. Since then, the priorities are determined according to a uniform lottery.

As we have emphasized, under quota policies, specifying the processing order is critical. Between 2009 and 2019, USCIS used a Reserve-Initiated processing rule. In 2020, in accordance with the 2017 *Buy American and Hire American Executive Order*, USCIS switched to a Unreserved-Initiated rule, in order to award visas to the most-skilled workers. Pathak, Rees-Jones, and Sönmez (2022) document this switch and give a detailed account of the consequences. In particular, they calculate the effect of this change on visa allocation between 2013-2017. They find that from (on average) 55,900 applicants with advanced degrees, 33,495 of them obtain a visa under the Reserve-Initiated rule, while 38,843 of them obtain a visa under the Unreserved-Initiated rule. This fact underscores how the complexity of quota policies can lead to issues in implementation, even in the simplest case with two groups.

We now use the structure of Proposition 10 to provide a diagnostic test that authorities can use to see the degree of effective affirmative action when employing a quota policy and apply it to the H1-B and Boston Public Schools settings.[41] Concretely, when an authority is considering designing its precedence order, it can simply compute the implied priority subsidy being afforded to each group. In the context of H1-B allocation, we assume the uniform

[40] See Pathak, Rees-Jones, and Sönmez (2022) for a detailed account of H-1B policies and reforms.

[41] We note that the H-1B lottery is a setting where the perfect information assumption is justified. First, the only object that is allocated is the visa and all applicants prefer obtaining the visa to not obtaining it. This removes any uncertainty over the preferences of the individuals. Second, the priorities are determined according to a uniform random lottery and the market is large. In particular, between 2013 and 2017, each year, 85,000 visas are allocated to an average of 137,017 reserve ineligible and 55,900 reserve eligible applicants. Thus, although there is uncertainty at the individual level, the distribution of lottery numbers conditional on reserve eligibility is essentially fixed.

Table 2: Equivalent Priorities for Different Precedence Orders

	# Applicants		# Reserve-eligible Visas	
	General	Reserve-eligible	R-I Rule	NR-I Rule
5-yr Average (2013-2017)	137,017	55,900	33,495	38,834
Equivalent Subsidy (α)			23	35

Note: Allocation of H-1B visas under Reserve-Initiated and Nonreserve-Initiated rules along with the equivalent subsidies that would induce these allocations.

random lottery of USCIS is implemented by drawing a number uniformly from the interval $[0, 100]$. In the counterfactual priority mechanism, there are no quotas for reserve category applicants, but they get a score subsidy of α, i.e. their random numbers are distributed uniformly on $[\alpha, 100+\alpha]$. Computing the implied α under both processing orders to compare the policies, we obtain Table 2. Note that even though both quota policies correspond to 20,000 visas being reserved for applicants with advanced degrees, there is an important difference in the number of visas allocated to advanced degree applicants and therefore in the subsidy required to achieve that allocation. In particular, the Unreserved-Initiated order leads to a 12-point subsidy increase relative to the Reserve-Initiated benchmark.

References

Pathak, P. A., A. Rees-Jones, and T. Sönmez (2022): "Immigration lottery design: engineered and coincidental consequences of H-1B reforms," *Review of Economics and Statistics*, pp. 1–43.