Int J Game Theory (2005) 33: 535–549
DOI: 10.1007/s00182-005-0215-7

International Journal of Game Theory

College admissions with affirmative action*

Atila Abdulkadiroğlu

Department of Economics, Columbia University, NY10027, New York

Abstract. This paper first shows that when colleges' preferences are substitutable there does not exist any stable matching mechanism that makes truthful revelation of preferences a dominant strategy for every student. The paper introduces student types and captures colleges' preferences for affirmative action via type-specific quotas: A college always prefers a set of students that respects its type-specific quotas to another set that violates them. Then it shows that the student-applying deferred acceptance mechanism makes truthful revelation of preferences a dominant strategy for every student if each college's preferences satisfy responsiveness over acceptable sets of students that respect its type-specific quotas. These results have direct policy implications in several entry-level labor markets (Roth 1991). Furthermore, a fairness notion and the related incentive theory developed here is applied to controlled choice in the context of public school choice by Abdulkadiroğlu and Sönmez (2003).

1. Introduction

A college admissions problem is a many-to-one, two-sided matching problem. In several real-life applications of this problem, colleges' preferences over sets of students are determined by gender, racial and ethnic composition. For example, in the British entry-level medical labor market, Edinburgh surgeons may specify that they will employ no more than one female house officer at the same time (Roth 1991). For some residency programs in the American resident matching market, preferences of hospitals are determined by the

*I am grateful to two anonymous referees and the editor, whose comments significantly enhanced the exposition of the paper. I am also thankful to Al Roth for his valuable feedback. I would also like to thank Ron Jones, Bahar Leventoğlu, Paul Milgrom, Tayfun Sönmez and seminar participants at Caltech, Duke University, University of Montreal, University of Rochester, Stanford Institute for Theoretical Economics (SITE) Summer Workshop, SUNY Albany, and Yale University for their helpful comments. I gratefully acknowledge Alfred P. Sloan Foundation research fellowship and NSF CAREER Award SES-04-49946.

composition of professional specialities of students (Roth and Peranson 1999, Milgrom 2003).

Similar cases arise in controlled public school choice in the US. Many school districts give parents the opportunity to choose the public school their child attends. However, in some states, choice is limited by court-ordered desegregation guidelines (Abdulkadiroğlu and Sönmez 2003). Another example is New York City, where certain city high schools have to admit students across the ability range, with quotas reserved for students with low, middle and high reading scores (Abdulkadiroğlu, Pathak and Roth 2005).[1]

Giving agents the incentive to reveal their true preferences (dominant strategy incentive compatibility) is not only of theoretical interest, but also of practical concern. These incentives are studied in the literature only when colleges' preferences are represented by a simple ordering of individual students (Dubins and Freedman 1981 and Roth 1982). However, in each of the examples above, preferences of hospitals/schools cannot be represented by a simple ordering of individual doctors/students. In this paper, we fill this gap by specifying a class of preferences for colleges to capture preferences for gender or racial and ethnic compositions, and studying agents' incentives within this class.

We first show that, when colleges' preferences are substitutable, no stable mechanism is dominant strategy incentive compatible for all students (Theorem 1). This result follows Milgrom (2003), who shows that, when colleges have substitutable preferences, the student-applying deferred acceptance mechanism (SA-DAA) is not dominant strategy incentive compatible for all students.

Next, we ask: Is there a non-trivial class of preferences that captures affirmative action constraints, and at the same time yields a stable mechanism that is dominant strategy incentive compatible for all students? Our answer is positive. We formalize preferences for affirmative action by introducing type-specific quotas as follows: There exists a finite type space for students, such as {male, female} in case of Edinburgh surgeons. Each student is of exactly one of these types. In addition to its capacity, each college has a type-specific quota. We assume that a college always prefers a set of students that respects its type-specific quotas to another set that violates them. We refer to this assumption as AA (resembling Affirmative Action).

We further impose a restricted responsiveness condition (RR) on colleges' preferences: A college's preferences over sets of students are responsive to its preferences over individual students, restricting the attention only to the sets of students that are acceptable for that college. If a set does not respect a college's type-specific quotas, it is not acceptable for that college. Therefore, RR imposes responsiveness only on those sets that respect a college's type-specific quotas. RR is a generalization of responsiveness when there is more than one type of students. A preference profile that satisfies RR may fail to be responsive (Example 1).

When colleges' preferences satisfy AA and RR, they are substitutable (Lemma 1), so that the set of stable matchings is non-empty (Proposition 1). Moreover, SA-DAA produces a stable matching that every student finds at least as desirable as any other stable matching (Proposition 2).

[1]Donald Hirch (1994, page 120) notes similar constraints in UK: City Technology Colleges are required to admit students across the ability range and their student body should be representative of the community in the catchment area.

Our first main result is that, when colleges' preferences satisfy AA and RR, SA-DAA is dominant strategy incentive compatible for all students (Theorem 2). Theorem 2 provides a positive result as opposed to Milgrom's negative result and our even more negative result in Theorem 1. It is also the first to extend dominant strategy incentive compatibility for all students beyond the class of responsive preferences. The positive dominant strategy incentive compatibility result with responsive preferences due to Dubins and Freedman (1981) and Roth (1982) is a corollary of Theorem 2 with a singleton type space.

Our second contribution is a fairness notion for the controlled school choice problem. By embodying AA and RR, our fairness notion provides a novel connection through stability between a controlled school choice problem and an associated college admissions with affirmative action problem in which colleges' preferences satisfy AA and RR. Therefore, it does not only enhance our understanding of the controlled school choice problem, but also yields an important mechanism for real-life applications of the problem.

We introduce our model and give our results in Section 2. We apply our theory to the American resident matching market and controlled public school choice in Section 3. We discuss our class of preferences further in Section 4.

2. The model and the results

2.1. College admissions

A **college admissions problem** consists of:

[1] A finite set of students $S = \{s_1, \ldots, s_n\}$.
[2] A finite set of colleges $C = \{c_1, \ldots, c_m\}$.
[3] A capacity vector $q = (q_{c_1}, \ldots, q_{c_m})$, where q_c is the capacity of college $c \in C$.
[4] For every student $s \in S$, a strict preference relation P_s over $C \cup \{s\}$. Let $P^S = (P_{s_1}, \ldots, P_{s_n})$ be the profile of these relations.
[5] For every college $c \in C$, a strict preference relation P_c over subsets of S. Let $P^C = (P_{c_1}, \ldots, P_{c_m})$ be the profile of these relations.

Each preference relation is complete and transitive. For $i \in C \cup S$, let $\alpha R_i \beta$ if $\alpha P_i \beta$ or $\alpha = \beta$. From now on, small letters represent individual agents and singleton sets of individuals, whereas capital letter represent (not necessarily singleton) sets. A small letter may also represent $\varnothing$.

A **matching** μ is a function from $C \cup S$ to the set of all subsets of $C \cup S$ such that

[i] $|\mu(s)| = 1$ for every student s, and $\mu(s) = s$ if $\mu(s) \notin C$;
[ii] $\mu(c) \subset S$ and $|\mu(c)| \leq q_c$ for every college c;
[iii] $\mu(s) = c$ if and only if $s \in \mu(c)$.

$\mu(s)$ denotes s's match at μ; $\mu(c)$ denotes the set of students college c is matched with at μ.

A set of students $S' \subset S$ is **acceptable for** $c \in C$ if $S' R_c \varnothing$. A college $c \in C$ is **acceptable for** s if $c R_s s$.

Let $Ch_c(S')$ denote the most preferred subset of $S' \subset S$ for college c, i.e. $Ch_c(S') \subset S'$ and for any other $\hat{S} \subset S'$ such that $\hat{S} \neq Ch_c(S')$, $Ch_c(S') P_c \hat{S}$. We refer to $Ch_c(S')$ as **c's choice among** S'.

A matching μ is **blocked by a student** s if $\mu(s)$ is not acceptable for s. It is **blocked by a college** c if c prefers a strict subset of $\mu(c)$ to $\mu(c)$, i.e. $\mu(c) \neq Ch_c(\mu(c))$. It is **blocked by a student-college pair** (s, c) if s and c are not matched by μ but would both prefer s to be matched to c, i.e. $cP_s\mu(s)$ and $s \in Ch_c(\mu(c) \cup s)$. A matching μ is **stable** if it is not blocked by any individual agent or any student-college pair.

College c has **substitutable preferences** if for any $S' \subset S$, $s' \in S'$, $s'' \in S - s'$, when $s' \in Ch_c(S')$, $s' \in Ch_c(S' - s'')$ as well. The set of stable matchings is nonempty when every college has substitutable preferences (Kelso and Crawford 1982[2]).

Consider the following algorithm, to which we refer as the student-applying deferred acceptance algorithm (SA-DAA):

Step 1: Each student applies to her most preferred college. Each college rejects all but those in its choice among its applicants.

Step $k \geq 2$: Each student who is rejected at step $k-1$ applies to her next most preferred college. Each college considers applicants that it has not rejected. It rejects all but those in its choice among these students.

The algorithm terminates when no student is rejected. Then every student is matched with the college to which she applies last and by which she is not rejected.

When all the colleges have substitutable preferences, this algorithm produces a stable matching that every student finds at least as desirable as any other stable matching (Theorem 6.8, Roth and Sotomayor 1990, p.176).

A (direct) **mechanism** requires agents to reveal their preferences, and selects a matching based on submitted preferences. A mechanism is **dominant strategy incentive compatible (DSIC) for an agent** if revealing her true preferences is a dominant strategy for that agent in the preference revelation game induced by that mechanism. A **stable mechanism** is a mechanism that selects a matching that is stable with respect to the submitted preference profile.

It is well known that no stable mechanism is DSIC for every agent (Theorem 3, Roth 1982). However, there exist restrictions on preferences that are sufficient for the existence of a stable mechanism that is DSIC for every student. To define these restrictions, we introduce additional definitions.

Each P_c induces a complete, transitive and strict preference relation for c over *singletons of students and* $\emptyset$. In particular $sP_cs'\emptyset P_cP_cs''$ reads as follows: College c prefers enrolling s only to enrolling s' only; c prefers enrolling s' only to leaving positions unfilled; and c prefers leaving positions unfilled to enrolling s'' only. We refer to this preference relation as c's **preferences over individual students**.

A college c's preference relation over groups of students is **responsive** (to its preferences over individual students) if, for any $S', S'' \subset S$, $s' \in S'$, $s'' \in S - S'$ or $s'' = \emptyset$, and $S'' = (S' - s') \cup s''$, $S' P_c S''$ if and only if $s' P_c s''$. If c has responsive preferences, then $Ch_c(S')$ is the set of at most q_c acceptable students that are ranked highest in S' according to c's preferences over individual students.

[2]Also see Alkan and Gale (2003), Echenique and Oviedo (2004), Hatfield and Milgrom (2005), and Ostrovsky (2005) for further discussion on substitutability and existence of stable matchings in more general matching frameworks.

A responsive preference relation is substitutable. Furthermore, when all colleges have responsive preferences, the direct mechanism that is coupled with SA-DAA is DSIC for all students (Theorem 5.16, Roth and Sotomayor 1990). However, this result does not generalize to substitutable preferences (Milgrom 2003). We obtain the following even more negative result:

Theorem 1. *When colleges can have any substitutable preferences, no stable mechanism is DSIC for every student.*

Proof: The proof is via a counterexample. Let $S = \{s_1, s_2, s_3\}$, $C = \{c_1, c_2\}$, $q_{c_1} = 2$ and $q_{c_2} = 1$. Consider the following preference profile P:

$$\begin{array}{llll} P_{s_1}: & c_1 & c_2 & \\ P_{s_2}: & c_2 & c_1 & \text{and} \\ P_{s_3}: & c_2 & c_1 & \end{array} \qquad \begin{array}{llllll} P_{c_1}: & s_3 & \{s_1, s_2\} & s_1 & s_2 & \\ P_{c_2}: & s_1 & s_2 & s_3 & & \end{array}$$

Accordingly, $c_1 P_{s_1} c_2 P_{s_1} s_1$, and $\varnothing P_{c_1} \{s_1, s_3\}$.

Both P_{c_1} and P_{c_2} are substitutable. There is a unique stable matching for P: $\mu(c_1) = s_3$, $\mu(c_2) = s_1$ and $\mu(s_2) = s_2$, i.e. s_2 remains unmatched.

Now let $P' = (P_{-s_2}, P'_{s_2})$ where P'_{s_2} reverses the ranking of colleges, i.e. $c_1 P'_{s_2} c_2 P'_{s_2} s_2$. There are two stable matchings for P' : $\mu'_1 = \mu$; $\mu'_2(c_1) = \{s_1, s_2\}$ and $\mu'_2(c_2) = s_3$. Note that μ'_1 is the college-optimal stable matching, whereas μ'_2 is the student-optimal stable matching.

Next, let $P'' = (P'_{-s_3}, P''_{s_3})$ where P''_{s_3} ranks c_1 as "unacceptable", i.e. $c_2 P''_{s_3} s_3 P''_{s_3} c_1$. There is a unique stable matching $\mu'' = \mu'_2$.

Suppose to the contrary that there is a stable matching mechanism m that is DSIC for every student. If $m(P') = \mu'_1$, then when the true preference profile is P', s_3 is better off by misrepresenting her preferences as P''_{s_3}, since then m picks the unique stable matching μ'' under P'', i.e. $m(P'') = \mu''$, and $\mu''(s_3) = c_2 P'_{s_3} c_1 = \mu'_1(s_3)$, a contradiction. Then $m(P') = \mu'_2$. In this case, when the true preference profile is P, s_2 is better off by misrepresenting her preferences as P'_{s_2}, since m picks the unique stable matching under P, i.e. $m(P) = \mu$, and $\mu'_2(s_2) = c_1 P_{s_2} s_2 = \mu(s_2)$, a contradiction. So no stable matching mechanism is DSIC for every student.[3] ■

Note that even if we assume responsive preferences, no stable mechanism is DSIC for every college (Proposition 2, Roth 1985).

Next, we ask: Is there a non-trivial class of preferences that captures affirmative action constraints, and at the same time yields a stable mechanism that is DSIC for all students? We provide a positive answer to this question in the next section.

2.2. College admissions with affirmative action

In addition to the five items in a college admissions problem, a **college admissions with affirmative action problem** consists of

[3] Under P', students prefer μ'_2 to μ'_1. Milgrom (2003) uses P and P' to show that SA-DAA is not DSIC when colleges can admit any substitutable preferences.

[6] A type space $T = \{\tau_1, ..., \tau_k\}$
[7] A type function $\tau : S \to T$; $\tau(s)$ is the type of student s
[8] For every college c, a vector of type-specific quotas $q_c^T = (q_c^{\tau_1}, ..., q_c^{\tau_k})$ such that $q_c^\tau \leq q_c$ for every c and every τ; and $\sum_{\tau \in T} q_c^\tau \geq q_c$. We refer to these quotas also as affirmative action constraints.

We use type-specific quotas to capture colleges' preferences for affirmative action. We interpret q_c^τ as the maximum number of slots that college c would like to allocate to type-τ students. First, let us give some definitions.

A list of students $S' \subset S$ **respects the capacity constraint at college** c if $|S'| \leq q_c$; S' **respects affirmative action constraints at** c if it respects type-specific quotas at c, that is $|\{s \in S' : \tau(s) = \tau\}| \leq q_c^\tau$ for every $\tau \in T$; S' **respects constraints at** c if it respects the capacity and affirmative action constraints at c. We impose the following restriction on preferences to capture affirmative action constraints.

Property AA: (Affirmative Action) For $c \in C$, P_c satisfies AA if for every $S', S'' \subset S$ such that S' respects constraints at c and S'' does not respect constraints at c, $S' P_c S''$.

Since Ø trivially respects constraints, AA implies that if $S' \subset S$ is acceptable for $c \in C$, then S' respects constraints at c. We discuss AA further in the last section. Next we impose responsiveness only on acceptable sets of students.

Property RR: (Restricted Responsiveness) For $c \in C$, P_c satisfies RR if for every $S', S'' \subset S$ such that $S' P_c \varnothing$, $S'' P_c \varnothing$, and $S'' = (S' - s') \cup s''$ for some $s' \in S'$, $s'' \in S - s'$ or $s'' = \varnothing$, we have $S' P_c S''$ if and only if $s' P_c s''$.

If a set does not respect the capacity or affirmative action constraints at c, it is not acceptable for c. Therefore, RR imposes responsiveness only on those sets that respect the capacity and affirmative action constraints. RR reduces to Martínez et al. (2000)'s q_F-responsiveness when the type space is a singleton. The following example shows that RR does *not* imply responsiveness.

Example 1. There are two female students $\{f_1, f_2\}$ and two male students $\{m_1, m_2\}$. A college c has two seats, $q_c = 2$; and it prefers to enroll at most one student of each gender, i.e. $q_c^f = 1$ and $q_c^m = 1$. Otherwise, its preference relation over groups of students is responsive to the following ranking: $f_1 P_c f_2 P_c m_1 P_c m_2 P_c \varnothing$. Then, $\{f_1, m_2\} P_c \{f_1, f_2\}$ but $f_2 P_c m_2$. So, P_c is not responsive although it satisfies RR.

However,

Lemma 1. *AA and RR imply substitutability.*

Proof: Suppose that P_c satisfies AA and RR. Let $S' \subset S$, $s' \in Ch_c(S')$, and $s'' \in S - s'$. Note that $Ch_c(S')$ respects constraints at c, as does $Ch_c(S') - s'$. Also $Ch_c(S') P_c (Ch_c(S') - s')$ by revealed preferences. Therefore $s' P_c \varnothing$ by RR.

Suppose to the contrary that $s' \notin Ch_c(S' - s'')$. Then $s' P_c \varnothing$, AA and RR imply (i) $|\{s \in Ch_c(S' - s'') : \tau(s) = \tau(s')\}| = q_c^{\tau(s')}$, and (ii) for all $s \in Ch_c(S' - s'')$ such that $\tau(s) = \tau(s')$, $s P_c s'$. But $Ch_c(S' - s'') \subset S'$, so that there exists $s \in Ch_c(S' - s'') - Ch_c(S')$ such that $\tau(s) = \tau(s')$. Then we obtain a

contradiction with RR since $s \in S'$, $\tau(s) = \tau(s')$, $sP_c s'$, $s' \in Ch_c(S')$, $(Ch_c(S') - s') \cup s$ respects constraints at c but $s \notin Ch_c(S')$. ■

Roth (1991) obtains a similar result in a many-to-two matching model, where students can be one of two types {male, female} and colleges may specify that they will employ no more than one female student. Then college preferences satisfying this constraint, but otherwise responsive to a simple rank-ordering, are substitutable (Proposition 6, Roth 1991). Lemma 1 generalizes this observation to a many-to-one matching framework when the number of student types may be greater than two. We can deduce the following result:

Proposition 1. *When colleges' preferences satisfy AA and RR, the set of stable matchings is nonempty.*

Substitutability is sufficient for the existence of a stable matching. So Proposition 1 follows directly from Lemma 1. Substitutability is also sufficient for the following result (Theorem 6.8 of Roth and Sotomayor 1990):

Proposition 2. *When colleges' preferences satisfy AA and RR, SA-DAA produces a stable matching that every student finds at least as desirable as any other stable matching.*

Now we present our first main result:

Theorem 2. *When colleges' preferences satisfy AA and RR, SA-DAA is DSIC for every student.*

Theorem 2 provides a positive result, which should be compared to Milgrom's negative result and our even stronger result in Theorem 1. It is also the first one to extend DSIC for all students beyond the class of responsive preferences. Furthermore, the DSIC result of Dubins and Freedman (1981) and Roth (1982) is a corollary of our theorem on a singleton type space.

We defer the proof to the appendix. However a short discussion follows here. Dubins and Friedman (1981) and Roth (1982) obtain DSIC for marriage problems (one-to-one matching models). When colleges' preferences are responsive, this result can be easily extended to the college admissions problem as follows: Associate with any college admissions problem a marriage problem in which (i) for every college c, there are q_c colleges, $c^1, \ldots, c^{q_c}$, each with capacity one and the same preferences as c; (ii) the set of students is the same, each student prefers c^x to $\tilde{c}^y$ if and only if she prefers c to $\tilde{c}$ in the original problem; she prefers c^n to c^{n+1}; and c^x is acceptable for s if and only if c is acceptable for s in the original problem. Then SA-DAA is DSIC for students in this marriage problem, which is equivalent to the original problem.

In our problem with type-specific quotas, the possibility of $\sum_{\tau \in T} q_c^\tau > q_c$ precludes such well-defined constructions of associated marriage problems when there are more than three types. The following example demonstrates this point:

Example 2. Consider a college c with two seats. Suppose that c prefers to enroll at most one student of each of the following types: White, Black, Asian. We can create at most two smaller colleges, c_1 and c_2, each with capacity one.

Suppose that c_1 is a "White" college, i.e. it enrolls white students only; and c_2 is a "Black" college. Then Asian students cannot be enrolled in c_1 and c_2 in the associated marriage problem. If c prefers Asian students to other type of students in the original problem, then the matching produced in the associated marriage problem will not be stable in the original problem. Therefore, we have to assign one of c_1 and c_2 to at least two types. The following cases are possible.

Case 1: c_1 is a "White" college. Then c_2 must be a "Black or Asian" college; it enrolls Black or Asian students. Then, the final allocation will satisfy the type-specific constraints, since no more than one student of each type can be enrolled in these colleges. However, stability is not assured, because c may prefer all Black or Asian students to all White students in the original problem, yet a white student is assigned to c via c_1 in the associated marriage problem.

Case 2: Each of c_1 and c_2 is assigned to at least two types. For example, c_1 is a "White or Asian" college, and c_2 is a "Black or Asian" college. Then, one type (in this example Asian) will appear in both c_1 and c_2. Consequently, since more than one Asian student may be matched with c (one with c_1 and one with c_2), the match may not respect type-specific quotas.

This simple example can be generalized by extending the type space and/or increasing the capacity and type-specific quotas at c. We provide a direct proof of our DSIC result in the appendix by extending Roth (1982).

3. Applications

3.1. Entry-level medical labor markets

A doctor in the UK can become eligible for full registration with the General Medical Council only if that doctor completes 12 months in a preregistration position, typically six months in a medical position and six months in a surgical position. In the Edinburgh case, surgeons may specify that they will employ no more than one female house officer in any six month period (Roth 1991, 2002). When every doctor applies for single positions and every hospital's preferences satisfy AA for the type set {male, female}, and are otherwise responsive to a linear ranking of doctors, i.e. satisfy RR, our results apply directly.[4]

Roth and Peranson (1999) observe in the American resident matching market[5] that "the director of a second-year postgraduate residency program arranges with the director of a prerequisite first-year program that his

[4]Note that the Edinburgh case is a many-to-two matching problem in which a doctor applies for two positions.

[5]For recent theoretical advances, see Ehlers (2004), Klaus and Klijn (2005).

residents will spend their first year in that prerequisite program. However if the second-year program then fails to match as many residents as were anticipated, this leaves vacancies in the first-year program that can be filled by other applicants.'' Our model and results directly apply to this problem: Let p be such a first-year program with total capacity q_p. Set the space of student types as $T = \{\text{first-year, second-year}\}$. Suppose that the second-year program anticipates $q_2 \leq q_p$ residents to be matched. Divide p into two smaller programs, p_1 and p_2 with capacities $q_{p_1} = q_p - q_2$ and $q_{p_2} = q_2$. Set the type-specific capacity constraints at p_1 as follows: $q_{p_1}^{first-year} = q_{p_1}$ and $q_{p_1}^{second-year} = 0$, so that p_1 admits only first-year students. Let p_1's preferences over first-year students coincide with p's preferences over first-year students. For p_2, $q_{p_2}^{first-year} = q_{p_2}^{second-year} = q_{p_2}$. Also, p_2's preferences are obtained by elevating second-year students in p's preferences; otherwise p_2's preferences over k^{th}-year students coincide with p's preferences over k^{th}-year students, $k = 1, 2$.

3.2. Controlled choice in public schools

Abdulkadiroğlu and Sönmez (2003) introduce a new class of problems, namely controlled choice problems, in the context of public school choice. A **controlled choice problem** consists of (1) a finite set of students $S = \{s_1,...,s_n\}$, (2) a finite set of schools $C = \{c_1,...,c_m\}$, (3) school capacities $q = (q_{c_1},...,q_{c_m})$, (4) for every student $s \in S$, a strict preference relation P_s over $C \cup \{s\}$, (5) for every school $c \in C$, a strict priority ranking P_c of students in S, (6) a type space $T = \{\tau_1,...,\tau_k\}$, (7) a type function $\tau : S \to T$, (8) for every school c, a vector of type-specific quotas $q_c^T = (q_c^{\tau_1},...,q_c^{\tau_k})$ such that $q_c^\tau \leq q_c$ for every c and every τ; and $\sum_{\tau \in T} q_c^\tau \geq q_c$. We refer to these quotas also as **controlled choice constraints.**

Controlled choice attempts to provide choice to students while maintaining racial and ethnic balance at schools. In some states, choice is limited by court-ordered desegregation guidelines, whereas such guidelines are adopted voluntarily in some other schools districts. Controlled choice constraints capture the restrictions imposed by such desegregation guidelines.

A controlled choice problem is essentially a college admissions with affirmative action problem with one distinction: Priorities may not represent school preferences in a controlled choice problem. For example, they may be determined by proximity to a school. One of our main contributions in this paper is a *fairness* notion that associates a controlled choice problem with a problem of college admissions with affirmative action.

Definition. A matching μ is **fair** in a controlled choice problem if

i. the list of students at every school respects capacity and controlled choice constraints under μ; and for every $s \in S$, $\mu(s)$ is acceptable for s;
ii. there do not exist students $s, s' \in S$ and a school $c \in C$ such that $cP_s\mu(s)$ and
[a] $\mu(c) \cup s$ respects capacity and controlled choice constraints at c; or
[b] $\mu(s') = c$, $(\mu(c) - s') \cup s$ respects the controlled choice constraints at c, and s has a higher priority at c than s', i.e. sP_cs'

We say that in case (a) s **exhibits justified envy at** c, and in case (b) s **exhibits justified envy for** s' **at** c.

Fairness or no-envy is a well-studied property in the literature (see for example Tadenuma and Thomson (1991), also see Young (1995) and Thomson (forthcoming) for an excellent survey). Here, s's envy at c is justified only when either (a) s can be placed at c without violating the capacity and controlled choice constraints at c or (b) s can be placed at c by removing a lower priority student s' at c, where the resulting matching does not violate controlled choice constraints at c.

If for some $c \in C$, $\mu(c)$ violates c's capacity or controlled choice constraints, then μ is not fair. Therefore, (i) embodies the property AA. Also, (ii) is equivalent to saying that for every $c \in C$, c's preferences are responsive over sets of students that respect its capacity and controlled choice constraints. In other words, (ii) embodies the property RR. So, we can associate with each controlled choice problem a college admissions with affirmative action problem in which (1) student priorities at schools reflect school preferences over individual students, (2) controlled choice constraints reflect affirmative action constraints, and (3) school preferences satisfy AA and RR. Then the following result follows immediately:

Theorem 3. *A matching in a controlled choice problem is fair if and only if it is stable in the associated college admissions with affirmative action problem.*

The proof follows from a simple comparison of the definitions of stability and fairness. When colleges' preferences are responsive, Balinski and Sönmez (1999) observe that fairness in the Turkish college admissions, which is a one-sided matching problem, is equivalent to stability in a corresponding two-sided matching problem. Their result is a corollary to Theorem 3 with singleton type space. Furthermore, Theorem 3 yields the following mechanism for the controlled choice problem: Run SA-DAA in the associated college admissions with affirmative action problem. Our previous results imply the following:

Theorem 4: *In the class of controlled choice problems, SA-DAA produces a fair matching that every student finds at least as desirable as any other fair matching. Furthermore, it is DSIC for all students.*

4. Further discussion

We derive our formulation of preferences for affirmative action from certain applications. However, one may imagine more complex environments. For example, in a recent work Hatfield and Milgrom (2005) provide a class of preferences that allow for overlapping affirmative action constraints. In their formulation, a hospital can fill a minority slot or a female slot (but not both) by hiring a female minority doctor.

There are at least two potential improvements for our formalization of preferences, each of which deserves attention and is left for future research.

First, we assume that the type of each student is a one-dimensional variable. However, a college might have preferences for affirmative action along

various dimensions. For example, a college might prefer a class that is racially balanced as well as balanced in terms of gender. The following example shows that multi-dimensional type spaces introduce complementarity among students.

Example 3. Let $S = \{bm, bf, wm, wf\}$. A student xy is of race "x" and gender "y". College c has a capacity of two and prefers to enroll at most one student of each race and at most one student of each gender. Its preferences are given by

$$P_c: \quad \{bm, wf\} \quad \{bf, wm\} \quad bm \quad wf \quad bf \quad wm$$

Note that $Ch_c(S' = \{bf, wm, wf\}) = \{bf, wm\}$, whereas $bf \notin Ch_c(S' - wm) = wf$. So, c's preferences are not substitutable. As opposed to Hatfield and Milgrom (2005), a female minority student here fills both a minority slot and a female slot.

Since we lose substitutability, the existence of a stable matching in the case of a multi-dimensional type space does not follow from previous results. However, substitutability is not a necessary condition. We leave the identification of conditions that guarantee existence with multi-dimensional type spaces for future research.

Second, imposing type-specific quotas alone does not guarantee desegregation unless these quotas are chosen appropriately at each school by the district authority. Consider a school with racial quotas of 75 for majority students, and 55 for minority students. Enrolling 75 majority students and no minority students would not violate the fairness notion above, although the resulting class is fully segregated. We plan to study the consequences of minimum quotas in future as well.

Appendix

A Proof of Theorem 2

We extend the proof of Roth (1982) to our general model. First we introduce more notation. A matching μ **respects constraints** if for every $c \in C$, $\mu(c)$ respects constraints at c. For any $S' \subset S$ and $\tau \in T$, let $\#_\tau(S') = |\{s \in S' : \tau(s) = \tau\}|$ be the number of type-τ students in S'. **The quota for type τ at college c is met under** μ if $\#_\tau(\mu(c)) = q_c^\tau$.

By AA and RR, a matching μ is stable if (i) for every $s \in S$, $c \in C$, $\mu(s) R_s s$ and $\mu(c) = Ch_c(\mu(c))$; (ii) there do not exist $s, s' \in S$, $c \in C$ such that $\mu(s') = c$, $cP_s\mu(s)$, $(\mu(c) - s') \cup s$ respects constraints at c, and $sP_c s'$. In this case, we say that (s, c) blocks μ.

Under AA and RR, $Ch_c(S')$ can be found as follows: Order students in S' according to P_c. Include one acceptable student in $Ch_c(S')$ at a time in this order if the resulting set does not exceed the capacity of c and respects type specific quotas. If the capacity is met at c, no more students are included in $Ch_c(S')$. If the type-specific quota for type τ is met at c, no more students of type-τ are included in $Ch_c(S')$.

For $P = (P^S, P^C)$, $DA^S(P)$ denotes the matching produced by SA-DAA for P. Given P, let $\mu = DA^S(P)$ and $\mu_k(c)$ be the set of students that are tentatively matched with c at the end of step k of $DA^S(P)$. We say that s **is rejected by** c **in favor of** s'**at step** k **of** $DA^S(P)$ if (i) either $s \in \mu_{k-1}(c)$ or s applies to c at step k, (ii) $s' \in \mu_k(c)$ and $s \notin \mu_k(c)$, (iii) $(\mu_k(c) - s') \cup s$ respects constraints at c and (iv) $sP_c\varnothing$. By the definition of $DA^S(P)$, we have $s'P_cs$ for any such s'.

Given P, define the available set of colleges for s as follows: $A(s;P) = \{c \in C \cup s : \exists \tilde{\mu},\ \tilde{\mu} \text{ is stable for } P,\ \tilde{\mu}(s) = c\}$. Let $c(s;P)$ denote the best alternative in $A(s;P)$ with respect to P_s.

Lemma 2. *For every $s \in S$, $DA^S(P)$ matches s with $c(s;P)$.*

Proof: Since s is acceptable for herself for every $s \in S$, no $s \in S$ rejects herself in $DA^S(P)$. Suppose that no $s \in S$ is rejected by any $c \in A(s;P)$ at steps $1, ..., k-1$ of $DA^S(P)$. Suppose to the contrary that some $s \in S$ is rejected by some $c \in A(s;P)$ at step k of $DA^S(P)$. Define $S' = \{s' \in \mu_k(c) : \tau(s') = \tau(s)\}$ and $S'' = \mu_k(c) - S' = \{s'' \in \mu_k(c) : \tau(s'') \neq \tau(s)\}$. Note that $s'P_cs$ for every $s' \in S'$.

Since $c \in A(s;P)$, there exists a stable matching $\tilde{\mu}$ such that $\tilde{\mu}(s) = c$. Stability of $\tilde{\mu}$ implies that, for every $s' \in S'$, either $\tilde{\mu}(s')P_{s'}c$ or $\tilde{\mu}(s') = c$. If $\tilde{\mu}(s')P_{s'}c$, then s' is rejected by $\tilde{\mu}(s') \in A(s';P)$ before step k in $DA^S(P)$, a contradiction. So $\tilde{\mu}(s') = c$ for all $s' \in S'$. Therefore, $\#_{\tau(s)}(\mu_k(c)) \leq \#_{\tau(s)}(\tilde{\mu}(c))$.

If $\#_{\tau(s)}(\mu_k(c)) = q_c^{\tau(s)}$, then $\#_{\tau(s)}(\tilde{\mu}(c)) = q_c^{\tau(s)}$, so that there exists $s' \in S'$ such that $\mu_k(s') = c$ and $\tilde{\mu}(s') \neq c$. Since $s'P_cs$, stability of $\tilde{\mu}$ implies $\tilde{\mu}(s')P_{s'}c$. But then s' is rejected by $\tilde{\mu}(s') \in A(s';P)$ before step k in $DA^S(P)$, a contradiction. So (i) $\#_{\tau(s)}(\mu_k(c)) < q_c^{\tau(s)}$. Then (i) implies that (ii) $|\mu_k(c)| = q_c$, otherwise s would not be rejected, and (iii) $s''P_cs$ for every $s'' \in S''$, otherwise, since the quota limit for type $\tau(s)$ is not met at c, s would not be rejected.

For any $\tau'' \neq \tau(s)$, if $\tilde{\mu}(s'') = c$ for every $s'' \in S''$ such that $\tau(s'') = \tau''$, then $\#_{\tau''}(\mu_k(c)) \leq \#_{\tau''}(\tilde{\mu}(c))$ follows immediately. Now, suppose that $\tilde{\mu}(s'') \neq c$ for some $s'' \in S''$. If $\tilde{\mu}(s'')P_{s''}c$, then s'' is rejected by $\tilde{\mu}(s'') \in A(s'';P)$ before step k in $DA^S(P)$, a contradiction. So $\tilde{\mu}(s'') \neq c$ implies $cP_{s''}\tilde{\mu}(s'')$. Then stability of $\tilde{\mu}$ and $s''P_cs$ imply that $\#_{\tau(s'')}(\tilde{\mu}(c)) = q_c^{\tau(s'')}$. This implies $\#_{\tau(s'')}(\mu_k(c)) \leq \#_{\tau(s'')}(\tilde{\mu}(c))$. So, (iv) $\#_{\tau''}(\mu_k(c)) \leq \#_{\tau''}(\tilde{\mu}(c))$ for all $\tau'' \neq \tau(s)$.

Then (i), (ii) and (iv) imply $|\tilde{\mu}(c)| > |\mu_k(c)| = q_c$, a contradiction. ■

Let P_{-i} denote the preference relations of all agents except that of agent $i \in S \cup C$. Fix some $s \in S$. Let $\mu = DA^S(P_s, P_{-s})$ and $\mu' = DA^S(P'_s, P_{-s})$. Let Q_s be such that $\mu'(s)Q_sc$ for all $c \neq \mu'(s)$. We refer to Q_s as **a simple misrepresentation for** P'_s. Let $v = DA^S(Q_s, P_{-s})$.

Lemma 3. *If Q_s is a simple misrepresentation for P'_s, then $\mu'(s) = v(s)$.*

Proof: Let $P = (P_s, P_{-s})$, $P' = (P'_s, P_{-s})$ and $P'' = (Q_s, P_{-s})$. Suppose that μ' is not stable under P''. Then we obtain one of the following contradictions with the stability of μ' under P': (i) If there exists $\hat{s} \in S$ such that $\hat{s}P''_{\hat{s}}\mu'(\hat{s})$, then $\hat{s}P'_{\hat{s}}\mu'(\hat{s})$, a contradiction; (ii) if there exists $\hat{s} \in S$ and $c \in C$ such that $cP''_{\hat{s}}\mu'(\hat{s})$ and $\hat{s} \in Ch_c(\mu'(c) \cup \hat{s})$, then $cP'_{\hat{s}}\mu'(\hat{s})$ and $\hat{s} \in Ch_c(\mu'(c) \cup \hat{s})$, a contradiction. So μ' is stable under P''.

Then ($\mu'(s)Q_sc$ for all $c \neq \mu'(s)$) and (μ' is stable under P'') imply $c(s;P'') = \mu'(s)$. Then we obtain $\mu'(s) = v(s)$ by Lemma 2. ■

In other words, for any misrepresentation P'_s, there is a simple misrepresentation Q_s that works as well as P'_s. So, it suffices to check simple misrepresentations only. Then,

Lemma 4. *If Q_s is a simple misrepresentation such that $v(s)R_s\mu(s)$, then for every $s' \in S$, $v(s')R_{s'}\mu(s')$.*

Proof: Suppose to the contrary that for some s', $\mu(s')P_{s'}v(s')$. Since $s' \neq s$, s' states the same preferences, so that s' applies to and is rejected by $\mu(s')$ at some step of $DA^S(Q_s, P_{-s})$. Let k be the first step of $DA^S(Q_s, P_{-s})$ at which some student, say s', is rejected by $\mu(s')$.

Define $S'' = \{ s'' \in S : s'' \text{ does not apply to } \mu(s') \text{ in } DA^S(P_s, P_{-s}) \}$. Then $\mu(s'')P_{s''}\mu(s')$ for every $s'' \in S''$. Also, there exists at least one $s'' \in S''$ such that s'' applies to $\mu(s')$ at step k of $DA^S(Q_s, P_{-s})$ and s' is rejected by $\mu(s')$ in favor of s'' at step k. But then s'' is rejected by $\mu(s'')$ before step k in $DA^S(P)$, a contradiction. ■

The next lemma is an addition to the steps of Roth's argument. This lemma holds trivially in Roth's model.

Lemma 5. *Let Q_s be a simple misrepresentation such that $v(s)R_s\mu(s)$. Then, $|v(c)| = |\mu(c)|$ for every $c \in C$.*

Proof: By Lemma 4, if a student does not apply to a college c in $DA^S(P_s, P_{-s})$, she does not apply to c in $DA^S(Q_s, P_{-s})$ either. That is, every $c \in C$ receives at least as many applications in $DA^S(P_s, P_{-s})$ as it does in $DA^S(Q_s, P_{-s})$. So, $|v(c)| \leq |\mu(c)|$ for every $c \in C$. Again, by Lemma 4, the number of unmatched students under v is less than or equal to the number of unmatched students under μ. This implies $\sum_{c \in C} |v(c)| \geq \sum_{c \in C} |\mu(c)|$. So, for every $c \in C$, $|v(c)| = |\mu(c)|$. ■

Proof of Theorem 2. Given a preference profile P, suppose that $s \in S$ is not acceptable for $c \in C$. Obtain P'_s from P_s by deleting c from P_s, i.e. by making c unacceptable for s. Then $DA^S(P_s, P_{-s}) = DA^S(P'_s, P_{-s})$, so that we can assume without loss of generality that a college c is acceptable for a student s only if s is acceptable for c under P. Consequently, a student that is not acceptable for a college c does not apply to c in DA^S.

We do not need to check unsuccessful misrepresentations. Fix $s \in S$. Let Q_s be a simple misrepresentation. Suppose that either $v(s)P_s\mu(s)$ or $v(s) = \mu(s)$. We will show that $v(s)P_s\mu(s)$ is not possible.

For any s', we say that s' **makes a match** (with $\mu(s')$) **at step** k of $DA^S(P_s, P_{-s})$ if s' applies to $\mu(s')$ at step k.

Let t be the final step of $DA^S(P_s, P_{-s})$. Consider a student s' who makes a match with $c = \mu(s')$ at step t. We will show that $\mu(s') = v(s')$.

First, note the following: (i) No student is rejected by c at step t. Otherwise, the rejected student would be matched after step t, so that t would not be the final step of $DA^S(P_s, P_{-s})$, a contradiction. (ii) There is always an empty slot at c before step t. Otherwise, in order to match s' to c, some other student would be rejected at step t, a contradiction with (i). By (ii) and a similar logic, (iii) the quota limit for type $-\tau(s')$ is not met at any step $r < t$. Then by (ii) and

(iii), (iv) no type $\tau(s')$ student is rejected by c in $DA^S(P_s, P_{-s})$. Therefore, by (ii) and (iv), (v) if a student s'' is rejected by c at some step in $DA^S(P_s, P_{-s})$, then $\tau(s'') \neq \tau(s')$ and the quota limit for type $\tau(s'')$ is met at c at the time of the rejection of s''.

Now, suppose to the contrary that $c \neq v(s')$. By Lemma 4, $v(s')P_{s'}c$ so that s' does not apply to c in $DA^S(Q_s, P_{-s})$. By Lemma 5, $|v(c)| = |\mu(c)|$. So there exists some $s'' \in S - s'$ such that $\mu(s'') \neq c$ and $v(s'') = c$. Then by Lemma 4, $cP_{s''}\mu(s'')$. If $\tau(s'') = \tau(s')$, then by (iv), s'' would apply to and not be rejected by c in $DA^S(P_s, P_{-s})$, a contradiction. So, $\tau(s'') \neq \tau(s')$. Also, $cP_{s''}\mu(s'')$ and $\mu(s'') \neq c$ imply that s'' is rejected by c at some step $r < t$ of $DA^S(P_s, P_{-s})$. Then by (v), s'' is rejected by c in $DA^S(P_s, P_{-s})$ because of the quota limit for type $\tau(s'')$. Therefore, in $v = DA^S(Q_s, P_{-s})$, s'' fills in the slot of another $\hat{s}$ at c such that $\tau(\hat{s}) = \tau(s'') \neq \tau(s')$, $c = \mu(\hat{s}) \neq v(\hat{s})$. So, the slot that is not filled by s' at c under v remains unfilled, which implies $|v(c)| < |\mu(c)|$, a contradiction. So $\mu(s') = v(s')$ for every s' who makes a match at step t.

The rest of the proof is by induction: For $r < t$, suppose that $\mu(s') = v(s')$ for any s' who makes her match at step $r+1$ or at a later step of $DA^S(P_s, P_{-s})$. We have just showed that this is true for $r = t - 1$. We will show that $\mu(s') = v(s')$ for any s' who makes her match at step r as well.

Now, suppose to the contrary that s' makes her match at step r of $DA^S(P_s, P_{-s})$ and $c = \mu(s') \neq v(s')$. By Lemma 4, $v(s')P_{s'}c$, so that s' does not apply to c in $DA^S(Q_s, P_{-s})$. Define $S'' = \{\, s'' : v(s'') = c \neq \mu(s'') \,\}$. By Lemma 5, $|v(c)| = |\mu(c)|$. So $S'' \neq \varnothing$. Then by Lemma 4, $cP_{s''}\mu(s'')$ for every $s'' \in S''$. So there exists $\hat{s} \in S''$ who is rejected by c in favor of s' at step r of $DA^S(P_s, P_{-s})$. Since s' makes her match with c at step r of $DA^S(P_s, P_{-s})$, $\hat{s}$ makes her match at a later step $r' > r$ of $DA^S(P_s, P_{-s})$. Then by the induction hypothesis, $\mu(\hat{s}) = v(\hat{s})$, a contradiction with $\hat{s} \in S''$. Therefore, $\mu(s') = v(s')$.

Then, the induction on r proves that $\mu(s') = v(s')$ for every $s' \in S$, in particular for s. Thus, s cannot successfully manipulate DA^S by misrepresenting her preferences. ■

References

Abdulkadiroğlu A, Pathak P, Roth AE (2005) The New York city high school match. American Economic Review Papers and Proceedings (forthcoming)

Abdulkadiroğlu A, Sönmez T (2003) School choice: a mechanism design approach. American Econ. Rev. 93(3):729–747

Alkan A, Gale D (2003) Stable schedule matching under revealed preferences. J. Econ. Theory 112:289–306

Balinski M, Sönmez T (1999) A tale of two mechanisms: student placement. J. Econ. Theory 84(1):73–94

Dubins LE, Freedman DA (1981) Machiavelli and the Gale-Shapley algorithm. American Math. Monthly 88:485–494

Echenique F, Oviedo J (2004) Core many-to-one matchings by fixed-point methods. J. Econ. Theory 115(2):358–376

Ehlers L (2004) In search of advice for participants in matching markets which use the Deferred-Acceptance algorithm. Games Econ. Behav. 48:249–270

Hatfield J, Milgrom P (2005) Auctions, matching and the law of auctions matching and the law of aggregate demand. American Econ. Rev. (forthcoming)

Hirch D School: a Matter of Choice. Paris: Publication Service, OECD, 1994

Kelso A, S., Jr. Crawford VP (1982) Job matching, coalition formation, and gross substitutes. Econometrica 50:1483–1504

Klaus B, Klijn F (2005) Stable matchings and preferences of couples. J. Econ. Theory 121:75–106
Martínez R, Jordi M, Alejandro N, Oviedo J (2000) Single agents and the set of many-to-one stable matchings. J. Econ. Theory 91:91–105
Milgrom P (2003) Matching with contracts, mimeo
Ostrovsky M (2005) Stability in supply chain networks,mimeo
Roth AE (1982) The economics of matching: stability and incentives. Math. Oper. Res. 7:617–628
Roth AE (1985) The college admissions problem is not equivalent to the marriage problem. J. Econ. Theory 36:277–288
Roth AE (1991) A natural experiment in the organization of entry-level labor markets: regional markets for new physicians and surgeons in the United Kingdom. American Econ. Rev. 81(3):414–440
Roth AE (2002) The economist as an engineer: game theory, experimentation, and computation as tools for design. Econometrica 70(4):1341–1378
Roth AE, Peranson E (1999) The redesign of the matching market for american physicians: some engineering aspects of economic design. American Econ. Rev. 89(4):748–780
Roth AE, Sotomayor MAO (1990) Two-sided matching: a study in game theoretic modeling and analysis. New York: Cambridge University Press
Tadenuma K, Thomson W (1991) No-envy and consistency in economies with indivisible goods. Econometrica 59(6):1755–1767
Thomson W The Theory of fair allocation. Princeton University Press, (forthcoming)
Young P (1995) Equity: In theory and practice. Princeton University Press