SCHOOL CHOICE: IMPOSSIBILITIES FOR AFFIRMATIVE ACTION

FUHITO KOJIMA

Abstract. This paper investigates the welfare effects of affirmative action policies in school choice. We show that affirmative action policies can have perverse consequences. Specifically, we demonstrate that there are market situations in which affirmative action policies inevitably hurt every minority student – the purported beneficiaries – under any stable matching mechanism, such as those used in New York City and Boston. Furthermore, we show that another famous mechanism, the top trading cycles mechanism, suffers from the same drawback. JEL Classification Numbers: C70, D61.
Keywords: Matching, Stability, School Choice, Affirmative Action, Deferred Acceptance, Top Trading Cycles.

1. Introduction

Racial desegregation is a long-standing problem in American society as well as many parts of the world. In the United States, affirmative action policies have been playing an important role in achieving this goal. Individuals in minority groups often receive preferential treatment in employment and school admission decisions.

Affirmative action policies have been widely used in public education although they have also received various criticisms.[1] Although recent Court decisions have prohibited the explicit use of "racial tie breakers" in public school admission to achieve racial integration,[2] indirect, or "color-blind," affirmative action policies are still regularly employed.

We study affirmative action policies in the context of the school choice problem as analyzed by Abdulkadiroğlu and Sönmez (2003) and Abdulkadiroğlu (2005).[3] Specifically,

Date: October 12, 2010.

Department of Economics, Stanford University. email: fuhitokojima1979@gmail.com. I am grateful to Toshiji Kawagoe, Al Roth, Tayfun Sönmez, Satoru Takahashi, Yosuke Yasuda, and especially Yusuke Narita for discussion and comments. I also thank Pete Troyan for his research assistance.

[1]Criticisms include, among others, the claim that affirmative action is a reverse discrimination and doubts about its effectiveness as desegregation measure.

[2]See *Parents Involved in Community Schools v. Seattle School District No. 1* and *Meredith v. Jefferson County Board of Education*.

[3]See Roth (1991) for a related model in the context of British labor market for doctors, and Ergin and Sönmez (2006) who study the Boston mechanism in school choice. See Roth and Sotomayor (1990), Roth (2007), and Sönmez and Ünver (2009) for more general surveys of the literature.

we investigate the consequences of adopting affirmative action policies on student welfare. Our main finding is that an affirmative action policy can have a perverse effect on student welfare, including minority students, who are the purported beneficiaries. More specifically, we establish impossibility theorems stating that there are situations where *affirmative action policies inevitably hurt every minority student* under any stable matching mechanism, such as those used in New York City and Boston. Moreover, this misfortune is unavoidable under alternative ways to implement the affirmative action policy: The impossibility results hold both when certain school seats are reserved for minority students (by imposing type-specific quotas on majority students) and when minority students are given preferential treatment in receiving priority in schools. Furthermore, similar impossibility results hold under another popular mechanism, the top trading cycles mechanism. These findings suggest that caution should be exercised in the use of affirmative action policies even if helping minority students is deemed desirable by society.

On the other hand, we also find that imposing affirmative action does not necessarily hurt majority students. In fact, there are cases in which a stronger affirmative action policy makes everyone better off, *including every majority student*. This observation appears to be contrary to popular beliefs and suggests that caution may be needed when assessing the cost of affirmative action policies as well as its benefit.

The analytical approach of this paper follows the tradition of impossibility studies in the matching literature. Roth (1982) shows that there exists no stable and strategy-proof mechanism. Sönmez (1997, 1999) studies two forms of manipulations – manipulations via capacities and via pre-arranged matches, respectively – and shows that immunity to either of these manipulations is incompatible with stability. While the analysis of the current paper is new to our knowledge, the formulations of our impossibility theorems are inspired by these studies.

The rest of the paper proceeds as follows. Section 2 sets up the model. Section 3 presents our results for stable mechanisms. Section 4 establishes results for the top trading cycles mechanism. Section 5 concludes.

2. Model

A market is tuple $G = (S, C, (\succeq_i)_{i\in S\cup C}, (\mathbf{q}_c)_{c\in C})$. S and C are finite and disjoint sets of students and schools. For each student $s \in S$, $\succeq_s$ is a preference relation over C and being unmatched (being unmatched is denoted by $\emptyset$). We assume that preferences are strict. We write $c' \succ_s c$ if and only if $c' \succeq_s c$ but not $c \succeq_s c'$. For each school c, $\succeq_c$ is a priority order over the set of students. We assume that priority orders are strict. We write $s' \succ_c s$

if and only if $s' \succeq_c s$ but not $s \succ_c s'$. If $c \succ_s \emptyset$, then c is said to be **acceptable** to s. The set of students are partitioned to two subsets; the set S^M of **majority students** and S^m of **minority students**.[4] For each $c \in C$, $\mathbf{q}_c = (q_c, q_c^M)$ is the **capacity** of c: The first component q_c represents the total capacity of school c, while the second component q_c^M represents the type-specific capacity for majority students.

A **matching** μ is a mapping from $C \cup S$ to $C \cup S \cup \{\emptyset\}$ such that

(1) $\mu(s) \in C \cup \{\emptyset\}$,
(2) For any $s \in S$ and $c \in C$, $\mu(s) = c$ if and only if $s \in \mu(c)$.
(3) $\mu(c) \subseteq S$ and $|\mu(c)| \leq q_c$ for all $c \in C$,
(4) $|\mu(c) \cap S^M| \leq q_c^M$ for all $c \in C$.

All conditions except for (4) are standard in the literature. Condition (4) requires that the number of majority students matched to each school c is at most its type-specific capacity q_c^M.

A matching μ is **stable** if

(1) $\mu(s) \succeq_s \emptyset$ for each $s \in S$, and
(2) if $c \succ_s \mu(s)$, then either
    (a) $|\mu(c)| = q_c$ and $s' \succ_c s$ for all $s' \in \mu(c)$, or
    (b) $s \in S^M$, $|\mu(c) \cap S^M| = q_c^M$, and $s' \succ_c s$ for all $s' \in \mu(c) \cap S^M$.

All conditions except for (2b) are standard. Condition (2b) describes a case in which a potential blocking is not realized because of a type-specific capacity constraint for the majority students: Student s wants to be matched with school c, but she is a majority student and the seats for majority students are filled by students who have higher priority than s at c.

A **mechanism** is a function ϕ that, for each market G, associates a matching $\phi(G)$. A mechanism ϕ is **stable** if $\phi(G)$ is a stable matching in G for any given G.

This model is a special case of the "controlled school choice" analyzed by Abdulkadiroğlu and Sönmez (2003). They present a more general model, where there are an arbitrary finite number of student types and there is a type specific capacity for each type of students. In this paper we focus on a very simple situation in which there are only two types of students S^M and S^m. We chose our simple model for expositional simplicity only, and this modeling choice is without loss of generality: Since our results are impossibility theorems, all claims that hold in our simple environment hold in their more general model.

[4]Although we use words such as "majority" and "minority," we do not necessarily assume that there are more majority students than minority students. Such an assumption changes none of the results of this paper.

Consider the following **deferred acceptance algorithm** (Gale and Shapley (1962); adapted to controlled school choice by Abdulkadiroğlu and Sönmez (2003)):

- Step 1: Start with a matching in which no student is matched. Each student s applies to her first choice school (call it c). The school c rejects s if (i) q_c seats are filled by students who have higher priority than s at c or (ii) $s \in S^M$ and q_c^M seats are filled by students in S^M who have higher priority than s at c. Each school c keeps all other students who applied to c.
- Step t: Start with the tentative matching obtained at the end of Step $t-1$. Each student s applies to her first choice school (call it c) among all schools that have not rejected s before. The school c rejects s if (i) q_c seats are filled by students who have higher priority than s at c or (ii) $s \in S^M$ and q_c^M seats are filled by students in S^M who have higher priority than s at c. Each school c keeps all other students who applied to c.

The algorithm terminates at a step in which no rejection occurs, and the tentative matching at that step is finalized. Since no student applies to a school that has rejected her again and at least one rejection occurs in each step as long as the algorithm does not terminate, the algorithm stops in a finite number of steps. Modifying the argument by Gale and Shapley (1962), Abdulkadiroğlu and Sönmez (2003) show that the outcome of the deferred acceptance algorithm is the student-optimal stable matching, a stable matching that is unanimously most preferred by all students among all stable matchings.

3. Consequences of Affirmative Action Policies under Stable Mechanisms

This section investigates welfare consequences of imposing affirmative action constraints under stable matching mechanisms. The main finding is that in some markets, the affirmative action constraint inevitably hurts minority students who it tries to help. In general, affirmative action policies can have counterintuitive welfare effects. The constraint hurts every student in some environments, while helping every student in others.

Market $\tilde{G} = (S, C, \succeq, \tilde{\boldsymbol{q}})$ is said to **have a stronger affirmative action policy than** $G = (S, C, \succeq, \boldsymbol{q})$ if, for every $c \in C$, $q_c = \tilde{q}_c$ and $q_c^M \geq \tilde{q}_c^M$. The definition requires that the type-specific capacity for the majority be smaller in $\tilde{G}$ than in G, so that a stronger restriction is imposed in the former.

A matching μ' is **Pareto inferior to μ for the minority** if (i) $\mu(s) \succeq_s \mu'(s)$ for every $s \in S^m$ and (i) $\mu(s) \succ_s \mu'(s)$ for at least one $s \in S^m$.

Definition 1. A matching mechanism ϕ is said to **respect the spirit of affirmative action** if there are no markets G and $\tilde{G}$ such that $\tilde{G}$ has a stronger affirmative action policy than G and $\phi(\tilde{G})$ is Pareto inferior to $\phi(G)$ for the minority.

The definition considers a change in market situations from G to $\tilde{G}$ such that a stronger affirmative action policy is introduced in the latter than in the former. Respecting the spirit of affirmative action requires that introducing a stronger affirmative action policy never hurts all minority students.

Note that there is a sense in which respect of the spirit of affirmative action is a weak requirement: It only excludes situations in which *every minority student* is made (weakly) worse off by a stronger affirmative action policy. Thus it allows for a situation where a stronger affirmative action policy hurts some minority students while other minority students are made better off. One reason that we consider such a weak requirement is that our results are impossibility theorems, thus showing an impossibility with respect to a weak requirement immediately implies an impossibility result for a stronger requirement.

Unfortunately, respect of the spirit of affirmative action turns out to be a demanding requirement in school choice. More specifically, the following result states that it is in conflict with stability.

Theorem 1. *There exists no stable mechanism that respects the spirit of affirmative action.*

Proof. Consider the following market $G = (S, C, P, \boldsymbol{q})$. Let $C = \{c_1, c_2\}, S = \{s_1, s_2, s_3\}$, and $S^M = \{s_1, s_2\}, S^m = \{s_3\}$, and

$$\begin{aligned} &\succeq_{c_1}\!: s_1, s_2, s_3 && \boldsymbol{q}_{c_1} = (q_{c_1}, q_{c_1}^M) = (2, 2) \\ &\succeq_{c_2}\!: s_2, s_3, s_1 && \boldsymbol{q}_{c_2} = (1, 1), \end{aligned}$$

where the notational convention here is that students are listed in order of priorities: At school c_1, for instance, student s_1 has the highest priority, s_2 has the second highest priority, and s_3 has the lowest priority. Student preferences are given by

$$\begin{aligned} &\succeq_{s_1}\!: c_1, \\ &\succeq_{s_2}\!: c_1, c_2, \\ &\succeq_{s_3}\!: c_2, c_1, \end{aligned}$$

where the notational convention is that schools are listed in order of preferences and schools not on the preference list is unacceptable.[5] There exists a unique stable matching μ in this market given by[6]

$$\mu(c_1) = \{s_1, s_2\},$$
$$\mu(c_2) = s_3.$$

Consider $\tilde{\boldsymbol{q}} = (\tilde{\boldsymbol{q}}_{c_1}, \boldsymbol{q}_{c_2})$, where $\tilde{\boldsymbol{q}}_{c_1} = (2, 1)$. The market $\tilde{G} = (S, C, P, \tilde{\boldsymbol{q}})$ has a stronger affirmative action policy than $G = (S, C, P, \boldsymbol{q})$. In market $\tilde{G}$, there is a unique stable matching $\tilde{\mu}$ given by

$$\tilde{\mu}(c_1) = \{s_1, s_3\},$$
$$\tilde{\mu}(c_2) = s_2.$$

Student s_3 is strictly worse off under $\tilde{\mu}$ than under μ. Therefore $\tilde{\mu}$ is Pareto inferior to μ for the minority. Since μ and $\tilde{\mu}$ are unique stable matchings of G and $\tilde{G}$, respectively, this completes the proof. □

In the example presented in the proof, it is not only the minority student but also the majority students that are weakly worse off in $\tilde{G}$. In other words, this example shows that a stronger affirmative action constraint can induce a Pareto inferior matching (for all students). Given this example, one might suspect that the result is not surprising: After all, affirmative action is a constraint, and as such, it is always bad for welfare. While this intuition may sound sensible, it is not correct in general. To see this point, we present an example where *the affirmative action constraint benefits everyone, including the majority students.*

Example 1. Let $C = \{c_1, c_2\}, S = \{s_1, s_2, s_3, s_4\}$. $S^M = \{s_1, s_2\}, S^m = \{s_3, s_4\}$.

$$\succeq_{c_1}: s_1, s_4, s_2, s_3 \qquad \boldsymbol{q}_{c_1} = (2, 2),$$
$$\succeq_{c_2}: s_3, s_4, \ldots \qquad \boldsymbol{q}_{c_2} = (1, 1),$$

where the notation $\ldots$ in the end of the priority of c_2 indicates that part of the priority is arbitrary (similar convention is used in the remainder of the paper). Student preferences

[5]The model allows for students to find some schools unacceptable for expositional simplicity, but this feature is not needed for our results: A modification can be made for all our results by introducing an additional school with a sufficient capacity that admit students who are unmatched in the current examples.

[6]We abuse notation and denote a singleton set $\{x\}$ by x whenever there is no concern for confusion.

are given by

$$\succeq_{s_1}: c_1,$$
$$\succeq_{s_2}: c_1,$$
$$\succeq_{s_3}: c_1, c_2,$$
$$\succeq_{s_4}: c_2, c_1.$$

The unique stable matching μ of this market is given by

$$\mu(c_1) = \{s_1, s_4\},$$
$$\mu(c_2) = s_3,$$
$$\mu(s_2) = \emptyset.$$

Suppose that the capacity of c_1 is changed to $\tilde{\boldsymbol{q}}_{c_1} = (2, 1)$, so that the new market has a stronger affirmative action policy than the original one. The student-optimal stable matching $\tilde{\mu}$ of this modified market is given by

$$\tilde{\mu}(c_1) = \{s_1, s_3\},$$
$$\tilde{\mu}(c_2) = s_4,$$
$$\tilde{\mu}(s_2) = \emptyset.$$

Every student is weakly better off under $\tilde{\mu}$ than under μ: Students s_1 and s_2 are indifferent, whereas s_3 and s_4 are strictly better off.

Furthermore, a similar point can be made for an *arbitrary* stable matching mechanism: There exist markets G and $\tilde{G}$ such that $\tilde{G}$ has a stronger affirmative action policy than G and, for any stable mechanism ϕ, matching $\phi(G)$ is Pareto inferior to $\phi(\tilde{G})$ for the minority.

Example 2. Let ϕ be an arbitrary stable mechanism. Consider the following market $G = (S, C, P, \boldsymbol{q})$. $C = \{c\}, S = \{s_1, s_2, s_3\}, S^M = \{s_1, s_2\}, S^m = \{s_3\}$,

$$\succeq_c: s_1, s_2, s_3 \qquad\qquad \boldsymbol{q}_c = (q_c, q_c^M) = (2, 2).$$

The above capacity indicates that effectively there is no affirmative action constraint. Suppose that $c \succ_s \emptyset$ for every $s \in S$. In this market there exists a unique stable matching μ given by

$$\mu(c) = \{s_1, s_2\}.$$

Consider $\tilde{\boldsymbol{q}} = \tilde{\boldsymbol{q}}_c$, where $\tilde{\boldsymbol{q}}_c = (2, 1)$. The market $\tilde{G} = (S, C, P, \tilde{\boldsymbol{q}})$ imposes affirmative action on G. There is a unique stable matching, denoted by $\tilde{\mu}$, given by

$$\tilde{\mu}(c) = \{s_1, s_3\}.$$

Clearly, μ is Pareto inferior to $\tilde{\mu}$ for the minority.

3.1. **An alternative affirmative action policy.** So far we have seen that affirmative action policy does not necessarily help the minority when the market is organized under a stable mechanism. A natural question is whether there is an alternative policy measure to attain the goal of affirmative action. To investigate this question, we consider another popular form of affirmative action policy. We say that $\tilde{G} = (S, C, \tilde{\succeq}, \boldsymbol{q})$ **has a stronger priority-based affirmative action policy** than $G = (S, C, \succeq, \boldsymbol{q})$ if, $\tilde{\succeq}_s = \succeq_s$ for all $s \in S$ and, for every $c \in C$ and $s, s' \in S$, $s \succeq_c s'$ and $s \in S^m$ imply $s \tilde{\succeq}_c s'$. This policy is often called preferential treatment and is based on a simple idea: A priority-based affirmative action policy promotes the ranking of a minority student at schools relative to majority students while keeping the relative ranking of each student within her own group fixed.

Definition 2. A matching mechanism ϕ is said to **respect the spirit of priority-based affirmative action** if there are no markets G and $\tilde{G}$ such that $\tilde{G}$ has a stronger priority-based affirmative action policy than G and $\phi(\tilde{G})$ is Pareto inferior to $\phi(G)$ for the minority.

This requirement is similar to and inspired by a condition called the respect of improvements (Balinski and Sönmez, 1999). A mechanism respects improvements if a student is never made worse off when her own ranking at some schools improves while relative rankings among other students are unchanged. The current definition similarly considers a change in the ranking of a student. Our requirement is different from theirs in that a condition is imposed on what happens to a minority student's welfare when another minority student's ranking improves relative to majority students.

Theorem 2. *There exists no stable mechanism that respects the spirit of priority-based affirmative action.*

Proof. Consider the following market, $G = (S, C, P, \boldsymbol{q})$. $C = \{c_1, c_2\}, S = \{s_1, s_2, s_3\}$, $S^M = \{s_1\}$, $S^m = \{s_2, s_3\}$.

$$\begin{aligned} &\succeq_{c_1}: s_1, s_2, s_3 && \boldsymbol{q}_{c_1} = (1, 1) \\ &\succeq_{c_2}: s_2, s_1, s_3 && \boldsymbol{q}_{c_2} = (1, 1). \end{aligned}$$

Student preferences are given by

$$\succeq_{s_1}: c_2, c_1$$
$$\succeq_{s_2}: c_1, c_2$$
$$\succeq_{s_3}: c_2, c_1.$$

In this market G, there are two stable matchings μ and μ' given by

$$\mu(c_1) = s_2,$$
$$\mu(c_2) = s_1$$
$$\mu(s_3) = \emptyset,$$

and

$$\mu'(c_1) = s_1,$$
$$\mu'(c_2) = s_2$$
$$\mu'(s_3) = \emptyset.$$

We consider the following cases.

Case 1. Suppose that $\phi(G) = \mu$. In that case, consider $\tilde{G}$ which changes c_2's priority as follows:

$$\tilde{\succeq}_{c_2} : s_2, s_3, s_1.$$

Market $\tilde{G}$ has a stronger priority-based affirmative action policy than G. In this market $\tilde{G}$, there is a unique stable matching $\tilde{\mu}$ given by

$$\tilde{\mu}(c_1) = s_1,$$
$$\tilde{\mu}(c_2) = s_2,$$
$$\tilde{\mu}(s_3) = \emptyset.$$

Clearly, $\tilde{\mu}$ is Pareto inferior to μ for the minority.

Case 2. Suppose that $\phi(G) = \mu'$. In that case, consider $\tilde{G}'$ which changes c_2's priority as follows:

$$\tilde{\succeq}'_{c_2} : s_1, s_2, s_3.$$

Market G has a stronger priority-based affirmative action policy than $\tilde{G}'$. In market $\tilde{G}'$, there is a unique stable matching $\tilde{\mu}'$ given by

$$\tilde{\mu}'(c_1) = s_2,$$
$$\tilde{\mu}'(c_2) = s_1,$$
$$\tilde{\mu}'(s_3) = \emptyset.$$

Clearly, μ' is Pareto inferior to $\tilde{\mu}'$ for the minority although G has a stronger priority-based affirmative action policy than $\tilde{G}'$, completing the proof.

□

4. The Top Trading Cycles Mechanism

In this section, we analyze an alternative class of matching mechanisms that has attracted attention in recent years. Given school priorities and student preferences, consider the following **top trading cycles mechanism**:

- Step 1: Start with a matching in which no student is matched. For each school c, set its total counter at its total capacity q_c and its majority-specific counter at its type-specific capacity q_c^M. Each school points to a student who has the highest priority at that school. Each student s points to her most preferred school that still has a seat for her, that is, a school whose total counter is strictly positive and, if $s \in S^M$, its majority-specific counter is strictly positive. There exists at least one cycle (if a student points to $\emptyset$, it is regarded as a cycle). Every student in a cycle receives the school she is pointing to and is removed. The counter of each school is reduced by one. If the assigned student is in S^M, then the school matched to that student reduces its majority-specific counter by one. If no student remains, terminate. Otherwise, proceed to the next step.
- Step t: Start with the matching and counter profile reached at the end of Step $t-1$. Each school points to a student who has the highest priority at that school. Each student s points to her most preferred school that still has a seat for her, that is, a school whose total counter is strictly positive and, if $s \in S^M$, its majority-specific counter is strictly positive. There exists at least one cycle (if a student points to $\emptyset$, it is regarded as a cycle). Every student in a cycle receives the school she is pointing to and is removed. The counter of each school is reduced by one. If the assigned student is in S^M, then the school matched to that student reduces its majority-specific counter by one. If no student remains, terminate. Otherwise, proceed to the next step.

This algorithm terminates in a finite number of steps because at least one student is matched at each step as long as the algorithm has not terminated and there are a finite number of students. The top trading cycles matching is the matching reached at the termination of the above algorithm.

The current version of the top trading cycles algorithm was introduced by Abdulkadiroğlu and Sönmez (2003) for the school choice problem.[7] While it does not necessarily produce a stable matching, the mechanism has a number of desirable properties. First, it always produces a Pareto efficient matching. Second, it is group-strategy proof, that is, no coalition of students can jointly misreport preferences in such a way that every student in the coalition is made weakly better off with at least one strictly better off. Based on these advantages, the top trading cycles algorithm has been considered for use in a number of school districts in the United States, such as Boston and San Francisco.

Unfortunately, the next result shows that the top trading cycles mechanism also has a problematic feature regarding affirmative action.

Theorem 3. *The top trading cycles mechanism does not respect the spirit of affirmative action.*

Proof. Let $C = \{c_1, c_2, c_3\}, S = \{s_1, s_2, s_3, s_4\}, S^M = \{s_1, s_2\}, S^m = \{s_3, s_4\}$,

$$\begin{aligned} &\succeq_{c_1}\!: s_1, s_2, s_3, s_4 && \boldsymbol{q}_{c_1} = (2, 2), \\ &\succeq_{c_2}\!: s_2, s_3, \ldots && \boldsymbol{q}_{c_2} = (1, 1), \\ &\succeq_{c_3}\!: s_4, \ldots && \boldsymbol{q}_{c_3} = (1, 1). \end{aligned}$$

Student preferences are given by

$$\begin{aligned} &\succeq_{s_1}\!: c_1, \\ &\succeq_{s_2}\!: c_1, c_3, \\ &\succeq_{s_3}\!: c_3, \\ &\succeq_{s_4}\!: c_2. \end{aligned}$$

In this market, the top trading cycles mechanism results in a matching μ given by

$$\begin{aligned} \mu(c_1) &= \{s_1, s_2\}, \\ \mu(c_2) &= s_4, \\ \mu(c_3) &= s_3. \end{aligned}$$

[7]The original top trading cycles algorithm was defined in the context of the housing market and is attributed to David Gale by Shapley and Scarf (1974).

Now suppose that the capacity of c_1 is given by $\tilde{\boldsymbol{q}}_{c_1} = (2, 1)$, so that the new market has a stronger affirmative action policy than the original market. In this market, the top trading cycles mechanism results in a matching μ' given by

$$\tilde{\mu}(c_1) = s_1,$$
$$\tilde{\mu}(c_2) = s_4,$$
$$\tilde{\mu}(c_3) = s_2,$$
$$\tilde{\mu}(s_3) = \emptyset.$$

Every student is weakly worse off under $\tilde{\mu}$ than under μ: Students s_1 and s_4 are indifferent, whereas s_2 and s_3 are strictly worse off. Note that s_3 is a minority student. □

This result shows that the top trading cycles mechanism does not guarantee that an affirmative action policy has an intended effect to help the minority. Thus the difficulty of affirmative action policies is not confined to stable mechanisms. Another remark is that every student is made weakly worse off by the affirmative action policy in the example used in the proof. Thus it is possible that the policy unambiguously hurts welfare.

One might hope that minority students are not hurt by priority-based affirmative action policies under the top trading cycles mechanism. Unfortunately, the next result shows that this is not the case.

Theorem 4. *The top trading cycles mechanism does not respect the spirit of priority-based affirmative action.*

Proof. Let $C = \{c_1, c_2, c_3\}, S = \{s_1, s_2, s_3, s_4\}$, $S^M = \{s_1, s_2\}$, and $S^m = \{s_3, s_4\}$.

$$\begin{aligned} &\succeq_{c_1}\!: s_4, s_2, \dots && \boldsymbol{q}_{c_1} = (1, 1) \\ &\succeq_{c_2}\!: s_1, s_3, \dots && \boldsymbol{q}_{c_2} = (1, 1) \\ &\succeq_{c_3}\!: s_1, s_4, s_2, s_3 && \boldsymbol{q}_{c_3} = (1, 1). \end{aligned}$$

Students preferences are given by

$$\succeq_{s_1}\!: c_1,$$
$$\succeq_{s_2}\!: c_2,$$
$$\succeq_{s_3}\!: c_2,$$
$$\succeq_{s_4}\!: c_3.$$

In this market, the top trading cycles mechanism results in matching μ given by

$$\mu(c_1) = s_1,$$
$$\mu(c_2) = s_3,$$
$$\mu(c_3) = s_4,$$
$$\mu(s_2) = \emptyset.$$

Now suppose that the priority of c_3 is changed to

$$\succsim_{c_3} : s_4, s_1, s_2, s_3.$$

The new market has a stronger priority-based affirmative action policy than the previous one. In this market, the top trading cycles mechanism results in matching $\tilde{\mu}$ given by

$$\tilde{\mu}(c_1) = s_1,$$
$$\tilde{\mu}(c_2) = s_2,$$
$$\tilde{\mu}(c_3) = s_4,$$
$$\tilde{\mu}(s_3) = \emptyset.$$

Thus, although the new market has a stronger prioirity-based affirmative action policy than the original one, the resulting assignment is Pareto inferior for the minority: Student s_3 is made worse off under $\tilde{\mu}$ than under μ, while s_4 is matched to the identical school under μ and $\tilde{\mu}$. This completes the proof. □

5. Conclusion

This paper investigated welfare implications of affirmative action policies in the context of school choice. Our main finding is that there are environments in which affirmative action policies inevitably hurt every minority student – the purported beneficiaries – under any stable matching mechanism. This result seems to be bad news as stable matching mechanisms are used in New York City and Boston and are being considered in other school districts.

The above negative conclusion is robust to specific measures for implementing affirmative action. More specifically, our impossibility theorems hold both when certain school seats are reserved for minority students and when minority students are given preferential treatment in receiving priority in schools. Furthermore, the same impossibility results hold under the top trading cycles mechanism.

The findings of this paper suggest that caution should be exercised when employing affirmative action policies. Even if affirmative action is deemed desirable by society,

the policies may hurt purported beneficiaries. In that sense this paper is in agreement with extensive studies that point out the potential shortcomings of affirmative action policies.[8] The difficult problem of affirmative action seems to require further theoretical and empirical research.

References

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.

Balinski, M., and T. Sönmez (1999): "A tale of two mechanisms: student placement," Journal of Economic Theory, 84, 73–94.

Ehlers, L. (2010): "Controlled School Choice," Unpublished mimeo.

Ergin, H., and T. Sönmez (2006): "Games of School Choice under the Boston Mechanism," Journal of Public Economics, 90, 215–237.

Fryer, R., and G. Loury (2005): "Affirmative action and its mythology," The Journal of Economic Perspectives, 19(3), 147–162.

Gale, D., and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," American Mathematical Monthly, 69, 9–15.

Roth, A. E. (1982): "The Economics of Matching: Stability and Incentives," Mathematics of Operations Research, 7, 617–628.

——— (1991): "A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians and Surgeons in the U.K.," American Economic Review, 81, 415–440.

——— (2007): "Deferred Acceptance Algorithms: History, Theory, Practice and Open Questions," forthcoming, *International Journal of Game Theory*.

Roth, A. E., and M. A. O. Sotomayor (1990): Two-sided matching: a study in game-theoretic modeling and analysis. Econometric Society monographs, Cambridge.

Shapley, L., and H. Scarf (1974): "On cores and indivisibilities," Journal of Mathematical Economics, 1, 23–37.

Sönmez, T. (1997): "Manipulation via Capacities in Two-Sided Matching Markets," Journal of Economic Theory, 77, 197–204.

[8]See Ehlers (2010) that studies difficulty of controlled school choice policies different from the current ones. Fryer and Loury (2005) discuss issues with affirmative action policies more generally.

——— (1999): "Can Pre-arranged Matches be Avoided in Two-Sided Matching Markets?," Journal of Economic Theory, 86, 148–156.

Sönmez, T., and M. U. Ünver (2009): "Matching, Allocation, and Exchange of Discrete Resources," forthcoming, Handbook of Social Economics, eds. Jess Benhabib, Alberto Bisin, and Matthew Jackson, Elsevier.