Discussion Paper Series – CRC TR 224

Discussion Paper No. 666
Project B 04

Transparent Matching Mechanisms

Markus Möller[1]

March 2025

[1]University of Bonn, Email: mmoelle2@uni-bonn.de

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Transparent Matching Mechanisms*

Markus Möller

February 2025

Abstract

In a standard one-to-one agent-object matching model, I consider a central matching authority that publicly announces a strategy-proof mechanism and then initiates a matching. Following Akbarpour and Li (2020), the authority's commitment to the announced mechanism is limited to mechanisms rendering participants' observations indistinguishable from it. I call an announced mechanism *transparent* if any deviation from it would be detected.

The main findings identify trade-offs regarding transparency and other desirable properties: Under stability or efficiency, strategy-proof mechanisms are transparent if and only if they are dictatorial. At the same time, the agent-proposing Deferred Acceptance (DA) mechanism is tantamount to committing to stability, while efficient mechanisms often fail to commit to efficiency. This transparency trade-off between stability and efficiency persists when strategy-proofness is guaranteed.

Keywords: Matching, Transparency, Strategy-Proofness, Stability, Efficiency.
JEL Codes: C78, D47, D82.

*Thanks to Alexander Westkamp, Stephan Lauermann, Aram Grigoryan, two anonymous referees, and the co-editor for providing valuable feedback. Support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2126/1 - 390838866 and through CRC TR 224 (Project B04) is gratefully acknowledged.
University of Bonn. E-mail: mmoelle2@uni-bonn.de

1 Introduction

This paper examines to what extent participants can be confident that a central matching authority follows its announced assignment rules. Several real-world examples reveal that the authority's conduct may deviate from the rules it has promised. In 2020, Boston Public Schools (BPS) incorrectly rejected dozens of students from the city's most prestigious exam schools, attributing the instance to internal miscommunication. Despite an external audit, a similar error reappeared in 2023.[1] Another prominent example is the bribery scandal at U.S. colleges, which surfaced in 2019, where university officials used fabricated athletic credentials and other tactics to influence admissions in favor of particular applicants.[2] Likewise, misconduct occurred at the National Resident Matching Program (NRMP) and Chicago Public Schools, uncovered only after thorough third-party investigations or explicit audits.[3] In contrast, the 2020 BPS admission flaw was detected by a student's tutor who noticed a discrepancy between the student's grades and the official admissions scores. These incidents raise a central question: Under which admission mechanisms can deviations remain undetected by participants?

Motivated by these concerns, I employ a canonical one-to-one object-allocation framework with privately known preferences and no monetary transfers. In this setting, an authority publicly announces a *strategy-proof* direct mechanism but then may implement a different mechanism behind the scenes, producing a publicly observed matching. Following Akbarpour and Li (2020), the authority's commitment is limited to *safe deviations*—those for which each agent's individual observation can still be explained by some recombination of other agents' reports under the announced mechanism. However, unlike Akbarpour and Li (2020), I do not impose assumptions on the authority's objectives; its deviations may be intentional or purely erroneous. I call a mechanism *transparent* if it permits no safe deviations.

The main analysis investigates how transparency interacts with three fundamental properties widely regarded as desirable in practical settings: *strategy-proofness,*

[1] *The Boston Globe, August 31, 2020* and *The Boston Globe, April 12, 2023*.

[2] *Press release of the U.S. Attorney's Office, District of Massachusetts, March 12, 2019*.

[3] In 1995, the NRMP faced claims that it had broken its promise to use a mechanism not manipulable by residents (Williams, 1995)—a claim later independently verified by Roth and Peranson (1997). In Chicago, an independent audit of the 2016–2017 admission process uncovered privileged treatment, documentation errors, and additional screening of applicants (Grigoryan and Möller, 2024; Schuler, 2018).

stability, and efficiency. All three have been extensively examined in the literature and were central to policy decisions influenced by transparency concerns. In 2005, for instance, BPS considered two candidates, the stable DA and the efficient TTC. The committee ultimately chose DA, stating that "*the behind-the-scenes mechanized trading [in TTC] makes the student assignment process less transparent*" (Leshno and Lo, 2021). Yet while both DA and TTC are strategy-proof, stability and efficiency themselves are mutually incompatible (Balinski and Sönmez, 1999).

To identify potential trade-offs in terms of transparency, the analysis adopts an informational benchmark motivated by incomplete privacy. The idea is that while privacy protection typically ensures that students' reports remain confidential, their outcomes and certain priority criteria (e.g., walk-zones, special abilities) can be difficult to hide. Students thus may gauge their relative standing in school's priority score ranking; and indeed, as the 2020 BPS case shows, such partial information sometimes suffices for participants to detect deviations.

Specifically, I assume that agents' preferences remain private, but other features—such as the set of agents, objects, and scores—are public knowledge. However, the key insights of the paper also extend to the school-choice framework of Abdulkadiroğlu and Sönmez (2003) under alternative informational benchmarks. With the exception of Proposition 3, all priority-based allocation results carry over, including settings where participants only learn their own priority scores and objects' capacities and eventually observe only their own assignment and objects' cutoffs (i.e., for each object the score of the lowest-scoring agent assigned to it).[4]

Main Results The first set of results indicates a trade-off between strategy-proofness and transparency. Specifically, I show that any strategy-proof mechanism that is stable or efficient is transparent if and only if it is dictatorial. In particular, DA is transparent exactly when it collapses to a serial dictatorship (Proposition 1). Likewise, any strategy-proof and efficient mechanism is transparent if and only if it is equivalent to a *sequential dictatorship* (Theorem 1). In light of these findings, the remaining insights focus on the trade-off between stability and efficiency under weaker transparency considerations.

[4]Naturally, many interesting scenarios lie outside this framework. One such scenario involves bribery that influences agents' private information or instances where agents are bribed directly to cover a deviation.

Concretely, I ask whether the authority's announcements can genuinely uphold its claims of stability and efficiency. A central result of this paper establishes that announcing DA is equivalent to a commitment to stability: a deviation from DA is safe precisely when the deviation is itself a stable mechanism for the market's underlying priorities (Theorem 3). By contrast, for efficient mechanisms, the authority can often remain undetected even when it induces inefficient outcomes (Theorem 2).

The transparency advantage of stability over efficiency persists when the authority can guarantee that all its deviations are strategy-proof. Indeed, while every deviation from DA is detected in this alternative setup, TTC is not transparent if the priority structure contains cycles (Proposition 2). The absence of these cycles is directly linked to a condition making TTC transparent (Proposition 3) and that is weaker than some other well-known acyclicity condition characterizing TTC^s with regard to various desirable properties.

Before turning to the main analysis, I briefly review the related literature and outline the organization of the paper.

Related Work This paper is among the first in matching markets to relax the authority's assumption of full commitment. However, there are recent studies that offer complementary perspectives on this topic in allocation problems.

Independently of this work, Grigoryan and Möller (2024) explore how much information is needed to detect deviations. They introduce an index defined by the minimum-sized group of individuals whose information is sufficient to detect any deviation. Under the *Immediate Acceptance (IA)* mechanism, two agents suffice to detect any deviation; by contrast, under DA and TTC or sequential dictatorships, detecting a deviation often requires access to all agents' information. Relatedly, Pycia and Ünver (2024) show that for group-strategy-proof and efficient mechanisms, deviations from Arrovian efficiency can be verified by comparing a single agent's relative outcome ranking plus an unknown challenger alternative.

In Hakimov and Raghavan (2023) transparency arises from designing information structures implementable via sequential public disclosure of interim cutoffs and private feedback. They show that there exist transparent information structures for DA and TTC in which each agent reports only one object at a time. However, more generally, neither private feedback nor cutoffs alone suffice to achieve a transparent protocol.

Unlike in these concurrent works, the notion studied here is a feature of the

mechanism, since communication is entirely private, mechanisms are direct and agents' strategic behavior is taken into account. The present paper also finds necessary and sufficient conditions for entire classes of strategy-proof mechanisms and examines commitment to desirable properties.

Akbarpour and Li (2020) analyze a framework with sequential private communication between an authority and agents, that focuses on credible Bayes–Nash implementation under imperfect information. Both Akbarpour and Li (2020) and Woodward (2020) study partial commitment in auctions where every deviation is authority-initiated. The authority's objectives are publicly known, which is a suitable assumption in contexts where auctioneers aim primarily to maximize revenue—giving bidders a clearer view of the auctioneer's incentives. By contrast, transparency does not require deviations to be incentive compatible for the authority, nor does it assume any specific or known objective on the authority's part. In line with the motivating applications, it thus remains agnostic about possible motives or mistakes behind deviations.

More broadly, this paper contributes to our understanding of the structure and verifiability of matching mechanisms (Gonczarowski and Thomas, 2024; Gangam et al., 2023), and it connects to the literature modeling limited commitment via agents' observable outcomes (Dequiedt and Martimort, 2015; Baliga et al., 1997; Bester and Strausz, 2000, 2001).

The paper proceeds as follows. Section 2 presents the formal model and the definition of transparency. Sections 3 and 4 analyze efficient and stable mechanisms, respectively. Section 5 extends the analysis to the case in which strategy-proofness is guaranteed. The Appendix contains all proofs omitted from the main text.

2 The Basic Framework

2.1 Preliminaries

There are finite sets of agents I and indivisible objects $X \cup \{\varnothing\}$, where $\varnothing$ denotes the outside option and $|I| \geq 2$ and $|X| \geq 2$.

Each agent $i \in I$ has a strict preference relation P_i over $X \cup \{\varnothing\}$, where R_i is the corresponding weak preference relation.[5] Object $x \in X$ is *acceptable* if $xP_i\varnothing$ for i,

[5]That is, R_i is a complete, transitive and antisymmetric binary relation. For each pair of objects

and *unacceptable* otherwise. Let P_i be i's preference ranking and let $P \equiv (P_i)_{i \in I}$ be a (preference) profile with the corresponding domain $\mathcal{P}$. Also, for any $J \subseteq I$, denote by $P_J = (P_j)_{j \in J}$ the profile restricted to J. We denote by $-i$ the set of all agents except i.

For each object $x \in X$ and for each agent $i \in I$, assign a score $s_i^x \in \mathbb{R}^{++}$ ensuring $s_i^x \neq s_j^x$ for any $j \neq i$. For each pair of agents $i, j \in I$, we say i has higher priority or score at $x \in X$ than j if and only if $s_i^x > s_j^x$. Thus, for each $x \in X$, the collected scores $s^x = (s_i^x)_{i \in I}$ induce a strict (priority) score ranking over the agents. Let $s = (s^x)_{x \in X}$ be a *score structure*.

A matching $\mu : I \to X \cup \{\varnothing\}$ assigns each agent exactly one object, with no two agents receiving the same object from X. Any agent receiving the outside option $\varnothing$, and any object in X not assigned to an agent, is called *unassigned*. Let $\mathcal{M}$ collect the set of all possible matchings.

We now recall several standard definitions used throughout the analysis. A matching μ is *non-wasteful* if no unassigned object $x \in X$ is strictly preferred by an agent i over $\mu(i)$. It is *individually rational* if, for every i, object $\mu(i) R_i \varnothing$. A matching μ is *blocked* by i at $x \in X$, if there exists j such that $x \, P_i \, \mu(i)$, $\mu(j) = x$, and $s_i^x > s_j^x$. A matching μ is *stable* for s if it is not blocked, individually rational, and non-wasteful. A matching μ is *(Pareto) efficient* if there is no other matching ν such that $\nu(i) \, R_i \, \mu(i)$ for all $i \in I$, and $\nu(j) \, P_j \, \mu(j)$ for some $j \in I$.

A *mechanism* $g : \mathcal{P} \to \mathcal{M}$ maps each profile to a matching. Denote $g_i(P)$ as i's match under $g(P)$. A mechanism g is *individually rational* if it yields only individually rational matchings, and *non-wasteful* if it generates only non-wasteful ones. Stable and efficient mechanisms are defined analogously, and they are also individually rational and non-wasteful by definition.

A mechanism g is *strategy-proof* if, for all P, there is no $i \in I$ and P_i' such that $g_i(P_i', P_{-i}) \; P_i \; g_i(P)$. In words, a mechanism is strategy-proof if no single agent can do better by misrepresenting her preferences. A mechanism g is *group strategy-proof* if, for all P, there is no $J \subseteq I$ and P_J' such that $g_i(P_J', P_{-J}) \, R_i \, g_i(P)$ for each $i \in J$, and $g_j(P_J', P_{-J}) \, P_j \, g_j(P)$ for at least one $j \in J$. In words, group-strategy-proofness requires that no group of agents can jointly misrepresent their preferences to weakly improve everyone's assignment while strictly benefiting at least one member. We say a mechanism g is *non-bossy* if for all P, there is no $i \in I$, and P_i', such that

$x, y \in X \cup \{\varnothing\}$, we write $x R_i y$ if either $x P_i y$ or $x = y$.

$g_i(P) = g_i(P'_i, P_{-i})$, but $g(P) \neq g(P'_i, P_{-i})$. In other words, there is no agent that is *bossy* in the sense that changing this agent's preferences affects other agents' assignments but not her own.

2.2 A Transparency Framework

Consider a central matching authority that makes an *announcement* g public to all agents. Once agents report their preferences P, the authority uses a mechanism $\tilde{g}$ to induce a publicly observable outcome $\tilde{g}(P)$. Yet the mechanism $\tilde{g}$ itself and the profile P remain confidential. Hence, from the perspective of individual agents, the outcomes under $\tilde{g}$ may be indistinguishable from those following from announcement g. We now formalize when agents can detect that the authority did not use g.

Assume that each agent i knows the elements I, X, s and how g operates. Also, let i's *observation* $o_i\left(P, \tilde{g}(P)\right)$ be the ordered pair $\left(P_i, \tilde{g}(P)\right)$. A mechanism $\tilde{g}$ is a *deviation* if there is a profile P for which $\tilde{g}(P) \neq g(P)$. Following Akbarpour and Li (2020), an observation $o_i\left(P, \tilde{g}(P)\right)$ under $\tilde{g}$ has an *innocent explanation* for agent i if there exists P'_{-i} such that $o_i\left(P, \tilde{g}(P)\right) = o_i\left((P_i, P'_{-i}), g(P_i, P'_{-i})\right)$. In other words, an observation has an innocent explanation if the observation could follow from a configuration of other agents' preferences under g. When i does not have an innocent explanation for her observation $o_i\left(P, \tilde{g}(P)\right)$, then we say that i *detects* the deviation $\tilde{g}$ from g. A deviation is *safe* if for every agent i and every profile P, observation $o_i\left(P, \tilde{g}(P)\right)$ has an innocent explanation.

We define the following transparency notion, which requires that any deviation can be detected by at least one agent.

Definition 1. A mechanism g is *transparent* if it has no safe deviations.

In the remainder of the paper, we apply this notion to the two canonical classes of stable and efficient mechanisms. Before doing so, consider the following observation that we rely on repeatedly in the analysis.

Lemma 1. *If g is non-wasteful and individually rational, then any safe deviation $\tilde{g}$ from g is non-wasteful and individually rational.*

It is easy to see that the statement follows from each agent observing her preference ranking and the entire matching.

3 Efficient Mechanisms

This section studies the transparency features of efficient mechanisms. First, we demonstrate that efficient mechanisms may admit safe deviations that lead to inefficient outcomes. To motivate such a deviation, consider a public school assignment setting where the authority seeks to enforce hidden distributional constraints (e.g., balanced representation across regions, genders, or socioeconomic groups).[6] If these constraints are incompatible with efficiency, the authority may still announce an efficient mechanism to encourage participation, but then deviates in order to comply with its hidden distributional objectives.

Example 1. Let $I = \{i, j\}$ and $X = \{x, y\}$ and consider s such that $s_i^x > s_j^x$ and $s_j^y > s_i^y$. The authority announces TTC^s, which is known to be efficient, strategy-proof (Roth, 1982; Abdulkadiroğlu and Sönmez, 2003) and induced via the TTC algorithm operating on s. We now construct a safe deviation $\tilde{g}$ from TTC^s, for which the preference profile in the table below will be central:

P_i	P_j	P'_i	P'_j
y	x	x	y
x	y	y	x
$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$

Specifically, consider a deviation $\tilde{g}$ that differs from TTC^s only with respect to the outcome obtained for profile P, where

$$\tilde{g}(P) = \{(i, x), (j, y)\} \neq \{(i, y), (j, x)\} = TTC^s(P).$$

Since i and j prefer to exchange x and y under $\tilde{g}(P)$, deviation $\tilde{g}$ is not efficient.

To see that $\tilde{g}$ is safe, note that if agent j reports P'_j, we have $o_i\left(P, \tilde{g}(P)\right) = o_i\left((P_i, P'_j), TTC^s(P_i, P'_j)\right)$. Similarly, from j's perspective, if agent i reports P'_i, then $o_j\left(P, \tilde{g}(P)\right) = o_j\left((P'_i, P_j), TTC^s(P'_i, P_j)\right)$. Since for any other scenario, $\tilde{g}$ coincides

[6] See, for instance, the work on matching under regional constraints (Kamada and Kojima, 2015), affirmative action (Abdulkadiroğlu and Sönmez, 2003); (Abdulkadiroğlu et al., 2005); (Kojima, 2012); Hafalir et al. (2013), matching under complex constraints (Westkamp, 2013), or diversity constraints (Ehlers et al., 2014).

with TTC^s, all remaining observations trivially have innocent explanations. Thus, deviation $\tilde{g}$ is safe and inefficient. □

Recall that the TTC algorithm requires that at each step t any unassigned object points to its highest-scored unassigned agent. A common interpretation of this pointing is that the agent becomes the object's *owner*, enabling her either to obtain it via a *self cycle* (a cycle of length 1) or to trade it in a *trading cycle*. If an agent observes that another agent's score for the assigned object is high enough to envision her securing it via a self cycle, then this self cycle scenario provides a straightforward explanation for why the observer herself did not get that object. For instance, at profile P in Example 1, x and y are acceptable to their respective owners i and j. Rather than inducing the trading cycle in which i trades x with j for y, the authority can safely assign them via two self cycles (i.e., i points to x, j points to y, and vice versa). The same reasoning applies if the trading cycle involves more agents and objects. Conversely, if a deviation results in an observation from which an agent deduces that she should have been the owner of a more-preferred object, then the deviation is not safe.

Next, we introduce *sequential dictatorship mechanisms* Pápai (2000, 2001) which do not rely on trading cycles and are therefore inherently resistant to the kind of safe deviations just discussed. For each $\tilde{X} \subseteq X$ and P_i, let $top(P_i, \tilde{X})$ be the highest-ranked object on P_i among the outside option and all objects not in $\tilde{X}$. Moreover, for each P, let $\pi_P : \{1, \dots, |I|\} \to I$ be a bijection representing a dictatorial ordering of the agents such that for each $n \in \{1, \dots, |I|\}$, $\pi_{P,n}$ denotes the n-th dictator at P.

Definition 2. A mechanism g is a *sequential dictatorship*, if there are dictatorial orderings $\{\pi_P\}_{P \in \mathcal{P}}$ such that for any pair $P, \tilde{P}$ and for each $n \in \{1, \dots, |I|\}$, the following two conditions are satisfied:

(i) $g_{\pi_{P,n}}(P) = top\left(P_{\pi_{P,n}}, \cup_{l=1}^{n-1} g_{\pi_{P,l}}(P)\right)$, and

(ii) if $g_{\pi_{P,m}}(P) = g_{\pi_{\tilde{P},m}}(\tilde{P})$ for each $m < n$, then $\pi_{P,n} = \pi_{\tilde{P},n}$.

According to condition (i), a sequential dictatorship recursively defines matchings such that, for each profile, the respective dictator is assigned her most preferred object still available after all previous dictators have been assigned. Condition (ii) implies that the first dictator is always the same, while the next dictator's identity depends only on previous dictators' assignments and not on their detailed preferences.

The first main result of this section fully characterizes efficient and transparent announcements as sequential dictatorships.

Theorem 1. *Take any efficient announcement g. Then, g is transparent if and only if it is a sequential dictatorship.*

To see why sequential dictatorships are not transparent, note that at each step exactly one agent is guaranteed to pick her favorite object from the remaining ones. Observing the first dictator's assignment reveals the identity of the second dictator, whose assignment then reveals the third dictator, and so forth. Consequently, given her observation, each agent can trace the correct ordering of dictators and the objects that each dictator should have been able to choose from. Therefore, if the authority deviates from some profile, the first agent who realizes she must have been the dictator at a given stage, yet did not receive the supposed best available object, has no innocent explanation for her observation. Accordingly, the deviation is not safe.

The proof of the converse statement is divided into two parts. First, we consider efficient announcements that are not group-strategy-proof. By Pápai (2000), group-strategy-proofness is characterized by strategy-proofness and non-bossiness. Now, if an efficient strategy-proof mechanism is not group-strategy-proof, then strategy-proofness implies that there is a bossy agent at a profile in which this agent receives her top choice. The safe deviation we construct in this part of the proof, reproduces this agent's bossiness at that very profile: it preserves her assignment while altering other agents' assignments as if the agent were bossy. Such a deviation can be safe because efficiency of the announcement implies that the bossy agent can innocently explain her observation with a scenario in which every other agent has also received her top choice. At the same time, other agents can just attribute their observations to the bossy agent's possible preference shift.

The second part examines efficient and group-strategy-proof announcements. Each such mechanism coincides with a *Top-Cycle (TC)* mechanism (Pycia and Ünver, 2017), whose corresponding TC algorithm operates similarly to TTC but allows for more complex pointing rules.[7] If the mechanism is not a sequential dictatorship, one can find a profile and step of the TC algorithm where a trading cycle forms (i.e.,

[7]For a brief discussion of the TC algorithm, see Appendix A. The characterization of group-strategy-proof and efficient mechanisms of Pycia and Ünver (2017) extends to the setting with outside options as described in (Pycia and Ünver, 2017, Supplement, p.6) and Pycia and Ünver (2014). The same holds for the characterization of Pápai (2000) as shown in (Pycia and Ünver, 2014).

it contains at least two agents). We then construct a safe and inefficient deviation centered around this trading cycle by extending the key ideas as discussed for TTC^s in the context of Example 1.

Together, these arguments lead directly to the following three-way equivalence for efficient and group-strategy-proof mechanisms.

Theorem 2. *If g is efficient and group-strategy-proof, then the following three statements are equivalent:*

1. *g is transparent.*
2. *g is a sequential dictatorship.*
3. *g admits only efficient safe deviations.*

With very similar arguments, a characterization akin to Theorem 1 and Theorem 2 holds for the entire class of TC mechanisms in the many-to-one framework.[8] Also, by the same reasoning as described for the one-to-one setting, any efficient mechanism that is not group-strategy-proof is not transparent. However, the example below illustrates that group strategy-proofness cannot be relaxed to strategy-proofness in the statement of Theorem 2.

Example 2. Let $I = \{i, j, k\}$ and $X = \{x, y\}$. Denote $\hat{\mathcal{P}} = \{\hat{P} \in \mathcal{P} \mid \hat{P}_i = P_i\}$, where $P_i : x, y, \varnothing$ and consider g such that given any $P \notin \hat{\mathcal{P}}$, agents select their favorite objects among the remaining ones according to ordering i,j,k, while for any $P \in \hat{\mathcal{P}}$, the ordering changes to i,k,j.

Clearly, g is strategy-proof and efficient. To see that g is not group-strategy-proof, consider profiles $P = (P_i, P_{-i})$ and $P' = (P'_i, P_{-i})$ with $P_i : x, y, \varnothing$ and $P_i = P_j = P_k$, while $P'_i = x, \varnothing, y$. Announcement g leads to $g_i(P) = g_i(P')$ while $g_k(P) = y$ and $g_k(P') = \varnothing$, revealing that g is bossy. Consequently, g is not group-strategy-proof.

Next, take any deviation $\tilde{g}$ from g. First, under any safe $\tilde{g}$, it is clear that $g_i(P^*) = \tilde{g}_i(P^*)$ for any P^*. Now, consider a profile $\bar{P}$, where x is not ranked highest on $\bar{P}_i$. Hence, $o_j\left(\bar{P}, \tilde{g}(\bar{P})\right)$ reveals to j that $\bar{P} \notin \hat{\mathcal{P}}$. This means j detects $\tilde{g}$ unless $\tilde{g}_j(\bar{P}) = g_j(\bar{P})$ and therefore, we obtain $\tilde{g}(\bar{P}) = g(\bar{P})$.

[8] TC mechanisms remain efficient and group-strategy-proof in the many-to-one environment (Pycia and Ünver, 2011; Abdulkadiroğlu and Sönmez, 2003).

By Lemma 1, we know that if g is efficient and $\tilde{g}$ is safe, then $\tilde{g}$ is individually rational and non-wasteful. We show that this implies that $\tilde{g}$ is efficient. Concretely, consider a problem $\bar{P}$, where x is i's top-choice on $\bar{P}_i$. First, since $g_i(\bar{P}) = \tilde{g}_i(\bar{P}) = x$, if only one of $\bar{P}_j$ and $\bar{P}_k$ ranks y acceptable, then individual rationality would be violated whenever $\tilde{g}(\bar{P}) \neq g(\bar{P})$. Thus, $g(\bar{P}) = \tilde{g}(\bar{P})$ and therefore $\tilde{g}(\bar{P})$ is efficient. Second, if $\bar{P}_j$ and $\bar{P}_k$ both rank y over $\varnothing$, then by non-wastefulness of g either j or k must receive y implying efficiency of $\tilde{g}(\bar{P})$.

When $|X| < |I|$, then similar arguments apply to markets of any size and to mechanisms that are not dictatorial. By contrast, when $|X| \geq |I|$, any non-group-strategy-proof mechanism induced by a TC algorithm admits an inefficient safe deviation. Such a deviation can be constructed by applying the key ideas from Theorems 1 and 2.

4 Stable Mechanisms

This section analyzes the transparency properties of stable mechanisms. Since announcements are strategy-proof, we focus on the *agent-proposing DA*, the unique strategy-proof stable mechanism (Roth, 1984; Dubins and Freedman, 1981). Denote the DA operating on s with DA^s.

We begin with an elementary observation. Denote $\Sigma^s(P)$ as the set of stable matchings for P for s and let $\Sigma^s(P_i) = \bigcup_{\tilde{P}_{-i}} \Sigma^s(P_i, \tilde{P}_{-i})$ be the set of matchings that, given s and P_i, satisfy the stability conditions from the perspective of i. Combining these definitions directly leads to

Lemma 2. $\Sigma^s(P) = \bigcap_i \Sigma^s(P_i)$.

Note that Lemma 2 implies that stability can be verified agent-by-agent. This observation is central to prove the following core result of this paper.

Theorem 3. *A deviation $\tilde{g}$ from DA^s is safe if and only if $\tilde{g}$ is stable for s.*

Proof. Consider a deviation $\tilde{g}$ that is not stable for s. Thus, there exists P such that $\tilde{g}(P) \notin \Sigma^s(P)$. By Lemma 2, there exists i such that $\tilde{g}(P) \notin \Sigma^s(P_i)$. Since DA^s is stable for s, we have $DA^s(P_i, P'_{-i}) \in \Sigma^s(P_i)$ for any P'_{-i}. Consequently, since $\tilde{g}(P) \notin \Sigma^s(P)$, observation $o_i\left(P, \tilde{g}(P)\right)$ has no innocent explanation. Thus, $\tilde{g}$ is not safe.

Conversely, take any deviation $\tilde{g}$ that is stable for s. To establish that $\tilde{g}$ is safe, we derive an innocent explanation for any i and P and for any of her observations $o_i\left(P, \tilde{g}(P)\right)$. Consider any P'_{-i}, such that for each $j \neq i$, the object $\tilde{g}_j(P)$ is j's top choice on preferences P'_j. In this case, the fact $\tilde{g}(P) \in \Sigma^s(P)$ also implies that $\tilde{g}(P) \in \Sigma^s(P_i, P'_{-i})$. However, given profile (P_i, P'_{-i}) all agents but i receive their top choice under $\tilde{g}(P)$. Therefore, no other matching in $\Sigma^s(P_i, P'_{-i})$ Pareto dominates $\tilde{g}(P)$ (i.e., no stable matching makes all agents weakly better off, and at least one agent strictly better off). Since DA^s produces this agent-optimal stable matching, we have $\tilde{g}(P) = DA^s(P_i, P'_{-i})$. Thus, $o_i\left(P, \tilde{g}(P)\right)$ has an innocent explanation. Since i and P were arbitrary, $\tilde{g}$ is a safe deviation. □

While Theorem 3 severely restricts the scope for deviations from DA^s, it also implies that DA^s is transparent if and only if there is a unique stable matching for each profile. As shown next, under these conditions, DA^s reduces to a *serial dictatorship* (Satterthwaite and Sonnenschein, 1981; Svensson, 1994).

Proposition 1. *DA^s is transparent if and only if it is a serial dictatorship.*

To prove the result, one first shows that for DA^s to be a serial dictatorship, all objects must induce the same priority ranking. Because a serial dictatorship is simply a special case of a sequential dictatorship with a fixed dictatorial ordering, its transparency follows immediately. The need for such a strong restriction on s becomes clear from Example 1: in that example, DA^s coincides with TTC^s, the constructed $\tilde{g}$ remains stable under s, and agent i holds the highest score on x while agent j holds the highest score on y. As detailed in Appendix B, the straightforward priority and preference structures in Example 1 already hint at how the same logic extends to larger markets under DA^s.

However, although the logic behind DA^s transparency shares many features with that of efficient announcements, it is crucial to note that the constraints imposed on safe deviations by Lemma 2 have no counterpart under efficient mechanisms. In particular, unlike instability, inefficiency typically cannot be verified agent by agent, because it stems from mutual gains from trade that cannot be inferred from a single agent's report and outcome alone. Consequently, the scope for an authority to deviate safely from an efficient mechanism may be much broader.

In the remainder of this section, we briefly discuss how this section's findings extend to the public school assignment and college admission contexts (Abdulkadiroğlu

and Sönmez, 2003; Balinski and Sönmez, 1999). Concretely, assume each object has a publicly known capacity and that each agent's observation only comprises her own preferences, scores, assignment and a single-number statistic per object. Suppose the authority announces DA^s and additionally promises that the disclosed statistics correspond to the objects' *cutoffs*: specifically, if an object's capacity is filled, its cutoff is the lowest object-specific score among the agents assigned to it; otherwise, it is set to zero.

As argued in the following, even in this coarser environment, any deviation from stability for s will be detected. First, an agent's observation under the deviation can have an innocent explanation only if the disclosed statistic for her assignment does not exceed her own score on that object. Consequently, the disclosed statistics must be weakly lower than the genuine cutoffs of the induced matching. Second, because DA^s is stable for s, there must exist a configuration of other agents' assignments, preferences, and scores such that the resulting matching meets the stability constraints dictated by the agent's own assignment, preferences, scores, and the disclosed statistics. However, that matching is then stable under its genuine (higher) cutoffs, and thus remains stable for s. Hence, by similar reasoning as in Theorem 3 any deviation not stable for s is detected. It is also clear that the converse part of the proof of Theorem 3 applies, so the remaining results of this section follow immediately.

5 Strategic Agents

In this section, the authority commits *ex ante* to using a strategy-proof mechanism. This choice is motivated by the idea that in case all deviations are intentionally chosen by the authority, sophisticated agents could anticipate them. So the authority may decide to voluntarily guarantee strategy-proofness. Hence, we assume that any safe deviation must be strategy-proof itself. In this setup, we revisit DA^s and TTC^s. Starting with DA^s, we can directly apply Theorem 3 to get the following corollary:

Corollary 1. *DA^s is transparent.*

If s satisfies the acyclicity condition from Kesten (2006), then TTC^s is equivalent to DA^s and thus transparent by Corollary 1. These insights extend to the many-to-one framework of Abdulkadiroğlu and Sönmez (2003).[9] However, it later turns out that

[9]The weaker condition in Ergin (2002) coincides with the condition in Kesten (2006) only in the

Kesten's condition is not necessary. Instead, we now introduce a new cycle condition that captures the transparency of TTC^s more directly.

Definition 3. A *replacement cycle* in s consists of four agents $i, j, k, l \in I$ and two objects $x, y \in X$, all distinct, such that one of the following holds:

(1) $s_i^x > s_k^x > \{s_j^x, s_l^x\}$ and $s_j^y > s_l^y > \{s_i^y, s_k^y\}$, or

(2) $s_i^x > \{s_k^x, s_l^x\} > s_j^x$ and $s_j^y > \{s_k^y, s_l^y\} > s_i^y$.

In other words, there are two ways a replacement cycle can arise: Condition (1) requires that there are two pairs of agents who mutually outrank each other on two objects. To satisfy condition (2), there must be one pair that completely encloses another pair on two objects, with the top and bottom agents of the enclosing pair swapping positions across the objects.

Proposition 2. *If s has a replacement cycle, then TTC^s is not transparent.*

To build intuition for the statement, imagine a setting with four agents i, j, l, k and two objects x, y, where s contains a replacement cycle as in Definition 3 (1). Consider a problem where all objects are acceptable, i and j trade x and y, and the deviation $\tilde{g}$ swaps their assignments. The replacement cycle in s now ensures that agents k and l have sufficiently high scores at x and y to "replace" i at x and j at y if i and j respectively misreport x and y as unacceptable. In short, this replacement is necessary to prevent $\tilde{g}$ from producing waste or making violations of the score-based pointing rules of TTC^s apparent to i or j. By contrast, without it, $\tilde{g}$ must violate strategy-proofness to remain safe, since i and j could strategically forgo x and y to restore their TTC outcomes.

To see this more explicitly, assume we modify s so that i and j have the highest scores on x and y leaving anything else unchanged. Now suppose i misreports x as unacceptable. By individual rationality of TTC^s, strategy-proofness of $\tilde{g}$ would require that i receives the outside option. Then, for $\tilde{g}$ to be safe, j's top choice x cannot be wasted. Moreover, since j has a higher score at x than k and l, assigning one of them to x would mean j detects the deviation. So, j must receive x. But i still ranks y as her top choice and since i owns x, j can only acquire x by trading

one-to-one framework (Haeringer and Klijn, 2009; Kesten, 2006).

y in exchange for x with i. Hence, any safe $\tilde{g}$ must assign y to i which violates strategy-proofness $\tilde{g}$.

Given these arguments, consider the following condition that ensures s contains no replacement cycles.

Definition 4. A score structure s satisfies the *imperfect replacement property* if, for any four agents $i, j, k, l \in I$, there exist no two distinct objects $x, y \in X$ such that:

(1) $s_i^x > \{s_k^x, s_l^x\} > s_j^x$, and

(2) $s_l^y > \{s_i^y, s_j^y, s_k^y\}$ or $\{s_i^y, s_j^y, s_k^y\} > s_l^y$.

This property is weaker than the acyclicity condition of Kesten (2006), yet still sufficient for transparency of TTC^s.

Proposition 3. *If s satisfies the imperfect replacement property, then TTC^s is transparent.*

By definition, the imperfect replacement property holds in any market with at most three agents. Hence, by Proposition 3 the acyclicity conditions from Kesten (2006) and Mandal and Roy (2022) are not necessary for transparency of TTC^s, because they restrict top trading cycles in TTC to at most two agents.[10] Finally, the condition of Mandal and Roy (2022) is satisfied whenever there are at most two objects. Thus, by Proposition 2, it is not sufficient for transparency of TTC^s.

Appendix A Proofs of Section 3 and 5

This section contains the proofs of Section 3 and Section 5. In particular, subsection A.1 contains the proofs of Theorems 1 and 2 from Section 3, and subsection A.2 contains the proofs of Propositions 2 and 3 from Section 5.

We briefly describe some basic concepts and notation needed in this section. Since TTC^s is a special case of a TC mechanism, we only introduce these concepts once and use them in both subsections. Given any $J \subseteq I$, a submatching restricted to J is a mapping $\sigma : J \to X \cup \{\varnothing\}$. Let $\hat{I}_\sigma \equiv I \setminus \{J\}$ and let $\hat{X}_\sigma \subseteq X$ be the set of

[10] The acyclicity condition of Mandal and Roy (2022) characterizes priority structures for which TTC^s is *obviously strategy-proof* (Li, 2017; Mandal and Roy, 2022). See also Troyan (2019) for a similar condition.

unassigned agents and objects under σ. Note that the outside option $\varnothing$ is not in $\hat{X}_\sigma$ because it is always available. Nonetheless, recall that if an agent is mapped to $\varnothing$, we treat her as assigned, so she does not appear in $\hat{I}_\sigma$.

For any input P and any sequence of steps $t = 1, 2, \ldots$ denote by $\sigma^{t-1}(P)$ the submatching of agents and objects matched at the beginning of step t. We say a submatching σ is *on-path* (on TC or TTC) if there exists a profile P and a step t such that $\sigma = \sigma^{t-1}(P)$.

A.1 Proofs of Theorem 1 and Theorem 2

In the following, we prove Theorem 1 and Theorem 2. Lemma 3 presented first, implies the sufficiency parts of both statements. The converse direction for Theorem 2 follows from Lemma 4, while Theorem 1 needs Lemma 5 in addition. Thus, we first establish transparency of sequential dictatorships.

Lemma 3. *If g is a sequential dictatorship, then g is transparent.*

Proof. Consider an arbitrary sequential dictatorship g. Let $\tilde{g}$ be an arbitrary deviation from g and select P such that $\tilde{g}(P) \neq g(P)$. Given π_P, for each $i \in I$, denote $n_i = \pi_P^{-1}(i)$. Also, let $I' = \{i' \in I |\ g_{i'}(P) \neq \tilde{g}_{i'}(P)\}$ and select $i \in I'$ such that, for all $i' \in I' \setminus \{i\}$, it holds $n_i \leq n_{i'}$. We show that $o_i\left(P, \tilde{g}(P)\right)$ has no innocent explanation.

First, note that for each $\tilde{P}_{-i}$ such that $g_k(P) = g_k(P_i, \tilde{P}_{-i})$ for all k with $n_k < n_i$, Definition 2 (ii) implies that $n_i = \pi^{-1}_{(P_i, \tilde{P})}(i)$. However, then Definition 2 (i) means that $g_i(P) = g_i(P_i, \tilde{P}_{-i})$. This implies that given $g_i(P) \neq g_i(P_i, \tilde{P}_{-i})$, i has no innocent explanation for $o_i\left(P, \tilde{g}(P)\right)$. Therefore, $\tilde{g}$ is not safe and thus $\tilde{g}$ is transparent. □

Next, we proceed with the sufficiency parts of Theorem 1 and Theorem 2.

Lemma 4. *If g is efficient, group-strategy-proof and not a sequential dictatorship, then there exists a safe and inefficient deviation $\tilde{g}$ from g.*

In the proof, we rely on the TC algorithm and the characterization by Pycia and Ünver (2017). We briefly recapture the basics needed to describe the pointing specification under the TC algorithm.

First, for each on-path submatching σ, each unassigned object $x \in \hat{X}_\sigma$ points to an unassigned agent $i \in \hat{I}_\sigma$, thereby making i either the *owner* or the *broker* of x. If i owns x at $\sigma^{t-1}(P)$, then from step t onward, i can obtain x by forming a self cycle

(where i points to x and x points back to i) or by trading x via a *trading cycle* with another agent. By contrast, if i is the broker of x, then i can only trade x in a trading cycle; i cannot form a self cycle with x unless i's status changes to become the owner of x. No agent is ever the owner or broker of the outside option $\varnothing$.

Second, we use the TC algorithm for outside options as described in Pycia and Ünver (2014) and (Pycia and Ünver, 2017, Supplement, p.5). Because we have a *common* outside option $\varnothing$, we apply a slightly modified version in which $\varnothing$ does not point to anyone, and any owner who assigns herself to $\varnothing$ is immediately assigned to it.

Third, to ensure that the TC algorithm induces a group-strategy-proof and efficient g, the pointing must be *consistent* (Pycia and Ünver, 2017) across on-path submatchings. The interested reader is kindly referred to an excellent description of the TC algorithm along with a rigorous discussion and interpretation of these consistency conditions in Pycia and Ünver (2017) and Pycia and Ünver (2014). We use some of the implications of consistency in the proof below. Especially, we use that once an agent has become an owner of an object, this ownership persists as long as the agent is still unassigned.

Proof. Consider any efficient and group-strategy-proof g that is not a sequential dictatorship. Our goal is to construct a deviation $\tilde{g}$ from g that is safe but inefficient. Since g is equivalent to a TC mechanism, it can be induced via the TC algorithm with consistent pointing rules.

We first make two preliminary observations. First, according to Pycia and Ünver (2017) (Theorem 6) and Pycia and Ünver (2014) (Proposition 3), given any on-path submatching σ, if there is a single agent who owns all objects in $\hat{X}_\sigma$, then there is no broker at σ. Second, if there is no on-path submatching with strictly more than one owner, then it is easy to see that g is equivalent to a sequential dictatorship as defined by Definition 2.

These observations thus imply that since g is not a sequential dictatorship, there exists an on-path submatching σ^* with two agents $i, j \in \hat{I}_{\sigma^*}$ and two objects $x, y \in \hat{X}_{\sigma^*}$ such that i owns x and j owns y at σ^*. Let K be the set of agents assigned under σ^*. Consider preferences P_K such that for each $k \in K$, the top choice under P_k is $\sigma^*(k)$. Because g is non-bossy, for any profile $(P_K, \tilde{P}_{\hat{I}_{\sigma^*}})$ where $\tilde{P}_{\hat{I}_{\sigma^*}}$ is chosen arbitrarily, the TC algorithm reaches $\sigma^{t^*-1}(P_K, \tilde{P}_{\hat{I}_{\sigma^*}}) = \sigma^*$ at some step t^*.

For the construction of a deviation $\tilde{g}$, the following preferences of agent i and j are central. For i consider P_i, P_i' such that

- yP_ix and for all $x' \in X \cup \{\varnothing\} \setminus \{x, y\}$: xP_ix',
- xP'_iy and for all $x' \in X \cup \{\varnothing\} \setminus \{x, y\}$: yP'_ix'.

For agent j, let the preferences P_j, P'_j be

- xP_jy and for all $x' \in X \cup \{\varnothing\} \setminus \{x, y\}$: yP_jx',
- yP'_jx and for all $x' \in X \cup \{\varnothing\} \setminus \{x, y\}$: xP'_jx'.

Also, denote $L = I \setminus \{K \cup \{i, j\}\}$ and let P_L be specified arbitrarily. Given $P_{-ij} = (P_K, P_L)$, let $P' = (P'_i, P'_j, P_{-ij})$.

Next, consider $\tilde{g}$ such that $\tilde{g}(P) = g(P')$, and for all $\tilde{P} \neq P$, be $\tilde{g}(\tilde{P}) = g(\tilde{P})$. We first establish that $\tilde{g}$ is a deviation from g. Given agents in K report P_K, we reach σ^* in step t^* and thus for all $k \in K$,

$$\tilde{g}_k(P) = g_k(P') = g_k(P).$$

Moreover, given P at step t^*, there is a trading cycle

$$x \to i \to y \to j \to x,$$

which implies $g_i(P) = y$ and $g_j(P) = x$. By contrast, if agents report P', then there are two self cycles

$$x \to i \to x \quad \text{and} \quad y \to j \to y,$$

at step t^*, thus $g_i(P') = x$ and $g_j(P') = y$ and consequently

$$\tilde{g}(P) = g(P') \neq g(P).$$

We conclude that $\tilde{g}$ is a deviation. In addition, it directly follows that $\tilde{g}$ is not efficient, since agents i and j prefer to trade x and y given P under $\tilde{g}(P)$.

The final step is to show that $\tilde{g}$ is safe. For each $i' \in I$, we need to find innocent explanations for observation $o_{i'}(P, \tilde{g}(P))$. Since $\tilde{g}(P) = g(P')$ it is clear that for each $i' \neq i, j$, one obtains

$$o_{i'}(P, \tilde{g}(P)) = o_{i'}(P', g(P')).$$

To find innocent explanations for the remaining agents i and j, consider profiles (P'_i, P_{-i}) and (P'_j, P_{-j}). Since for each $k \in K$ the assignment is identical under the

deviation, we must have $\sigma^{t^*-1}(P_i', P_{-i}) = \sigma^{t^*-1}(P_j', P_{-j}) = \sigma^*$. Given $\sigma^{t^*-1}(P_j', P_{-j})$ at step t^*, there is a self cycle $y \to j \to y$ that assigns y to j. This implies that i is assigned x, since i owns x at σ^*, i is unassigned at $\sigma^{t^*}(P_j', P_{-j})$, and x is her favorite among the remaining objects at step $t^* + 1$. A symmetric argument applies to profile (P_i', P_{-i}), where i forms a self cycle with x at step t^*. Hence, for both $i^* \in \{i, j\}$ it holds

$$g_{i^*}(P') = g_{i^*}(P_i', P_{-i}) = g_{i^*}(P_j', P_{-j}).$$

Finally, by non-bossiness of g, for all $l \in L$, it is clear that

$$g_l(P') = g_l(P_i', P_{-i}) = g_l(P_j', P_{-j}).$$

Hence,

$$\tilde{g}(P) = g(P') = g(P_i', P_{-i}) = g(P_j', P_{-j})$$

which directly leads to

$$o_i\left((P_j', P_{-j}), g(P_j', P_{-j})\right) = o_i\left(P, \tilde{g}(P)\right) \quad \text{and} \quad o_j\left((P_i', P_{-i}), g(P_i', P_{-i})\right) = o_j\left(P, \tilde{g}(P)\right).$$

We thus conclude that each $i' \in I$ has an innocent explanation for $o_{i'}\left(P, \tilde{g}(P)\right)$.

Since the remaining observations under $\tilde{g}$ coincide with those under g, we obtain that $\tilde{g}$ is a safe and inefficient deviation from g. This completes the proof. □

This concludes the proof for Theorem 2. The final lemma of this section shows that if g is efficient and not group-strategy-proof, then we can find a safe deviation and thereby finish the proof for Theorem 1.

Lemma 5. *If g is efficient and not group-strategy-proof, then g is not transparent.*

Proof. Since g is not group-strategy-proof, but strategy-proof, we know that g is bossy. Thus, there exists an agent $i \in I$ and two profiles P and $P' = (P_i', P_{-i})$ such that $g(P) \neq g(P')$ and $g_i(P) = g_i(P')$. Because g is strategy-proof, for any $P^* = (P_i^*, P_{-i})$ where $g_i(P)$ is ranked highest on P_i^*,

$$g_i(P^*) = g_i(P) = g_i(P').$$

Thus, since $g(P) \neq g(P')$, it is either true that $g(P^*) \neq g(P)$ or $g(P^*) \neq g(P')$ or both. Let $g(P^*) \neq g(P)$. A symmetric argument applies to the case where $g(P^*) \neq g(P')$.

Next, consider a deviation $\tilde{g}$ with $\tilde{g}(P^*) = g(P') \neq g(P^*)$ and $\tilde{g}(\tilde{P}) = g(\tilde{P})$ for all $\tilde{P} \neq P^*$. We show that $\tilde{g}$ is safe. Since all observations under g and $\tilde{g}$ coincide except for P^*, we only need to find innocent explanations for $o_j\left(P^*, \tilde{g}(P^*)\right)$ for each agent $j \in I$. Concretely, if $j = i$, consider $\hat{P}_{-i}$ such that for each agent $j \neq i$, $\hat{P}_j$ ranks $g_j(P)$ at the top position. Then, since the unique efficient matching under $(P_i^*, \hat{P}_{-i})$ is $g(P)$, we have $g(P_i^*, \hat{P}_{-i}) = g(P)$. Thus, $o_i\Big((P_i^*, \hat{P}_{-i}), g(P_i^*, \hat{P}_{-i})\Big) = o_i\left(P^*, \tilde{g}(P^*)\right)$. Second, for each $j \neq i$, we have $o_j\left(P^*, \tilde{g}(P^*)\right) = o_j\left(P', g(P')\right)$. Hence, $\tilde{g}$ is a safe deviation from g. This completes the proof. □

A.2 Proofs of Proposition 2 and Proposition 3

In the following, let $TTC^s = g$.

Proof of Proposition 2. By hypothesis, we have agents i, j, k, l and objects x, y that satisfy the replacement cycle inequalities from Definition 3. We focus on the case in which Definition 3 (1) is satisfied. Similar arguments apply for Definition 3 (2).

For the construction of the safe and strategy-proof deviation from g, we need the following preferences. Let $\hat{\mathcal{P}}$ be the set of all profiles P such that, P_i ranks y first, and x second, P_j ranks x first and y second, P_k and P_l only find x and y acceptable and for each $i' \notin \{i, j, k, l\}$, $P_{i'}$ ranks $\varnothing$ first. The rest of these rankings are specified arbitrarily. Moreover, given any such P_i for agent i, let $\bar{P}_i$ rank y last while keeping the same relative order of all other objects as under P_i. Similarly, for P_j, define $\bar{P}_j$ so that x is ranked last, while the relative ranking among the remaining objects is the same as under P_j.

Consider the following deviation $\tilde{g}$. For any profile $P \in \hat{\mathcal{P}}$, let $\tilde{g}(P) = g(\bar{P}_i, \bar{P}_j, P_{-ij})$ and for any profile $P' \notin \hat{\mathcal{P}}$, set $\tilde{g}(P') = g(P')$. Since for every $P \in \hat{\mathcal{P}}$, we have $g(P) \neq g(\bar{P}_i, \bar{P}_j, P_{-ij})$, it follows that $\tilde{g}$ is indeed a deviation.

To see that $\tilde{g}$ is safe, we can concentrate on profiles $P \in \hat{\mathcal{P}}$. Recall that i has the highest score for x and j has the highest score for y. Thus, for each $i' \neq i, j$, the profile $(\bar{P}_i, \bar{P}_j, P_{-ij})$ supports an innocent explanation for $o_{i'}\big(P, \tilde{g}(P)\big)$, since in this case i is assigned to x, j to y and all remaining agents receive the outside option. Since the same holds for profiles $(P_i, \bar{P}_j, P_{-ij})$ and $(\bar{P}_i, P_j, P_{-ij})$, innocent explanations for $o_i\big(P, \tilde{g}(P)\big)$ and $o_j\big(P, \tilde{g}(P)\big)$ follow directly.

It remains to show that $\tilde{g}$ is strategy-proof. First, for any $i' \notin \{i, j, k, l\}$, the deviation $\tilde{g}$ yields exactly the same assignments as g. Since g is strategy-proof, i'

cannot gain by misreporting under $\tilde{g}$. Now fix any arbitrary profile P'. Consider the following arguments for agents in $\{i, j, k, l\}$:

For agent l and each P', either $g_l(P') = \tilde{g}_l(P')$ or $\tilde{g}_l(P')\ P_l\ g_l(P')$. First, for each P' such that $g(P') = \tilde{g}(P')$, we have $g_l(\hat{P}_l, P'_{-l})\ R'_l\ \tilde{g}_l(\hat{P}_l, P'_{-l})$ for all $\hat{P}_l \in \mathcal{P}_l$. Second, for each P' such that $g(P') \neq \tilde{g}(P')$, we have $\tilde{g}_l(P')\ R'_l\ \tilde{g}_l(\hat{P}_l, P'_{-l})$ for all $\hat{P}_l$. Together, this implies that l has no incentive to deviate under $\tilde{g}$. Similar arguments apply to k.

For agent i and each P', either $g_i(P') = \tilde{g}_i(P')$ or $g_i(P')\ P_i\ \tilde{g}_i(P')$. First, for each P' such that $g(P') = \tilde{g}(P')$, we have $g_i(\hat{P}_i, P'_{-i})\ R'_i\ \tilde{g}_i(\hat{P}_i, P'_{-i})$ for all $\hat{P}_i \in \mathcal{P}_i$. Also, for each P' such that $g_i(P') \neq \tilde{g}_i(P')$, we obtain $\tilde{g}_i(P')\ R'_i\ \tilde{g}_i(\hat{P}_i, P'_{-i})$ for all $\hat{P}_i$. Similar arguments apply to j. This completes the proof. □

Next, we show that the imperfect replacement property is sufficient for transparency of TTC^s.

Proof of Proposition 3. Suppose that s satisfies the imperfect replacement property. We show that there is no safe, strategy-proof deviation from g. Let $\tilde{g}$ be an arbitrary deviation from g and assume that $\tilde{g}$ is strategy-proof. We demonstrate that $\tilde{g}$ cannot be safe. Recall that by Lemma 1 any safe deviation $\tilde{g}$ from g must be non-wasteful and individually rational.

The following preferences profile will be central to the arguments: Among all $P \in \mathcal{P}$, choose the P with the smallest t such that the realized submatching at the end of step t, $\sigma^t(P)$, implies $g(P) \neq \tilde{g}(P)$. Denote $\sigma^{t-1}(P) = \sigma_{\min}$. For each P', we say we are at $\sigma_{\min}$ under P' if we are at the beginning of step t^* and $\sigma^{t^*-1}(P') = \sigma_{\min}$.

Next, consider the TTC algorithm under P. Let I^t be the set of agents who own an object at step t, and let $\hat{I}^t \subseteq I^t$ be those agents assigned at step t for whom $\tilde{g}_{i'}(P) \neq g_{i'}(P)$. Since we are at $\sigma_{\min}$, it follows from strategy-proofness and efficiency of g, that $g_{i'}(P) P_{i'} \tilde{g}_{i'}(P)$ for every $i' \in \hat{I}^t$.

To establish that $\tilde{g}$ cannot be safe, we distinguish cases according to the cardinality of I^t. To begin with, it is clear from Definition 4 that $|I^t| \leq 3$ must hold. In the following, for each i, let P_i^* be such that for all $x \in X$ with $x \neq g_i(P)$, we have $g_i(P) R_i^* \varnothing P_i^* x$.

Case 1 Let $|I^t| = 1$. In this case, there must be a self cycle at step t. Using similar arguments as in the proof of Lemma 3, we conclude that $\tilde{g}$ cannot be safe. Moreover,

we can apply the same reasoning in subsequent cases whenever an agent in $\hat{I}^t$ forms a self cycle. For any such instance, $\tilde{g}$ is not safe.

Case 2 Let $|I^t| = 2$. Hence, by Case 1, there exist $i, j \in \hat{I}^t$ such that a trading cycle $j \to g_j(P) \to i \to g_i(P) \to j$ forms at step t under P.

We first establish that $\tilde{g}_i(P_i^*, P_{-i}) \neq g_i(P)$ implies $\tilde{g}_j(P_i^*, P_{-i}) \neq g_j(P)$. To start, strategy-proofness of $\tilde{g}$ implies that $\tilde{g}_i(P) \neq g_i(P)$ leading to $\tilde{g}_i(P_i^*, P_{-i}) \neq g_i(P)$. In particular, $\tilde{g}_i(P_i^*, P_{-i}) = \varnothing$ by individual rationality of g. By contradiction, assume $\tilde{g}_j(P_i^*, P_{-i}) = g_j(P)$. Since $g_j(P)$ points to i at σ_{min} and i points to $g_i(P)$, whenever $g_j(P_i^*, P_{-i}) = g_j(P)$, then by definition of σ_{min} and the pointing rules of TTC, observation $o_i\left((P_i^*, P_{-i}), \tilde{g}(P_i^*, P_{-i})\right)$ has no innocent explanation if there is $l \in \hat{I}_{\sigma_{min}}$ such that $\tilde{g}_l(P_i^*, P_{-i}) = g_i(P)$. Thus, also $g_j(P) \neq \tilde{g}_j(P_i^*, P_{-i})$. Using a symmetric argument, we also know $\tilde{g}_i(P_j^*, P_{-j}) \neq g_i(P)$ since $\tilde{g}_j(P_j^*, P_{-j}) \neq g_j(P)$.

Given these arguments, note that strategy-proofness of $\tilde{g}$ requires $\tilde{g}_i(P_i^*, P_j^*, P_{-ij}) = \tilde{g}_j(P_i^*, P_j^*, P_{-ij}) = \varnothing$. Moreover, by non-wastefulness of $\tilde{g}$ and the definition of σ_{min}, there must exist l,$l' \in \hat{I}_{\sigma_{min}} \backslash \{i, j\}$ with $\tilde{g}_l(P_i^*, P_j^*, P_{-ij}) = g_i(P)$ and $\tilde{g}_{l'}(P_i^*, P_j^*, P_{-ij}) = g_j(P)$. However, $l, l' \notin I^t$ since $|I^t| \leq 2$. W.l.o.g, let $l \notin I^t$. Then, since s satisfies the imperfect replacement property, there is no $x \in X$ such that $s_l^x > s_i^x$. Therefore, i cannot have an innocent explanation for $o_i\left((P_i^*, P_j^*, P_{-ij}), \tilde{g}(P_i^*, P_j^*, P_{-ij})\right)$ with $\tilde{g}_l(P_i^*, P_j^*, P_{-ij}) = g_i(P)$, since $g_i(P)$ cannot be assigned to l whenever $g_i(P)$ is i's top choice. Hence, if $|\hat{I}^t| = 2$, then $\tilde{g}$ is not safe.

Case 3 Let $|I^t| = 3$. Note that whenever there is no cycle with at least three agents and objects, or $\hat{I}^t < 3$, then the arguments in Case 1 and Case 2 can be applied to conclude that $\tilde{g}$ is not safe. Thus, in the remaining scenario there are exactly three agents $\hat{I}^t = \{i, j, k\}$ that form a trading cycle $i \to g_i(P) \to j \to g_j(P) \to k \to g_k(P) \to i$ at step t under P.

By strategy-proofness, $\tilde{g}_i(P_i^*, P_{-i}) \neq g_i(P)$. Hence, $\tilde{g}_i(P_i^*, P_{-i}) = \varnothing$ by individual rationality of g. However, since $|\hat{I}| = |I^t| = 3$ and $\tilde{g}$ is non-wasteful, we can use similar arguments as in Case 2. Specifically, there must exist $l \in \hat{I}_{\sigma_{min}} \setminus \hat{I}^t$ with $l \notin I^t$, but $\tilde{g}_l(P_i^*, P_{-i}) = g_{i'}(P)$ for some $i' \in \hat{I}$. This implies that $g_{i'}(P) P_{i'} \tilde{g}_{i'}(P)$. Then, because s satisfies the imperfect replacement property, there is no $x \in X$ such that $s_l^x > s_i^x$. However, l can never be assigned to $g_{i'}(P)$ under g, as long as i' prefers $g_{i'}(P)$ to all objects in $\hat{X}_{\sigma_{min}}$. Therefore, i' cannot have an innocent explanation for observation

$o_{i'}((P_i^*, P_{-i}), \tilde{g}(P_i^*, P_{-i}))$ with $\tilde{g}_l(P_i^*, P_{-i}) = g_{i'}(P)$.

This completes the case distinction. Hence, whenever $\tilde{g}$ is strategy-proof, then it is not safe. □

Appendix B Proof of Proposition 1

If DA^s is a serial dictatorship, then for any given pair of agents $i, j \in I$ and objects $x, y \in X$, it holds $s_i^x > s_j^x$ if and only if $s_i^y > s_j^y$. Given any P, following the ordering of the induced score ranking for some $x \in X$, for each $n \in \{1, \ldots, |I|\}$, the n-th ranked agent is guaranteed her top choice among the remaining objects after all previous agents in line have left. The first ranked agent must receive her top choice under any stable matching in $\Sigma^s(P)$. Next, the second-ranked agent receives, under any stable matching in $\Sigma^s(P)$, her top choice among objects once the first agent is left, and so forth. For each P, it is clear that $\Sigma^s(P)$ is a singleton. Therefore, Theorem 3 implies that there exists no safe deviation from DA^s, and thus DA^s is transparent.

If DA^s is not a serial dictatorship, then there exist two agents $i, j \in I$ and two objects $x, y \in X$, such that $s_i^x > s_j^x$ and $s_j^y > s_i^y$. We first construct a deviation $\tilde{g}$ from DA^s, for which we need the following preferences. For each $k \neq i, j$, consider P_k that ranks $\varnothing$ first. Moreover, let P_i, P_i' be such that

- yP_ix and for all $x' \in X \cup \{\varnothing\} \setminus \{x, y\}$: xP_ix', and
- $xP_i'y$, and for all $x' \in X \cup \{\varnothing\} \setminus \{x, y\}$: $yP_i'x'$.

Similarly, consider P_j, P_j' such that

- xP_jy and for all $x' \in X \cup \{\varnothing\} \setminus \{x, y\}$: yP_jx', and
- $yP_j'x$ and for all $x' \in X \cup \{\varnothing\} \setminus \{x, y\}$: $xP_j'x'$.

Next, for the profile $P = (P_i, P_j, P_{-ij})$, let $\tilde{g}(P)$ yield $\tilde{g}_i(P) = x$, $\tilde{g}_j(P) = y$, and for all $k \neq i, j$, $\tilde{g}_k(P) = \varnothing$. Also, for any $P' \neq P$, let $\tilde{g}(P') = DA^s(P')$. Since for P, the DA algorithm yields $DA_i^s(P) = y$, $DA_j^s(P) = x$, we know that $\tilde{g}$ is a deviation.

It remains to show that $\tilde{g}$ is safe. Except for profile P, innocent explanations are immediate. For the remaining case with preferences P, we have

$$DA^s(P_i', P_j, P_{-ij}) = DA^s(P_i, P_j', P_{-ij}) = \tilde{g}(P).$$

Hence, for each agent $i' \in I$, observation $o_{i'}(P, \tilde{g}(P))$ has an innocent explanation. Thus, $\tilde{g}$ is a safe deviation and DA^s is not transparent.

References

Atila Abdulkadiroğlu and Tayfun Sönmez. School choice: A mechanism design approach. *American Economic Review*, 93(3):729–747, 2003.

Atila Abdulkadiroğlu, Parag A Pathak, Alvin E Roth, and Tayfun Sönmez. The boston public school match. *American Economic Review*, 95(2):368–371, 2005.

Mohammad Akbarpour and Shengwu Li. Credible auctions: A trilemma. *Econometrica*, 88(2):425–467, 2020.

Sandeep Baliga, Luis C Corchon, and Tomas Sjöström. The theory of implementation when the planner is a player. *Journal of Economic Theory*, 77(1):15–33, 1997.

Michel Balinski and Tayfun Sönmez. A tale of two mechanisms: student placement. *Journal of Economic Theory*, 84(1):73–94, 1999.

Helmut Bester and Roland Strausz. Imperfect commitment and the revelation principle: the multi-agent case. *Economics Letters*, 69(2):165–171, 2000.

Helmut Bester and Roland Strausz. Contracting with imperfect commitment and the revelation principle: the single agent case. *Econometrica*, 69(4):1077–1098, 2001.

Vianney Dequiedt and David Martimort. Vertical contracting with informational opportunism. *American Economic Review*, 105(7):2141–82, 2015.

Lester E Dubins and David A Freedman. Machiavelli and the gale-shapley algorithm. *The American Mathematical Monthly*, 88(7):485–494, 1981.

Lars Ehlers, Isa E Hafalir, M Bumin Yenmez, and Muhammed A Yildirim. School choice with controlled choice constraints: Hard bounds versus soft bounds. *Journal of Economic theory*, 153:648–683, 2014.

Haluk I Ergin. Efficient resource allocation on the basis of priorities. *Econometrica*, 70(6):2489–2497, 2002.

Rohith R. Gangam, Tung Mai, Nitya Raju, and Vijay V. Vazirani. A Structural and Algorithmic Study of Stable Matching Lattices of Multiple Instances. arXiv preprint:2304.02590, 2023.

Yannai A. Gonczarowski and Clayton Thomas. Structural complexities of matching mechanisms. Working Paper, 2024. URL https://arxiv.org/abs/2212.08709.

Aram Grigoryan and Markus Möller. A theory of auditability for allocation mechanisms. Working Paper, 2024. URL https://arxiv.org/abs/2305.09314.

Guillaume Haeringer and Flip Klijn. Constrained school choice. *Journal of Economic theory*, 144(5):1921–1947, 2009.

Isa E Hafalir, M Bumin Yenmez, and Muhammed A Yildirim. Effective affirmative action in school choice. *Theoretical Economics*, 8(2):325–363, 2013.

Rustamdjan Hakimov and Madhav Raghavan. Improving transparency and verifiability in school admissions: Theory and experiment. *Management Science*, forthcoming, 2023.

Yuichiro Kamada and Fuhito Kojima. Efficient matching under distributional constraints: Theory and applications. *American Economic Review*, 105(1):67–99, 2015.

Onur Kesten. On two competing mechanisms for priority-based allocation problems. *Journal of Economic Theory*, 127(1):155–171, 2006.

Fuhito Kojima. School choice: Impossibilities for affirmative action. *Games and Economic Behavior*, 75(2):685–693, 2012.

Jacob D Leshno and Irene Lo. The cutoff structure of top trading cycles in school choice. *The Review of Economic Studies*, 88(4):1582–1623, 2021.

Shengwu Li. Obviously strategy-proof mechanisms. *American Economic Review*, 107 (11):3257–87, 2017.

Pinaki Mandal and Souvik Roy. On obviously strategy-proof implementation of fixed priority top trading cycles with outside options. *Economics Letters*, 211:110239, 2022.

Szilvia Pápai. Strategyproof assignment by hierarchical exchange. *Econometrica*, 68 (6):1403–1433, 2000.

Szilvia Pápai. Strategyproof and nonbossy multiple assignments. *Journal of Public Economic Theory*, 3(3):257–271, 2001.

Marek Pycia and M Utku Ünver. Trading cycles for school choice. Working Paper, 2011.

Marek Pycia and M Utku Ünver. Incentive compatible allocation and exchange of discrete resources. Working Paper, 2014.

Marek Pycia and M Utku Ünver. Incentive compatible allocation and exchange of discrete resources. *Theoretical Economics*, 12(1):287–329, 2017.

Marek Pycia and M Utku Ünver. Ordinal simplicity and auditability in discrete mechanism design. CEPR Discussion Papers No. 18058, 2024.

Alvin E Roth. Incentive compatibility in a market with indivisible goods. *Economics letters*, 9(2):127–132, 1982.

Alvin E Roth. Stability and polarization of interests in job matching. *Econometrica: Journal of the Econometric Society*, pages 47–57, 1984.

Alvin E Roth and Elliott Peranson. The effects of the change in the nrmp matching algorithm. *JAMA*, 278(9):729–732, 1997.

Mark A Satterthwaite and Hugo Sonnenschein. Strategy-proof allocation mechanisms at differentiable points. *The Review of Economic Studies*, 48(4):587–597, 1981.

Nicholas Schuler. CPS OIG Uncovers Widespread Admissions Irregularities in K-8 Options for Knowledge Program. Office of Inspector General, Chicago Board of Education. Press Release, February 21, 2018.

Lars-Gunnar Svensson. Queue allocation of indivisible goods. *Social Choice and Welfare*, 11(4):323–330, 1994.

Peter Troyan. Obviously strategy-proof implementation of top trading cycles. *International Economic Review*, 60(3):1249–1261, 2019.

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

Kevin Jon Williams. A reexamination of the nrmp matching algorithm. national resident matching program. *Academic medicine: journal of the Association of American Medical Colleges*, 70(6):470–6, 1995.

Kyle Woodward. Self-auditable auctions. Working Paper, 2020.