Strategyproofness-Exposing Mechanism Descriptions

Yannai A. Gonczarowski
Harvard Economics & Computer Science

Ori Heffetz
Hebrew University & Cornell Economics

Clayton Thomas
Princeton University ↦ Microsoft Research

An Annotated Abstract

Our paper's abstract in blue

Elaboration & explanation in black

Abstract:

(static, direct-revelation)

A ***menu description*** presents a mechanism to player i in two steps.

Step (1) uses the reports of other players to describe i's *menu*: the set of i's potential outcomes.

Step (2) uses i's report to select i's favorite outcome from her menu.

(Main Question)

Can menu descriptions better expose strategyproofness, without sacrificing simplicity?

First main premise of our paper:
Menu descriptions provide a way to expose strategyproofness.
Indeed, while strategyproofness might be hard to infer from traditional descriptions of some mechanisms, it always holds for menu descriptions via a one-sentence proof: player i's menu in Step (1) cannot be affected by her report, and in Step (2), straightforward reporting guarantees her favorite outcome from the menu.

To begin, note that *every* strategyproof mechanism has a menu description [Hammond, 1979]. To see this, consider a description D of the outcome of the mechanism, and consider the following "brute force" menu description for player i:

Step (1): Iterate over all possible reports t_i' of player i, and let M denote the set of all outcomes for player i of the form $D(t_i', t_{-i})$.
Step (2): Award player i her favorite outcome (according to t_i) from M.

However, we believe such descriptions are indirect, unnatural, complicated, and impractical.

⇒ **Second main premise of our paper:**
Only *simple* menu descriptions are desirable.
What counts as a simple description is naturally subjective, multi-faceted, and context-dependent. As a guiding principle, we strive for menu descriptions that are comparable in simplicity to the corresponding traditional descriptions (which are typically the simplest known way to describe the outcome). We present new descriptions are (arguably, subjectively) nearly as simple as traditional ones. Then, we propose formal simplicity conditions, and use these conditions to reason about the limits of simple menu descriptions.

Examples of Menu Descriptions (alternative presentations of static, direct-revelation mechanisms)

Median Voting:
The median voting mechanism with three voters with single-peaked preferences.

Traditional Description:	Menu Description:
The three votes will be sorted from lowest to highest, and the *middle vote* of the three will be elected.	The "obtainable candidates" will be the votes of the other two players, and all candidates between them. Out of these "obtainable candidates," the one *closest to your own vote* will be elected.

Second-Price Auction:
A single-item, sealed-bid, second-price auction.

Traditional Description:	Menu Description:
The player who placed the *highest bid* will win the item. She will pay a price equal to the *second highest bid*.	Your "price to win" the item will be set to the *highest bid* placed by any *other* player. If your bid is higher than this "price to win," then you will win the item and pay this price.

(Main Results)

(1) **We propose** a new, simple menu description of Deferred Acceptance.

(namely, Serial Dictatorship and Top Trading Cycles)

(2) **We prove** that—in contrast with other common matching mechanisms—this menu description must differ substantially from the corresponding traditional description.

(3) **We demonstrate**, with a lab experiment on two elementary mechanisms, the promise and challenges of menu descriptions.

Our main results hold for **matching mechanisms**, say with (strategic) *applicants* and (non-strategic, fixed-preference) *institutions*. We consider Deferred Acceptance (DA): the applicant-optimal stable matching mechanism. DA has many advantages, but showing its strategyproofness from its traditional description conventionally requires a delicate and technical mathematical proof. Correspondingly, unlike the elementary examples above, it is far from clear how to characterize the menu in a simple way in DA.

Our **main positive theorem** provides provides a menu description of DA.
Our new description is comparable in simplicity to the traditional one, but its strategyproofness is far easier to show.

Traditional Description of DA:	Our New Menu Description of DA:
The applicants will be matched to institutions according to the *applicant-proposing deferred acceptance* algorithm [with this algorithm explained in detail].	Imagine running *institution-proposing deferred acceptance* with all institutions and all applicants *except you*, to obtain a hypothetical matching. You "earn admission" at every institution that ranks you higher than its hypothetically matched applicant. You will be matched to the institution that you *ranked highest* out of those at which you will have earned admission.

Next, we consider the additional canonical matching of Serial Dictatorship (SD) and Top Trading Cycles (TTC).

We observe that SD's traditional description is *already* a menu description; namely, for each applicant i simultaneously, SD runs as:

(1): Each applicant $1, \ldots, i-1$, in order, is matched to her top-ranked remaining institution.
(2): Applicant i is matched to her top-ranked remaining institution.
(3): Each applicant $i+1, \ldots, n$, in order, is matched to her top-ranked remaining institution.

This three-step outline *both* exposes strategyproofness to player i, and specifies the entire matching.

Our **second positive theorem** shows that, perhaps surprisingly, TTC has a simple description with this enhanced, three-step outline. In fact, a slight modification of the traditional description of TTC, specializing the order-of-operations to applicant i, suffices to expose one applicant's menu (and hence strategyproofness).

(1): Using only the preferences of applicants other than i, match as many cycles not involving applicant i as possible, and remove all matched applicants and institutions. Let M denote the set of remaining institutions.
(2): Now, match i to i's highest-ranked institution in M.
(3): Match the cycle created when i points to the institution from (2), and continue matching cycles until all applicants are matched.

Very briefly, our **impossibility theorems** prove:

(a): In a very strong sense, something like the above **three-step outline** for TTC is **impossible for DA**. In other words, it is impossible to find a menu description of DA within (a small tweak of) its traditional description.
(b): Simple descriptions of DA, as captured by a somewhat more restrictive formal condition than considered in (a), face a tradeoff: they **can convey strategyproofness** (with our new menu description); they **can convey feasibility**, i.e., that the outcome matching is one-to-one (with the traditional description); but they **cannot convey both**.

We conducted a preregistered, between-subjects lab experiment using the two pairs of descriptions in the elementary examples above.

Median Voting: We find a significant increase in rates of participants playing their dominant strategy: (70%; $N = 100$) under Traditional and (80%; $N = 100$) under Menu (equality-of-means $p = 0.01$). Furthermore, in Menu (but not Traditional), dominant strategy play is highly correlated with participants' *comprehension* of the mechanism. This may suggest that for the menu description of this mechanism—but not for the traditional description—understanding how the outcome is calculated drives an increased understanding of strategyproofness.

Second Price Auction: In contrast, here we find no significant difference in play between the two treatments. This may suggest that for some mechanisms, strategyproofness may be equally apparent from traditional and menu descriptions

Median Voting: Rates of dominant-strategy play by treatment.