American Economic Review: Papers & Proceedings 2017, 107(5): 225–229
https://doi.org/10.1257/aer.p20171028

Mistaken Play in the Deferred Acceptance Algorithm: Implications for Positive Assortative Matching[†]

By Alex Rees-Jones*

In their seminal work, Gale and Shapley (1962) presented the deferred acceptance algorithm (DAA), an algorithm that has subsequently motivated a large literature on mechanism design in two-sided matching markets. When utilizing this algorithm, two groups of market participants—call them *students* and *schools*—both submit their preferences over potential matches to a neutral intermediary. In each step of the algorithm, students are assigned to the most preferred school that has yet to reject them. Each school then assesses if it has admitted more students than its quota; if it has, it rejects its least preferred matched students until the quota is no longer violated. The algorithm iterates until all students have either been matched or been rejected from all ranked schools.

By applying this procedure, a market designer may avoid a common problem faced in matching markets: strategic misrepresentation of preferences. In many matching environments, ranking match partners in order of their desirability can be suboptimal. A sensible student might worry that all seats at an attainable school could be filled in the time spent being considered by more desirable, but unattainable, programs. These concerns are not present in the DAA. As established in Dubins and Freedman (1981) and Roth (1982), the DAA is *strategy-proof*: truthful preference reporting is a weakly dominant strategy for students. This feature is commonly viewed as especially appealing in the student-to-school matching setting, as it avoids the punishment of students who sincerely report their preferences without regard to strategic incentives.

*The Wharton School, University of Pennsylvania, 553 Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104 (e-mail: alre@wharton.upenn.edu).

[†]Go to https://doi.org/10.1257/aer.p20171028 to visit the article page for additional materials and author disclosure statement.

Despite this desirable theoretical property, misrepresentation of preferences appears to persist in the DAA. Lab experiments commonly reveal a sizable fraction of students making mistakes—that is, pursuing the dominated strategy of misrepresenting their preferences (see, e.g., Chen and Sönmez 2006; Pais and Pintér 2008; Calsamiglia, Haeringer, and Klijn 2010; Klijn, Pais, and Vorsatz 2013; and Featherstone and Niederle 2016). Outside of the lab, attempts at futile strategic misrepresentation are seen in both the Israeli Psychology match (Hassidim, Romm, and Shorrer 2016) and in the US Medical Residency Match (Rees-Jones 2016).

How might a market designer evaluate the consequences of these mistakes? While a large literature has emerged to provide a theoretical framework for assessing the outcomes that the DAA generates, this literature is built on the premise of optimal play. Little guidance currently exists on the consequences of relaxing this assumption. By definition, mistakes harm those who make them. However, the broader impact of these mistakes to aggregate social welfare remains an open question.

In this paper, I discuss the implications of these mistakes for *positive assortative matching* (PAM)—that is, the ability of the matching mechanism to sort the best students to the best schools. I demonstrate that, when student quality is imperfectly observed, the presence of these mistakes can facilitate PAM. In cases where PAM is socially valued, the welfare impact of mistakes can be critically determined by the signal of student quality that mistakes convey. Welfare losses can be severe if mistakes are most common among the best students. However, the presence of mistakes can be dramatically welfare-enhancing if mistakes are most common among the worst students. I proceed by presenting a simple example to build intuitions, and then by presenting a simulation study that calibrates the quantitative effects. I conclude by

Table 1—A Simple Example

Students' rank-orders		Schools' rank-orders		Probability of positive
Student A	Student B	A $\succ$ B	B $\succ$ A	assortative match
$1 \succ 2$	$1 \succ 2$	$\boxed{\{(A,1),(B,2)\}}$	$\{(A,2),(B,1)\}$	p
$1 \succ 2$	$2 \succ 1$	$\boxed{\{(A,1),(B,2)\}}$	$\boxed{\{(A,1),(B,2)\}}$	1
$2 \succ 1$	$1 \succ 2$	$\{(A,2),(B,1)\}$	$\{(A,2),(B,1)\}$	0
$2 \succ 1$	$2 \succ 1$	$\{(A,2),(B,1)\}$	$\boxed{\{(A,1),(B,2)\}}$	$1-p$

discussing the currently available field evidence on these mistakes' signal of student quality, and by highlighting other relevant welfare considerations that I have abstracted from in this argument—perhaps most importantly, fairness and equity concerns.

I. A Simple Example

To illustrate the potential for preference misrepresentation to facilitate PAM, let us begin with the simplest possible example. Consider a case where the market designer must match two students (A and B) to two schools (1 and 2). School 1 is better than school 2—a feature that is common knowledge—and both schools have a single position available. Student A is better than student B; however, this feature is *not* common knowledge. Instead, schools receive an imperfect signal of student ability, like those that are generated from standardized tests, grades, or letters of recommendation. Both schools receive the same signal and use it to determine their preferences over students. This signal correctly rank-orders the students with probability p.

In this exercise, matching A to 1 and B to 2 will be called the positive assortative match (denoted M^*), and is assumed to be preferable to matching A to 2 and B to 1. We will proceed by characterizing how the probability of achieving M^* varies depending on the reporting patterns of students. To begin, Table 1 reports the final match that will be realized depending on the signal the schools have received and the rank-orders submitted by the two students. Cases in which M^* is achieved are boxed.

In the case where the schools receive a correct signal about student quality, and thus submit the rank order $A \succ B$, notice that the presence of misrepresentation can only disrupt the match outcome. When both students submit their true preference ordering, M^* is achieved. However, if student A misrepresents his preference ordering, his request to match to school 2 will be granted and M^* is not achieved.

In the case where the schools receive an incorrect signal about student quality, and thus submit the rank order $B \succ A$, notice that the presence of misrepresentation can only improve the match outcome. When both students submit their true preference ordering, student B is erroneously given priority over student A due to the high signal of ability, and thus student B is matched to school 1. This outcome may only be avoided in the case where student B subsequently misrepresents his preferences.

Following this logic, the consequences of misunderstanding will be determined by the probability that the school's initial ranking was correct (p) as well as the relative propensity of both student A and student B to misrepresent their preferences (denoted ℓ_A and ℓ_B). It can be quickly verified that the probability of achieving M^* is determined by the equation

$$\begin{aligned}\Pr(M^*) = {} & (1-\ell_1)\cdot(1-\ell_2)\cdot p \\ & + (1-\ell_1)\cdot\ell_2\cdot 1 \\ & + \ell_1\cdot(1-\ell_2)\cdot 0 \\ & + \ell_1\cdot\ell_2\cdot(1-p).\end{aligned}$$

Absent misrepresentation, the probability of M^* is simply p, the probability that the schools correctly rank the students. Compared to this

baseline, $\Pr(M^*)$ is improved if and only if $\frac{\ell_2}{\ell_1} > \frac{p}{1-p}$.

These calculations illustrate the considerations that determine the link between preference misrepresentation and PAM. If preference misrepresentation is sufficiently associated with student quality—i.e., if bad students are sufficiently more likely to misrepresent their preferences—then the presence of misrepresentation can improve PAM, and thus social welfare in environments where PAM is valued. The necessary degree of association is determined by the strength of the other signals that the schools receive. Given an uninformative signal of student quality $\left(\frac{p}{1-p} = 1\right)$, the presence of misrepresentation is helpful so long as $\ell_2 > \ell_1$. As schools' signal of quality becomes more discerning, the relative propensity of the low-quality student to misrepresent must grow for the presence of misrepresentation to be socially valued.

II. Effect Sizes in Simulated Markets

To illustrate the potential for quantitatively important interactions between misrepresentation and PAM, I conduct a simulation exercise while varying features of these elements.

In these simulations, I match 100 students to 10 schools, each with a quota of 10 students. Schools are indexed by i, with school quality expressed on the unit interval as $q_i = \frac{11-i}{10}$. Students observe school quality, and prefer programs with higher quality values. Schools similarly attempt to form their preferences according to a ranking of student quality, where the true quality measure of student j is $q_j = \frac{101-j}{100}$. However, schools only receive an imprecise signal of student quality, and rank students according to that signal.[1]

Across simulations, I vary two features of these markets. First, I consider alternative assumptions on the fraction of students who misrepresent their preferences. Second, I consider alternative assumptions on the relationship between misrepresentation and student quality. Across three regimes, I model misrepresentation as pursued either by the highest quality students, by the lowest quality students, or by students selected at random. These regimes span the range of the potential for misrepresentation to signal student quality, and thus comparisons across these regimes may illustrate how the welfare effects of a given level of misrepresentation may vary depending on different assumptions on that parameter.

For each "percentage misrepresenting preferences" of the set $\{0, 10, 20, 30, 40, 50\}$, and for each of the three regimes associating misrepresentation to student quality, I simulate the market 10,000 times. Across all combinations, this results in 180,000 simulated markets.

The effects of misrepresentation on PAM are summarized in Figure 1, which plots the average quality of students matched to each school under the differing assumptions described above. For comparison, the dashed lines illustrate the maximal amount of PAM possible, as would be achieved with optimal behavior and perfect observation of student quality. Examining panel A, we confirm a similar intuition as was seen in the simple example: when misrepresentation signals low student quality, its presence in the market facilitates PAM. As seen in the figure, the inclusion of progressively more misrepresentation in this market leads to better quality students being matched to better quality schools. In contrast, when misrepresentation is made at random, and is thus unassociated with student quality, it harms PAM: higher rates of misrepresentation lead to lower student quality at the best schools and higher student quality at the worst. A similar, but substantially more quantitatively pronounced, pattern is observed when misrepresentation is pursued by the best students.

To assess how these effects might translate into social welfare calculations, I estimate a social welfare function over final pairings achieved. I assume that $W = \sum_{matched(i,j)} q_i \cdot q_j$—a specification that inherently values PAM—then normalize W by its average value when misrepresentation is not present. In Figure 2, I plot how social welfare evolves as increasing percentages of students misrepresent their preferences. To ease interpretation, I additionally demarcate the welfare values associated with two policy experiments. The higher (lower) dashed line delineates the average welfare associated with a 10 percent increase (decrease)

[1] Specifically, schools rank students based on the ordering of $q_j + \epsilon$, where $\epsilon \sim N(0, 1)$.

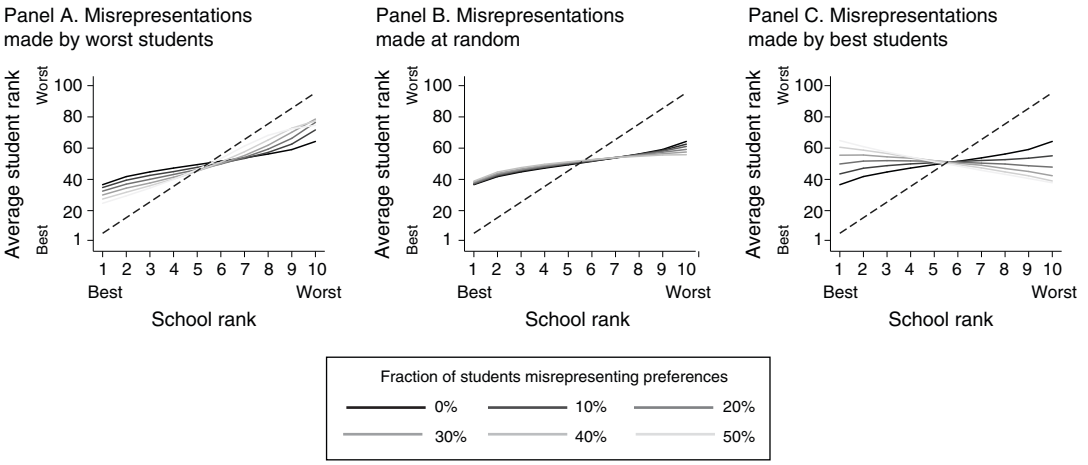

Figure 1. Positive Assortative Matching When Misrepresentation Is Present

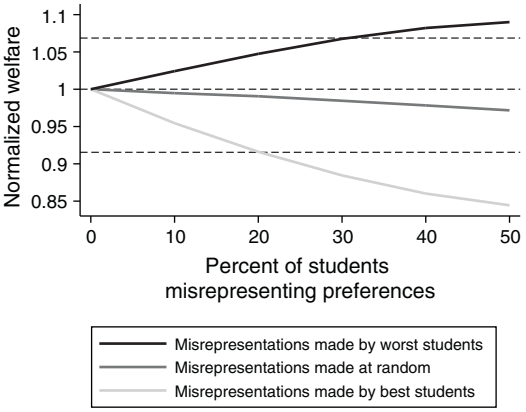

Figure 2. Misrepresentation and Welfare

in all program quotas when misrepresentation is not present. Welfare effects of these magnitudes may sensibly be considered large in most applications.

As illustrated in Figure 2, the welfare effect of a given level of misrepresentation varies substantially depending on the assumed relationship between misrepresentation and student quality. When mistakes are pursued by the lowest quality students, welfare of the market is improved—dramatically so when misrepresentation is prevalent. When mistakes are pursued by the highest quality students, welfare of the market is hindered—again, dramatically so when misrepresentation is prevalent. When misrepresentation is unassociated with student quality, the net impact is negative, although the quantitative importance is relatively minor compared to other regimes.

In summary, holding fixed a level of misrepresentation, different assumptions on the association between that misrepresentation and student quality will dramatically change welfare conclusions.

III. Discussion

Whether misrepresentation in the DAA helps or hinders PAM remains an open question, and critically depends on the signal value inherent in observing these mistakes. Few would insist that the best first-graders must be sophisticated game theorists. However, among the psychology and medical student populations studied in Hassidim, Romm, and Shorrer (2016) and Rees-Jones (2016), better students appear to be less likely to make mistakes. These results are tempered by Guillen and Hakimov's (2016) finding that optimal play in the closely related top trading cycle mechanism is more influenced by the provision of advice on optimal play than by assistance in understanding the algorithm—a finding that suggests a more limited potential for mistakes to serve as a signal. As this literature progresses, continued attempts to measure this association in the various contexts to which the DAA is applied will prove necessary.

While this paper is motivated by a desire to better understand the full welfare evaluation of mistakes, I caution the reader that I have focused attention on only a single element of welfare, and have made significant simplifying assumptions to clearly isolate PAM effects. My examples abstract from heterogeneity either in student's evaluations of schools or in schools' evaluation of students. I assume that schools do not differ in their ability to rank student quality, and do not deviate from truthful reporting themselves. I impose strong assumptions on the manner in which students misrepresent their preferences—a crucial component of these models that, like the association of mistakes with student quality, requires more measurement. Finally, I abstract entirely from considerations of fairness or equity. The abandonment of non-strategy-proof algorithms is often partially motivated by concerns that low-income or otherwise-disadvantaged students received little, or bad, guidance on optimal play. If suboptimal play persists in the DAA, similar concerns remain. All of these considerations merit careful evaluation in a complete welfare analysis of these markets.

As concerns of mistaken play persist, I encourage theoretical attention to these issues. However, until the relevant components have been measured and integrated into more complete welfare analysis, attempts to "nudge" in this environment must be pursued with caution; as demonstrated here, a well-intentioned nudge could be significantly socially harmful.

REFERENCES

Calsamiglia, Caterina, Guillaume Haeringer, and Flip Klijn. 2010. "Constrained School Choice: An Experimental Study." *American Economic Review* 100 (4): 1860–74.

Chen, Yan, and Tayfun Sönmez. 2006. "School Choice: An Experimental Study." *Journal of Economic Theory* 127 (1): 202–31.

Dubins, Lester, and David Freedman. 1981. "Machiavelli and the Gale-Shapley Algorithm." *American Mathematical Monthly* 88 (7): 485–94.

Featherstone, Clayton, and Muriel Niederle. 2016. "Boston versus Deferred Acceptance in an Interim Setting: An Experimental Investigation." *Games and Economic Behavior* 100: 353–75.

Gale, David, and Lloyd Shapley. 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly* 69 (1): 9–15.

Guillen, Pablo, and Rustamdjan Hakimov. 2016. "How to Get Truthful Reporting in Matching Markets: A Field Experiment." Social Science Research Center Berlin (WZB) Discussion Paper SP II 2015–208.

Hassidim, Avinatan, Assaf Romm, and Ran Shorrer. 2016. "'Strategic' Behavior in a Strategy-Proof Environment." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784659 (accessed January 12, 2017).

Klijn, Flip, Joana Pais, and Marc Vorsatz. 2013. "Preference Intensities and Risk Aversion in School Choice: A Laboratory Experiment." *Experimental Economics* 16 (1): 1–22.

Pais, Joana, and Agnes Pínter. 2008. "School Choice and Information: An Experimental Study on Matching Mechanisms." *Games and Economic Behavior* 64 (1): 303–28.

Rees-Jones, Alex. 2016. "Suboptimal Behavior in Strategy-Proof Mechanisms: Evidence from the Residency Match." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2662670 (accessed January 12, 2017).

Roth, Alvin. 1982. "The Economics of Matching: Stability and Incentives." *Mathematics of Operations Research* 7 (4): 617–28.