An experimental investigation of preference misrepresentation in the residency match

Alex Rees-Jones[a,b,1] and Samuel Skowronek[a]

[a]Operations, Information, and Decisions Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104; and [b]National Bureau of Economic Research, Cambridge, MA 02138

Edited by Alvin E. Roth, Stanford University, Palo Alto, CA, and approved September 24, 2018 (received for review February 21, 2018)

The development and deployment of matching procedures that incentivize truthful preference reporting is considered one of the major successes of market design research. In this study, we test the degree to which these procedures succeed in eliminating preference misrepresentation. We administered an online experiment to 1,714 medical students immediately after their participation in the medical residency match—a leading field application of strategy-proof market design. When placed in an analogous, incentivized matching task, we find that 23% of participants misrepresent their preferences. We explore the factors that predict preference misrepresentation, including cognitive ability, strategic positioning, overconfidence, expectations, advice, and trust. We discuss the implications of this behavior for the design of allocation mechanisms and the social welfare in markets that use them.

behavioral economics | mechanism design | matching | lying | deception

People often have strong incentives to lie about their preferences. These incentives are unfortunate, since market organizers must commonly make decisions based on the preferences that individuals report. Auction prices are often determined based on bids, but potential buyers may not bid their true valuation. Employees are often hired based on interviews, but job seekers may feign interest for the positions available. Students are often assigned to schools based on reported school preferences, but applicants may be incentivized to list an attainable school as their favorite. In environments like these, economists have devoted substantial effort to mitigating this problem by designing "strategy-proof" mechanisms that render truthful preference reporting incentive compatible. With such a mechanism in place, market participants who understand how outcomes are determined will see that there is no benefit to lying.

A growing body of evidence suggests that individuals misrepresent their preferences in incentive-compatible environments despite the futility of such efforts. Imperfect truth telling has been documented in laboratory experiments studying sealed-bid and clock auctions (1), in willingness-to-pay elicitations (2), and in applications of school-choice matching mechanisms (3–7). This work has informed recent theoretical advances aimed at characterizing mechanisms that are "obviously strategy-proof" to relatively unsophisticated decision makers (8). In many contexts, attendance to this criterion yields comparatively easy-to-understand mechanisms; however, in the context of stable two-sided matching mechanisms, no obviously strategy-proof options exist (9). An immediate implication is that, in matching environments where stability is required, we must rely on a degree of sophistication in market participants for optimal behavior to emerge.

Particularly in the context of student matching markets, these findings can be viewed as troubling. A key argument motivating the adoption of strategy-proof school-choice mechanisms is that they "level the playing field" (10). In algorithms with a nontruthful optimal strategy, strategically savvy—and disproportionately affluent—students are given an undue advantage at the expense of students who report their preferred schools truthfully. If strategy-proof mechanisms result in all participants reporting truthfully, this undesirable outcome is averted. However, if the inability to understand optimal strategies extends to cases where the optimal strategy requires no "gaming" of the system, an unleveled playing field remains. Understanding the prevalence and correlates of such mistakes then becomes crucial for assessing the fairness, and indeed the broader welfare consequences, of the allocations that these mechanisms generate.

Unfortunately, directly assessing the prevalence and correlates of preference misrepresentation is fundamentally challenging. In the field settings where these mechanisms are adopted, preferences are unobservable. Absent observing true preferences, the veracity of reported preferences cannot be directly assessed.* Experimenters have sidestepped this difficulty in the laboratory by using simplified matching scenarios to assign preferences. However, by restricting empirical investigations to the laboratory, such work can only document suboptimal behavior in unfamiliar and minimally incentivized tasks completed by populations different from the ones facing these mechanisms in the field. On the one hand, these external validity concerns potentially mitigate the worry that the observed failure of optimal reporting extends to the policy applications of primary concern. On the other hand, if misrepresentation persists in populations whose lives are affected by their performance in these mechanisms, the design and deployment of these mechanisms may require considerable revision.

Significance

Policymakers increasingly rely on matching algorithms to assign students to schools. Common algorithms can be "gamed" by students misrepresenting their preferences for schools, resulting in assignments that are unduly influenced by application strategies. In strategy-proof algorithms that incentivize students to tell the truth, this undesirable influence of strategic sophistication is argued to be eliminated. We conduct an online experiment among participants in a leading exemplar of strategy-proof market design: the assignment of new doctors to medical residencies. Our results suggest that many market participants do not understand that telling the truth is optimal. This illustrates that strategy-proof environments are not immune to the influence of strategic sophistication, and that practical tensions arise when using complex means to implement simple incentives.

Author contributions: A.R.-J. and S.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Data and program files are available at https://osf.io/tp6h5/.

[1]To whom correspondence should be addressed. Email: alre@wharton.upenn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1803212115/-/DCSupplemental.

*Despite this difficulty, some attempts to assess rates of truth telling in field settings have been made. To sidestep the difficulty of observing true preferences, researchers have relied on either unincentivized survey reports of self-assessed truthful behavior (11) or have examined specific types of reported preferences that are so anomalous that they cannot be plausibly explained by preference heterogeneity (12–14).

In this study, we aim to achieve the benefits of the laboratory-experimental approach to detecting failures of truth telling while simultaneously studying the behavior of a highly incentivized and highly trained population of direct policy relevance. We deploy a large-scale online experiment to 1,714 medical students participating in the 2017 National Resident Matching Program (NRMP), a system in which graduating medical students submit their preferences for residency programs to be used to determine their placements. The NRMP utilizes a modified version of the deferred acceptance algorithm (15, 16), a matching mechanism that is strategy-proof for students and is increasingly adopted for school assignment (17). The NRMP constitutes a flagship application of matching theory and remains one of the most carefully designed and extensively studied two-sided matching markets in existence.[†] Our online experiment puts NRMP participants through a simple incentivized matching task in which truth telling can be easily assessed. By deploying this study immediately after the NRMP match, and by transparently applying the same mechanism used by the NRMP, we are able to directly study the prevalence and correlates of preference misrepresentation in the precise population of interest.

We document widespread failure to pursue the incentivized strategy of truth telling. Over 23% of experimental participants misrepresent their preferences in our matching task, despite using this mechanism to make a career-altering decision mere days before.

We additionally examine the predictors of misrepresentation, shedding light on both the factors that contribute to this behavior and the features of individuals who bear the costs. The tendency for suboptimal behavior is associated with both the strength of the students' strategic positions (measured by randomly assigned test scores in the matching task) and by the students' cognitive reasoning abilities (measured by Raven's Matrices deployed after the matching task). Beyond metrics associated with student quality, the tendency for suboptimal behavior is associated with students' overconfidence, with the pursuit and availability of advice in the lead up to the NRMP match, and with students' trust in residency programs to rank students according to quality. These results identify the individuals who gain and lose from the complexity of the existing system, give guidance on the best practices for training market participants to engage with complex mechanisms, and critically inform the study and design of matching markets. We further discuss these implications in *Discussion*.

Study Population and Sample Recruitment

We solicited participation in our study by recruiting medical schools to present our recruitment materials to their NRMP participants, following recruitment protocol derived from previous survey investigations of medical students (19). To do so, we contacted representatives of all 147 medical schools accredited by the Association of American Medical Colleges (AAMC) located in the United States and Puerto Rico. As a result of our initial outreach and subsequent follow-up, we were able to successfully recruit 25 medical schools (*SI Appendix*, Fig. S1 and Table S1). These 25 schools vary widely in class size [minimum (min) = 41, maximum (max) = 328], location, and competitiveness. Compared with the full population of accredited medical schools, we find no statistically significant differences between participating and nonparticipating schools on total enrollment, average Medical College Admission Test (MCAT) performance, average undergraduate grade point average, acceptance rates, US News and World Report research rankings, or gender composition (*SI Appendix*, Table S2).

[†]Independent of its relation to the mechanism design literature, the NRMP is of intrinsic importance to a large labor market. In 2017 alone, the NRMP received 35,696 preference lists from applicants vying for 31,757 positions (18).

Shortly after the deadline for submission of residency preferences to the NRMP, participating schools forwarded our recruitment email to their graduating student body. This email asked students to participate in an anonymous 10-min survey about decision making in the NRMP match. Students were further told that they would earn an Amazon.com gift card valued between $5 and $50 with an expected value of $21 for participating in the survey. All data were collected before the NRMP's announcement of the results of the match.

Approximately 3,300 graduating medical students (17.1% of all graduating medical students from AAMC accredited schools) received an email with our survey link. Participant demographics are summarized in *SI Appendix*, Table S3. Our analysis is based on the 1,714 students (~51.9% of the students contacted) who both completed the survey and passed all exclusion criteria (*SI Appendix*, Fig. S2 and Table S4).

Experimental Design

All experimental materials are presented in the *SI Appendix*; we summarize the key measures below. Our materials were reviewed by the University of Pennsylvania Institutional Review Board (IRB) and ruled exempt from IRB review [as authorized by 45 CFR 46.101(b), category 2]. Informed consent was elicited on the first page of the web survey.

Incentivized Matching Task. Participating students were presented with an incentivized matching task. The prompt for this task explained: "In this exercise, you will go through a matching process much like the NRMP match. You will attempt to match to one of five hypothetical residency programs, and the payment you receive for taking this survey will depend on where you match. We will apply the standard algorithm that was used by the NRMP; as a reminder, an example of how this algorithm works is available here." The underlined term hyperlinked to NRMP training materials. Since students receive significant training and advice regarding this algorithm in the lead-up to participating in the NRMP match, we did not elaborate further on the functioning of this mechanism.

In each simulation, 50 students applied to five residency programs, each with 10 positions available. The preferences of both the programs and the other students are simulated according to guidelines communicated to the participant. We explained that all students agree on the same ranking of residency programs. We also explained that residency programs based their preferences on several factors, with students' Hypothetical Standardized Test (HST) scores being an important one. Based on the manner in which programs' preferences were simulated, every student had some possibility of matching to every program. This renders nontruthful preference reporting a strictly suboptimal strategy for maximizing expected payoff.

To communicate the desirability of different residency programs, participants were presented with a table (Fig. 1). For each program, this table reported both the average HST score of the admitted students and the value of the Amazon.com gift card that participants would receive if they matched. Participants were also told that they would earn $5.00 if they did not match to any program. The payment received from this matching process was the sole compensation provided for participation.

After this explanation of the matching task, students submitted their rank-order list (ROL) using a series of dropdown menus. Participants were told that they must apply to at least one program but could forego latter applications if they wished.

We will refer to ROLs that list all five residencies in order of their compensation as "optimal" or "truthful," and those that do not as "suboptimal" or "misrepresented." This labeling relies on the assumption that participants prefer more money to less. While that assumption is both standard and reasonable, under some conditions it could fail. For example, failure could arise if

Rank	Residency program	Average HST percentile score of admitted students	Amazon.com gift card value
1.	Maplecrest	80th percentile	$50.00
2.	Birch Hill	65th percentile	$25.00
3.	Elm South	50th percentile	$15.00
4.	Hickory Bridge	35th percentile	$10.00
5.	Pine Peak	20th percentile	$7.50

Fig. 1. Residency information for simulated residency match. This table was displayed to participants to communicate the desirability of different programs. The desirability was communicated in two ways: first, by the average HST scores of students admitted to each residency; and second, by the value of the gift card that participants would earn by matching to that program.

subjects prefer to earn less money because they value leaving money to the experimenter or to the simulated students that they compete against. We consider this possibility unlikely. Failure could also arise if subjects value not only the monetary payoffs but also anticipation or disappointment. We further discuss this latter possibility in our tests of possible correlates below. While it is necessary to rule out nonstandard preferences to ensure that misrepresented ROLs identify confusion about incentives, misrepresentation stemming from either source would be viewed as anomalous from the perspective of standard matching theory.

Correlates of Suboptimal Reporting. We preregistered our interest in five groups of correlates of suboptimal reporting, all proposed and discussed in prior literature (for a summary, see ref. 20). Not all of the variables that we examine are experimentally manipulated, and consequently not all analyses can be interpreted as estimating causal relationships. However, some of the associations help distinguish between potential factors driving the suboptimal behavior of interest. Furthermore, different predictors of misrepresentation suggest different welfare costs of this behavior, and the necessary approaches to reduce it. We motivate each factor of interest below and explain its measurement in the context of our study.

Student quality. The welfare consequences of misrepresentation can be significantly influenced by its correlation with student quality (21). Two distinct channels, conflated in the field but separable in our experiment, may generate such a correlation. First, students with comparatively low grades or test scores are often placed at a strategic disadvantage for obtaining a desirable match. This might result in attempts to misrepresent preferences as a means to compensate, or might lead students to fail to list desirable programs under the belief that they are unobtainable. Second, students in this position might also have comparatively low cognitive ability, which increases the probability of incorrectly identifying the optimal strategy in laboratory experiments utilizing this algorithm (22). Our experiment contains measures that allow us to study each channel separately.

To examine the impact of strategic positioning, participants were randomly assigned an HST percentile score. This score influenced each participants' ranking in residency program preferences, and thus their strategic position.

To examine the impact of cognitive ability, we presented participants with a test of spatial reasoning. We gave participants 5 min to complete seven Advanced Raven's Progressive Matrices (23), a test widely used to assess logical reasoning ability (24). Of course, medical students with low cognitive ability relative to their peers likely have substantially higher average cognitive ability than typical populations facing matching mechanisms (e.g., school children and their parents). Care is warranted when extrapolating our results onto other such populations.

Overconfidence. Overconfidence is a prevalent trait among physicians (25) and is commonly thought to broadly generate decision errors (26). Furthermore, recent research demonstrates that this bias affects suboptimal reporting in the related, but gameable, Boston mechanism (27). We generate a measure of overconfidence in the course of conducting our test of logical reasoning ability. After completing the Raven's Matrices, participants were asked to think about other medical students participating in this survey and to estimate the percentage of participants that they outperformed (slider scale = min: 0%, max: 100%). We code participants as overconfident if their forecast of their performance exceeds their actual percentile rank—in the language of Moore and Healy (28), this is a measure of overplacement. A secondary, but similar, measure of overconfidence is available from students' report and assessment of their MCAT performance. Participants were asked to report their MCAT score and then estimate the percentage of other MCAT takers who received a lower score than they received in the year that they took the MCAT (slider scale = min: 0%, max: 100%).

Desire to rank the expected outcome highly. If students derive utility from the anticipation of matching to a program that they rank highly, or if they expect to experience disappointment from matching to a program that they did not rank highly, then students may be motivated to submit nontruthful preference orderings that manage these anticipations. In this case, misrepresentation need not be irrational: in the presence of such belief-based utility functions, the deferred acceptance algorithm is not strategy-proof.

We test for the influence of expectations on misrepresentation by randomly varying the salience of the participants' expected match before they submit their ROLs. Before proceeding to the submission page, we randomly assigned half of the participants to indicate the residency where they expected to match. We reminded them of their expected match in the list submission prompt.

Pursuit and availability of advice. When mechanisms are sufficiently difficult to understand, participants may be significantly influenced by advice (or their tendency to seek it) (29, 30). To examine the role of advice, we requested that participants check all of the sources that provided them with advice regarding their NRMP submission from the following list: (*i*) current and/or past medical students who participated in the NRMP, (*ii*) participant's medical school, (*iii*) the NRMP website, and (*iv*) other sources. Participants then specified the advice they received from each entity in a free-response text box and rank ordered them based on the level of influence each had on their NRMP submission.

Mistrust of other market participants. In many mechanisms, a particular action (such as truth telling) may be an optimal strategy if and only if all other market participants similarly pursue optimal play. Note that this is not the case in the deferred acceptance algorithm that underlies the NRMP matching algorithm: truth telling is optimal regardless of the action of other market participants (31, 32). However, if participants misunderstand this distinction, or if they harbor mistrust of other market participants that leads them to doubt the credibility of the matching agency, suboptimal behavior could arise (33).

We asked participants whether they trusted the players in the NRMP matching market. Participants indicated (*i*) whether they trusted that the residencies that they rank ordered in their NRMP submission would rank order medical students based on a truthful assessment of their quality, (*ii*) whether they trusted other medical students to submit a truthful rank ordering of their preferences to the NRMP, and (*iii*) whether they trusted the NRMP to run the matching algorithm honestly (all questions, 1 = yes, 0 = no).

Results

We examine the data in three stages. First, we catalog the various ways participants submitted their ROL of the residency programs in the simulated match and document the monetary consequences of suboptimal behavior. Second, we provide evidence that behavior in our experiment is associated with known proxies for misunderstanding in the NRMP match. Third, we examine the correlates of suboptimal behavior.

Documenting Suboptimal Behavior. To apply optimally in the incentivized matching exercise, participants must rank residencies in order of their monetary value. Applications are suboptimal if participants shorten their ROL by not ranking all programs or permute their ROL by ranking their listed programs in an order that does not reflect the monetary payoffs.

We find that 23.3% (n = 399) of participants applied suboptimally. As shown in Fig. 2, 64.7% of participants who submitted a suboptimal ROL permuted the list of residency programs (15.1% of total N) but applied to all five programs, 28.3% (6.6% of total N) shortened their ROL, and 7.0% (1.6% of total N) both shortened and permuted their ROL. (See *SI Appendix*, Figs. S3 and S4 for analysis of the programs removed and misordered in suboptimal ROLs.)

Failure to submit the optimal ROL was costly. Participants who submitted a suboptimal ROL earned $18.20 on average, 21.2% less than the average earnings of participants who submitted an optimal ROL, $22.80 ($t$ = −5.43, P < 0.001). However, this difference cannot be entirely attributed to the effect of misrepresentation because participants' assigned HST scores affect both their earnings and their propensity to misrepresent preferences. As we document in *Examining the Correlates of Suboptimal Behavior*, misrepresentation becomes less common among students assigned comparatively high HST scores. The rate of misrepresentation varies between 28.6% in the second lowest decile and 14.0% in the second highest decile. The solid lines in Fig. 3 show that the average difference in experimental earnings between optimal and suboptimal participants is most dramatic for those assigned a comparatively high HST score, but persisted across the distribution of assigned strategic positions (for statistical tests, see *SI Appendix*, Table S5; for assessment of the rate of costly misrepresentation at the individual level, see *SI Appendix*, Fig. S5). While all students in the experiment are incentivized to truthfully report preferences, these results illustrate that the strength of incentives varies based on the student's position in the market. This variation in incentives is a key feature of this class of matching problems and a possible channel driving the hypothesized association between misrepresentation and student ability. A desirable student has a strictly larger set of possible match partners, which results in larger differences between the best and the worst outcomes that are possible from different reporting strategies.

Validation of Experimental Behavior. We conduct three validation exercises to confirm that behavior in our experiment proxies for misunderstanding of incentives in the real residency match.

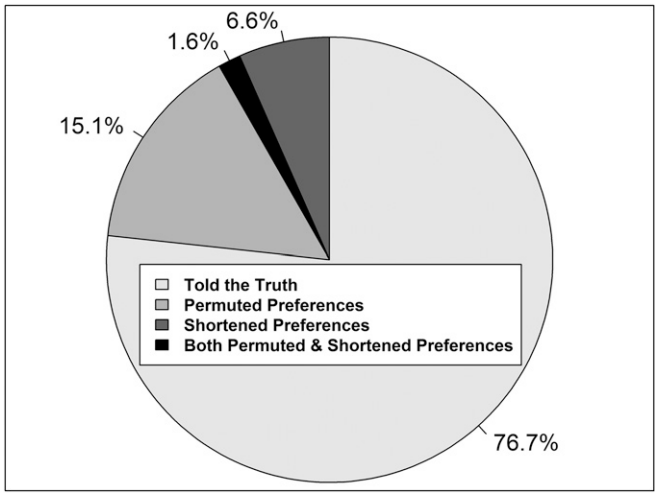

Fig. 2. Classification of truth-telling behavior.

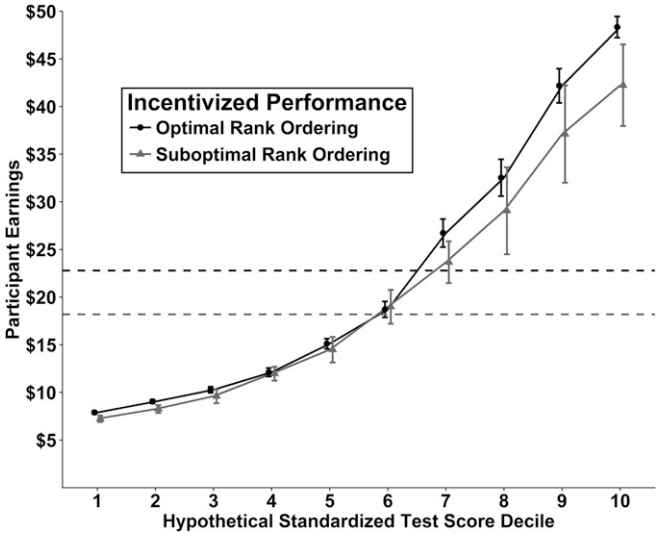

Fig. 3. Monetary losses associated with suboptimal preference reporting. This figure summarizes average experimental earnings as a function of both truth-telling status (optimal versus suboptimal rank ordering) and participants' randomly assigned test scores. The dashed lines represent the overall average earnings for participants who submitted suboptimal ($18.20) and optimal ($22.80) rank orderings. The solid lines denote average earnings within each decile of assigned test scores. Vertical lines at each point show 95% confidence intervals. See *SI Appendix*, Table S5 for statistical comparisons.

First, we test for differences in the rate of misrepresentation in our experiment as a function of self-reported truth-telling status in the NRMP. We find that students who report misrepresenting preferences to the NRMP are 9.4 percentage points more likely to misrepresent preferences in our experiment (22.6% vs. 33.0%, χ^2 = 6.19, P = 0.013).

Second, we test the correlation between misrepresentation in our experiment and the propensity for students to submit comparatively short preference lists to the NRMP. Short lists are a known proxy for suboptimal preference reporting and are actively discouraged in NRMP training materials (34). We regressed participants' likelihood to shorten their experimental ROL (1 = shortened, 0 = not) or to permute their experimental ROL (1 = permuted, 0 = not) on the number of programs participants ranked in their NRMP submission. We find that participants who submitted either shortened or permuted ROLs submitted shorter ROLs to the NRMP (shortened: B = −0.78, SE = 0.120, P < 0.001; permuted: B = −0.19, SE = 0.089, P = 0.038) (see *SI Appendix*, Fig. S6 for details).

Third, we examine differences in truth-telling rates across students who do, and do not, expect to match to their top-ranked program in the NRMP match. We find that participants who expected to match to their top NRMP match choice (n = 1,157; 67.5% of sample) were significantly less likely to submit an optimal ROL in the incentivized exercise (75.1%) compared with participants who did not hold this expectation (80.1%) (χ^2 = 5.19, P = 0.023). This result is consistent with our measure capturing a belief that optimal strategies involve strategically ranking attainable schools highly, a key component of optimal strategies in related, but manipulable, mechanisms (e.g., the Boston mechanism).

In summary, our experimental measure validates well with proxies for suboptimal preference reporting in the field.

Examining the Correlates of Suboptimal Behavior. Fig. 4 summarizes the full battery of tests of the correlates of suboptimal preference reporting. Plotted are estimated average marginal effects (AMEs) derived from a logit model predicting the outcome of submitting a truthful preference ordering. Fig. 4*B* presents the estimate for each

univariate model, predicting truth telling with only the single variable represented in that row. These results provide guidance on the features of students who do, or do not, face difficulties in pursing the optimal strategy. Fig. 4*A* presents estimates obtained from the complete model, including the entire battery of predictors. These provide clearer guidance of the role of each considered correlate, holding all else equal. We normalize all continuous variables in this analysis, so their coefficients may be interpreted as the association of a one-standard-deviation increase in the relevant variable. (*SI Appendix*, Fig. S7 reports these analyses using the self-reported measure of truth telling. In accordance with our preregistration plan, we treat these results as secondary.)

Student quality. Prior work examining unincentivized assessments of truth-telling status (11) or a subclass of egregious mistakes (12, 13) has provided evidence that students with better grades are less likely to misrepresent their preferences. We replicate this finding with our incentivized experimental measure. Participants with higher MCAT scores were significantly more likely to submit an optimal ROL: a one-SD increase in MCAT scores is associated with a five percentage point increase in the rate of truth telling (AME = 0.05, SE = 0.010, $P < 0.001$).

In field settings, an association with test scores can be jointly influenced by both response to a poor strategic position and by differences in logical reasoning ability. These channels are separable in our experiment, and we find evidence that both channels are active. Participants assigned to higher HST scores were more likely to submit an optimal ROL (AME = 0.04, SE = 0.010, $P < 0.001$). Furthermore, participants who performed better on the Raven's Matrices task were more likely to submit an optimal ROL (AME = 0.03, SE = 0.010, $P = 0.002$). As indicated in Fig. 4*A*, these estimates maintain comparable magnitudes and statistical significance while controlling for the full battery of correlates. In summary, the characteristics of high-performing students are useful individual predictors of truth-telling behavior, even when holding other factors constant.

Overconfidence. Examined in isolation, participants exhibiting overconfidence on the Ravens' task were two percentage points more likely to submit the optimal preference ordering, although we cannot reject the null hypothesis of no effect (difference of proportion $z = -0.91$, $P = 0.365$). This difference becomes greater in both magnitude and statistical significance in the complete model, at least partially due to eliminating the offsetting effect of our overconfidence measure's strong negative association with Raven's task performance ($r = -0.59$, $P < 0.001$). All else equal, overconfident participants were more likely to submit optimal ROLs compared with nonoverconfident participants (AME = 0.08, SE = 0.028, $P = 0.005$). Similarly, controlling for MCAT performance, participants who overestimate the percentile of their MCAT score submit an optimal ROL at a significantly higher rate (AME = 0.09, SE = 0.027, $P = 0.001$). Since overconfidence is typically associated with decision errors, the positive correlation documented here may be viewed as surprising. However, this positive relationship could naturally arise from our results on student quality, assuming that overconfidence leads students to overestimate the strength of their strategic position.

Desire to rank the expected outcome highly. We find no support for an effect of our expectations-salience manipulation. No significant difference is found in the propensity to report truthfully as a function of the expectations condition (difference of proportion $z = 0.33$, $P = 0.743$).

Pursuit and availability of advice. Medical students usually seek out and receive advice from many sources about how to maximize their chances for admission to a top residency program. Consistent with this tendency, 71.6% of participants report receiving advice from their medical school, 62.3% reported receiving advice from other students, 40.6% reported receiving advice from the NRMP website, and 23.6% reported receiving advice from other sources. We find that the pursuit and receipt of advice is significantly associated with the likelihood to submit an optimal ROL. Participants showed an increased likelihood to submit an optimal ROL when they reported receiving advice from their medical school (AME = 0.08, SE = 0.024, $P = 0.001$), other students (AME = 0.04, SE = 0.021, $P = 0.043$), the NRMP website (AME = 0.12, SE = 0.020, $P < 0.001$), or other sources (AME = 0.09, SE = 0.022, $P < 0.001$). As shown in Fig. 4*A*, the estimates associated with receiving advice from the NRMP website (AME = 0.10, SE = 0.021, $P < 0.001$) and other sources (AME = 0.07, SE = 0.023, $P = 0.002$) remain largely unchanged in the complete model while those associated with receiving advice from other students (AME = −0.01, SE = 0.023, $P = 0.565$) and from participants' medical school (AME = 0.036, SE = 0.025, $P = 0.158$) attenuate. Similar results are found when regressing truth-telling status on all advice sources simultaneously, excluding all other factors (*SI Appendix*, Table S6). We present extensive exploratory text analysis of the reports of advice received and show the effects of source influence in *SI Appendix*, Figs. S8–S12 and Table S7.

Mistrust of other market participants. While 97.3% of participants trusted the NRMP to run the algorithm honestly, 63.4% of participants did not trust other students to submit a truthful ROL and 42.0% of participants did not trust their residencies to rank order students fairly. We find that participants' likelihood to submit an optimal ROL decreased by five percentage points if they trusted residencies to rank order graduating medical students based on an honest assessment of their quality (AME = −0.05, SE = 0.020, $P = 0.017$), but that neither trust in other

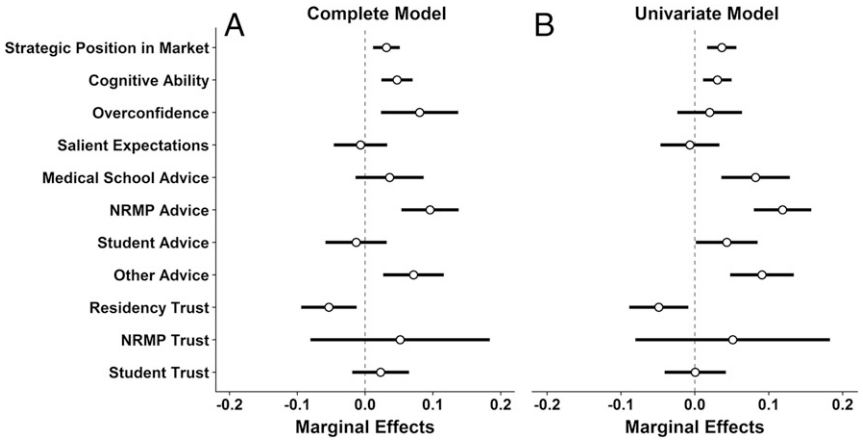

Fig. 4. Predictors of truth telling. Plotted are estimated average marginal effects derived from a logit model predicting whether participants reported truthful preferences. To illustrate the interpretation of effect sizes, note that a marginal effect of 0.1 corresponds to a 10 percentage point increase in the rate of truthful reporting. *A* presents estimates obtained from the complete model, including the entire battery of predictors. *B* presents the estimate for each univariate model, predicting truth telling with only the single variable represented in that row. HST score and Raven's performance are normalized. All other measures are binary. Horizontal lines at each data point represent 95% confidence intervals. See *SI Appendix*, Table S9 for the regression output. Sample for all regressions: 1,714.

students (AME = 0.00, SE = 0.021, $P = 0.982$) nor in the NRMP (AME = 0.05, SE = 0.067, $P = 0.446$) significantly affected performance. These effects remain largely unchanged in the complete model, or when regressing truth-telling status on all trust measures simultaneously, excluding all other factors (*SI Appendix*, Table S8).

Discussion

A large literature in economics has focused on the design of mechanisms that incentivize truth telling, and a large theoretical literature has assumed that behavior in these mechanisms is ultimately truthful. In this paper, we have demonstrated that highly trained and incentivized participants in a flagship application of mechanism design appear to misunderstand these incentives at a substantial rate. Furthermore, this behavior is critically tied to student quality, to overconfidence, to the pursuit and the sources of available advice, and to trust in residency programs.

An immediate implication of our results is that there is room for training programs to help medical students avoid harming themselves through attempts to game the system. As we document, students receiving advice from credible advisors are significantly more likely to behave optimally. At the same time, students reliant on the advice from other students—a potentially noncredible source—are no better, and potentially worse, at finding the optimal strategy. These results converge with evidence from the laboratory suggesting that trust in the "folk wisdom" of other market participants may be misplaced (35). Indeed, as we document in *SI Appendix*, Figs. S8 and S9, free-response descriptions of the advice provided from all sources reveal that a substantial fraction of recommended strategies are misguided. Attempts to better direct students to credible, high-quality advice are clearly needed.

Because different groups face different rates of misrepresentation, and because misrepresentation harms the outcomes of those who pursue it, the use of this mechanism will ultimately favor the groups who best understand it. To the extent that misunderstanding is driven by student ability, this can be desirable. Prior research highlights the potential for misunderstanding of the deferred acceptance algorithm to serve as a screening device and facilitate matching the best students to the best schools (21).[‡] However, our results suggest that factors beyond ability are favored through this channel. Overconfident students, students receiving credible advice, and students distrustful of residency programs are the net beneficiaries in our experiment—an outcome that is likely undesirable compared with the outcome that would arise under universal truth telling. Similar results can arise over more basic demographics: for example, in our data, women are eight percentage points more likely to misrepresent their preferences ($\chi^2 = 16.85$, $P < 0.001$), implying that men are the net beneficiaries of the presence of misrepresentation in this market. For reasons of both fairness and market efficiency, utilization of a mechanism that systematically rewards groups for factors independent of ability is typically viewed as undesirable. Further interventions to mitigate these effects are likely worthwhile, but to the extent that some residual misunderstanding is unavoidable, we encourage further research aimed at formally assessing the comparative performance of different matching mechanisms in the presence of persistent misunderstanding.

ACKNOWLEDGMENTS. We thank Katy Milkman, Maurice Schweitzer, and Ran Shorrer for helpful comments; Melissa Beswick for excellent research assistance; Vincent Conley for assistance in coding our experiment; and the Wharton Behavioral Lab and the Wharton Risk Center for financial assistance.

[‡]While screening on ability is the most natural consideration in the NRMP, screening on other dimensions can become important in other markets. For example, a recent study of the Hungarian college matching system finds that relatively affluent students are more likely to report preferences that suboptimally forego chances for scholarships, ultimately resulting in better targeting of financial aid to those in need (13).

1. Kagel JH, Harstad RM, Levin D (1987) Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica* 55:1275–1304.
2. Berry J, Fischer G, Guiteras RP (2015) Eliciting and utilizing willingness to pay: Evidence from field trials in Northern Ghana (Centre for Econ Policy Res, London), CEPR Discussion paper 10703.
3. Calsamiglia C, Haeringer G, Klijn F (2010) Constrained school choice: An experimental study. *Am Econ Rev* 100:1860–1874.
4. Chen Y, Sönmez T (2006) School choice: An experimental study. *J Econ Theory* 127:202–231.
5. Featherstone CR, Niederle M (2016) Boston versus deferred acceptance in an interim setting: An experimental investigation. *Games Econ Behav* 100:353–375.
6. Klijn F, Pais J, Vorsatz M (2013) Preference intensities and risk aversion in school choice: A laboratory experiment. *Exp Econ* 16:1–22.
7. Pais J, Pintér Á (2008) School choice and information: An experimental study on matching mechanisms. *Games Econ Behav* 64:303–328.
8. Li S (2017) Obviously strategy-proof mechanisms. *Am Econ Rev* 107:3257–3287.
9. Ashlagi I, Gonczarowski YA (2018) Stable matching mechanisms are not obviously strategy-proof. *J Econ Theory* 177:405–425.
10. Pathak PA, Sönmez T (2008) Leveling the playing field: Sincere and sophisticated players in the Boston mechanism. *Am Econ Rev* 98:1636–1652.
11. Rees-Jones A (2018) Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. *Games Econ Behav* 108:317–330.
12. Hassidim A, Romm A, Shorrer RI (2016) "Strategic" behavior in a strategy-proof environment. *Proceedings of the 2016 ACM Conference on Economics and Computation* (ACM, New York), pp 763–764.
13. Shorrer RI, Sóvágó S (2017) Obvious mistakes in a strategically simple college admissions environment: Causes and consequences. Available at https://ssrn.com/abstract=2993538. Accessed September 1, 2018.
14. Artemov G, Che Y-K, He Y (2017) Strategic 'Mistakes': Implications for Market Design Research. Mimeo.
15. Gale D, Shapley LS (1962) College admissions and the stability of marriage. *Am Math Mon* 69:9–15.
16. Roth AE, Peranson E (1999) The redesign of the matching market for American Physicians: Some engineering aspects of economic design. *Am Econ Rev* 89:748–780.
17. Pathak PA (2011) The mechanism design approach to student assignment. *Annu Rev Econ* 3:513–536.
18. National Residency Match Program (2017) Results and data: 2017 main residency match. (National Resident Matching Program, Washington, DC).
19. Benjamin DJ, Heffetz O, Kimball MS, Rees-Jones A (2014) Can marginal rates of substitution be inferred from happiness data? Evidence from residency choices. *Am Econ Rev* 104:3498–3528.
20. Hassidim A, Marciano D, Romm A, Shorrer RI (2017) The mechanism is truthful, why aren't you? *Am Econ Rev* 107:220–224.
21. Rees-Jones A (2017) Mistaken play in the deferred acceptance algorithm: Implications for positive assortative matching. *Am Econ Rev* 107:225–229.
22. Basteck C, Mantovani M (2018) Cognitive ability and games of school choice. *Games Econ Behav* 109:156–183.
23. Raven JC (1998) *Raven's Progressive Matrices* (Oxford Psychologists Press, Oxford).
24. Carpenter PA, Just MA, Shell P (1990) What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychol Rev* 97:404–431.
25. Saposnik G, Redelmeier D, Ruff CC, Tobler PN (2016) Cognitive biases associated with medical decisions: A systematic review. *BMC Med Inform Decis Mak* 16:138.
26. Bazerman MH, Moore DA (2013) *Judgment in Managerial Decision Making* (Wiley, New York), 8th Ed.
27. Pan S (2016) The instability of matching with overconfident agents: Laboratory and field investigations (Univ of Melbourne, VIC, Australia). Available at https://www.uts.edu.au/sites/default/files/2018-06/Pan2017.pdf. Accessed November 20, 2017.
28. Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psychol Rev* 115:502–517.
29. Guillén P, Hing A (2014) Lying through their teeth: Third party advice and truth telling in a strategy proof mechanism. *Eur Econ Rev* 70:178–185.
30. Guillén P, Hakimov R (2018) The effectiveness of top-down advice in strategy-proof mechanisms: A field experiment. *Eur Econ Rev* 101:505–511.
31. Dubins LE, Freedman DA (1981) Machiavelli and the Gale-Shapley Algorithm. *Am Math Mon* 88:485–494.
32. Roth AE (1982) The economics of matching: Stability and incentives. *Math Oper Res* 7:617–628.
33. Guillén P, Hakimov R (2017) Not quite the best response: Truth-telling, strategy-proof matching, and the manipulation of others. *Exp Econ* 20:670–686.
34. National Resident Matching Program (2016) Impact of length of rank order list on match results: 2002–2016 main residency match (National Resident Matching Program, Washington, DC).
35. Ding T, Schotter A (2015) Intergenerational advice and matching: An experimental study. Available at https://www.semanticscholar.org/paper/Intergenerational-Advice-and-Matching-%3A-An-Study-%E2%88%97-Ding-Schotter/f4ab23532b27dbecc57ef96b2a8b6a7dcbf8cfb4. Accessed March 10, 2015.