Suboptimal Behavior in Strategy-Proof Mechanisms: Evidence from the Residency Match

Alex Rees-Jones[*]

The Wharton School, University of Pennsylvania

September 18, 2015

Abstract: Strategy-proof mechanisms eliminate the possibility for gain from strategic misrepresentation of preferences. If market participants respond optimally, these mechanisms permit the observation of true preferences and avoid the implicit punishment of market participants who do not try to "game the system." Using new data from a flagship application of the matching literature—the medical residency match—I study if these potential benefits are fully realized. I present evidence that some students pursue futile attempts at strategic misrepresentation, and examine the causes and correlates of this behavior. These results inform the assessment of the costs and benefits of strategy-proof mechanisms, and demonstrate broad challenges in mechanism design.

Keywords: matching, deferred acceptance algorithm, suboptimal behavior.

JEL Classification Numbers: C78, D03.

[*] E-mail: alre@wharton.upenn.edu. I thank Aaron Bodoh-Creed, Dan Benjamin, Ori Heffetz, Judd Kessler, Miles Kimball, Ulrike Malmendier, Sean Nicholson, Ted O'Donoghue, Ran Shorrer, and seminar participants at UC Berkeley, the CHIBE/Roybal Retreat, Cornell University, George Mason University, The Stanford Institute for Theoretical Economics, and the University of Pennsylvania for helpful comments and advice. I thank Allison Ettinger and Matt Hoffman for their help in recruiting schools to the survey, Andrew Sung for excellent research assistance, and the National Institute on Aging (grant T32-AG00186) for generous research support.

A substantial literature in economics has explored mechanism design in two-sided matching markets. The defining characteristic of these markets is the need to accommodate the preferences of the two groups being matched: for example, when matching students to schools. Compared to the one-sided markets more commonly studied, these settings pose unique challenges to reaching desirable outcomes. Difficulty in coordinating on the timing of decisions often leads to "market unraveling" (Roth and Xing, 1994). Furthermore, decentralized approaches often result in unstable matches,[1] which have been empirically shown to be detrimental to the success of these markets (Roth, 1990; Roth, 1991). These problems can be avoided by employing a stable matching mechanism to assign a binding match based on preferences reported to a neutral intermediary at an agreed-upon time. However, the use of these mechanisms introduces the new challenge of managing the strategic incentives involved with preference reporting. If market participants can benefit from misrepresenting their preferences, we expect them to do so.

The student-proposing deferred acceptance algorithm (DAA) of Gale and Shapley (1962) provides a partial solution to the issue of strategic misreporting. For students, this mechanism is *strategy-proof*: truthful preference reporting is a weakly dominant strategy (Dubins and Freedman, 1981; Roth, 1982). Furthermore, truthtelling is approximately optimal for all market participants in sufficiently large markets (Immorlica and Mahdain, 2005; Kojima and Pathak, 2009; Avezedo and Budish, 2013). Strategy-proof mechanisms therefore provide a comparatively simple optimal strategy, which has been viewed as especially useful in the student-to-school matching setting. If optimal play is pursued, students may entirely avoid

[1] That is, matches in which a pair of agents both prefer to be assigned to each other instead of their realized pairing, or where a matched individual prefers to be unmatched.

devoting time or effort into figuring out how they should misrepresent their preferences. Students with a poor grasp of game theory are not punished for their failure to optimally "game the system," resulting in a level playing field between strategically sophisticated and strategically unsophisticated market participants (Pathak and Sonmez, 2008). These features, along with other desirable theoretical properties of the student-proposing DAA, have led a number of prominent market designers to assist in deploying this mechanism to the field (Roth and Peranson, 1999; Abdulkadiroğlu, Pathak, and Roth, 2005; Abdulkadiroğlu, Pathak, Roth, and Sonmez, 2005).

This paper explores empirically whether the benefits of strategy-proof mechanisms are fully realized. The typically expressed logic suggests that incentivizing truthful reporting will lead to truthful reports. However, even though the optimal strategy in the student-proposing DAA is simple, the strategic environment remains quite complex. In order to deduce the optimal strategy in this environment, students must draw upon a significant degree of game-theoretic sophistication. If any portion of the population lacks the necessary sophistication (or trusted advice from a sophisticated adviser) failures of optimal behavior might arise. Just as an otherwise-able student might misunderstand the strategic incentives faced in a non-strategy-proof mechanism, and fail to optimally "game the system," so too might a student do so in a strategy-proof mechanism. In this environment, the result would be misrepresentation of preferences despite the lack of scope for successful manipulation.

In this paper, I document the existence and nature of this suboptimal behavior in a classic setting from the matching literature: the process matching medical students to medical residencies. Analyzing a survey I administered to graduating medical students at 23 medical schools, I find that 17% of students self-assess their preference reporting strategy to be nontruthful, with 5% directly attributing this nontruthful behavior to strategic considerations. To validate these self-reports, I demonstrate that proxies for welfare are less predictive of the submitted preferences of students reporting nontruthful behavior, consistent with a disruption of utility maximization. Pursuit of strategic misrepresentation is more prevalent among men, among those with lower academic performance, and among those in more competitive specialties.

A growing literature in experimental economics has examined individual behavior in DAA-related mechanisms, and commonly finds a fraction of respondents with nontruthful reporting behavior (see, e.g., Chen and Somnez, 2006; Pais and Pintér, 2008; Calsamiglia, Haeringer, and Klijn, 2010; Klijn, Pais, and Vorsatz, 2013; Featherstone and Niederle, 2014). However, extending the study of this behavior outside of a controlled laboratory environment is challenging. While true preferences may be controlled or assigned—and thus observed—in the lab, the inability to observe true preferences is a defining characteristic of the field settings in which these matching mechanisms are deployed.[2] The validated self-classification approach presented in this paper offers a unique demonstration that failures of truthful reporting persist outside of the lab, in perhaps the most well studied and carefully designed two-sided matching mechanism currently in existence.

Beyond their implications specific to two-sided matching, these results permit a broader assessment of the limits of incentive compatibility. Economists commonly assume that optimal play can be expected when market participants are sufficiently intelligent, when sufficient information on the game is available, and when stakes are sufficiently high. The population

[2] Indeed, if true preferences were observed, designing a matching mechanism to incentivize truthful reporting would be unnecessary.

considered in this paper is far more educated than most, is acting in a setting with advice readily available and long institutional history with this mechanism, and is extremely invested in the outcome that this algorithm determines. On one hand, the low rate of nontruthful reporting found may be interpreted as a success: most participants appear to respond to incentives as they should. However, the persistence of suboptimal behavior in this setting, even at low rates, suggests the requisite levels of intelligence, information, and incentivization needed to ensure full compliance may never be achieved in practice. Some strategic misunderstanding may be unavoidable in these settings, necessitating attention to the comparative performance of mechanisms in the presence of suboptimal behavior, and the design of mechanisms that can minimize misunderstanding.[3]

This paper proceeds as follows. In section 1, I provide institutional details about the residency match, and discuss the survey data collected for this paper. Sections 2.1 and 2.2 present main results, and 2.3 addresses several robustness concerns. Section 3 concludes by discussing the implications of these results for mechanism choice in policy applications.

1 Institutional Setting and Data Collection

The data considered in this paper come from a survey of medical students participating in the 2012 National Resident Matching Program (NRMP). In this section, I provide a brief overview of the NRMP and the matching process, then present the details of data collection.

[3] For a fruitful approach to classifying mechanisms by their cognitive difficulty, leading to the design of mechanisms that are "obviously strategy-proof," see Li (2015).

1.1 Background on the Matching Process

The NRMP serves as a central clearinghouse for matching graduating medical students to U.S. residency programs. Its primary function is to collect the reported preferences of both students and residencies, and to use this information to determine the final matching. This has historically been done with mechanisms related to the DAA. In the 1951-1952 academic year, the NRMP implemented a matching algorithm equivalent to the school-proposing DAA, predating Gale and Shapley's seminal study of this mechanism by a decade (Roth, 2008). In the time since, this market has been the frequent subject of matching research (e.g., Roth, 1984; Roth, 1996; Agarwal, 2015). The NRMP's interaction with market designers ultimately lead them to invite Alvin Roth to assist in a redesign of the matching algorithm. This algorithm, implemented in 1998, is based on the student-proposing DAA, with several modifications to accommodate idiosyncrasies of the medical market (for full details, see Roth and Peranson, 1999). While these modifications complicate the strategic environment and render it not formally strategy-proof, simulations in Roth and Peranson (1999) demonstrate that the mechanism preserves incentives for truthful preference reporting for essentially all students.[4]

Medical students typically participate in this matching process in their fourth and final year of medical school. In preparation for participating in the match, students directly apply to a number of residency programs. Interested programs invite the student to visit and interview with program representatives. Both the students and the residencies use these interviews to gather information about their potential match partners. Following this

[4] Across 5 years of match data, their simulations suggest that the number of students who could benefit from misrepresentation ranged from 0 to 9 per year, out of approximately 20,000-25,000 applicants in the studied years.

interview process, students and residency representatives both determine their preferences over possible matches. These preferences are submitted to the NRMP at a coordinated time, and a binding match is announced several weeks later.

1.2 Implementation of Data Collection

To better understand the behavior of students in this submission process, I administered a large-scale survey of medical students during the 2012 residency match. This survey was conducted in collaboration with Daniel Benjamin, Miles Kimball, and Ori Heffetz, and has also been used to assess the performance of subjective well-being data as a utility proxy (Benjamin, Heffetz, Kimball, and Rees-Jones, 2014). In the lead-up to the 2012 residency match, we contacted virtually all 122 U.S. medical schools with full accreditation from the Liaison Committee on Medical Education. As a result of our outreach, 23 medical schools agreed to participate. At participating schools, an email was forwarded to students immediately after the NRMP preference submission deadline (February 22nd in the survey year). This email explained that the school was participating in a study of decision making in the residency match, and contained a link to the survey website. 579 students voluntarily completed this survey. Furthermore, students who completed this survey were asked to participate in a follow-up around one to two weeks later. The follow-up survey repeated all questions from the initial survey, facilitating an assessment of response error. This survey was completed by 133 respondents.

The timing of both surveys fell between the submission of preferences and the announcement of the match results. This timing was essential. First, it ensured that the decision was fresh in the respondents' minds; the median survey response was completed 11 days after preferences were submitted. Second, this timing ensured that students' information set was essentially identical to that which they had at their moment of choice. It is possible that the additional information conveyed by learning the outcome of the match would lead students to reconsider their preferences (either for rational or psychological reasons); the timing of this survey avoids this confounding factor.

The primary survey data of relevance to this study is a battery of questions about the truthfulness of the student's reporting behavior. In addition, the survey elicited students' top 4 choices from their rank order list, along with predictions about a number of attributes associated with these residencies.[5] Analysis in this paper is restricted to the 561 respondents who reported a preference ordering including at least two residencies. The details of survey items used will be presented as they are analyzed in the following section. Complete information on the survey's implementation—including recruitment materials and procedures, screenshots of the survey instrument, and analysis of selection into survey participation—is available in Benjamin, Heffetz, Kimball, and Rees-Jones (2013).

2 Main Results

This section presents survey evidence on the existence and nature of nontruthful preference reporting in the residency match. Section 2.1 presents the primary assessment of the prevalence of nontruthful reporting and the characteristics of those who pursue it. Section

[5] Note that while this only reveals a portion of a students' preference ordering, it is likely to be portion that is relevant for final assignments. In 2012, 83.6 percent of NRMP participants graduating from U.S. medical schools were matched to one of their top four choices (NRMP, 2012).

2.2 assesses the relationship between submitted preference orderings and available welfare proxies. Section 2.3 considers several robustness concerns relevant for interpreting these results.

2.1 Self-Assessments of Preference Reporting Behavior

The primary question of relevance to truthful reporting was the following: "When forming the ranking of residencies to submit to the NRMP, some candidates submit an ordering that is not the true order of how desirable they find the programs. When forming your list, did you report the exact ordering of your true preferences?" The available multiple-choice responses were "Yes," "No – I chose my list strategically," "No – I tried to report my true preferences, but I made a mistake," or "No – Other reason" with a request to list the other reason. All respondents are subsequently given a free-response opportunity to explain the motivations and reasoning behind their divergence between true and submitted preferences.

Table 1 presents a tabulation of the response to this question. The first row provides the distribution of responses for the full sample. The vast majority (83%) of survey respondents feel that their submitted preferences do accurately reflect their true preferences. The remaining 17% indicated that they pursued nontruthful reporting practices in one of the three categories provided. 5% of respondents report that their true preferences and submitted preferences differ due to an attempt at "strategic behavior;" since successful strategic manipulation is impossible for essentially all applicants, this may be viewed as evidence that a misunderstanding of strategic incentives influences at least a small portion of responses. Less than 1% of respondents—only two individuals—report that they felt they made a mistake, suggesting that conscious errors are not a primary determinant of the nontruthful behavior observed.[6] The remaining 11% of respondents reporting nontruthful behavior indicated that this was due to some "other reason." The reasons provided by respondents often described some combination of locational constraints and constraints imposed by family or a significant other. While it is possible that these subjects harbor a misunderstanding of the mechanism, these free responses suggest an alternative explanation for their reported deviation between reported and true preferences. Some of these survey respondents may have understood the term "preferences" in a particularly narrow sense, drawing a distinction between their preferences formed without regard for non-academic concerns and preferences that take into account all competing outside factors.[7] Given this concern with interpretation, I will generally focus attention on respondents directly reporting strategic manipulation, as this group more clearly demonstrates a misunderstanding of the mechanism.

In the remaining rows of table 1, I tabulate assessments of preference reporting behavior by all available subject characteristics. The first group of characteristics capture basic demographic information: gender, relationship status, participation in the couples' match, and age. Among these categories, I find evidence of differences in the distribution of responses by gender (Fisher's exact test p-value = 0.037), relationship status (Fisher's exact test p-value = 0.041), and dual-match participation (Fisher's exact test p-value = 0.065). The differences seen among these distributions reveal a notable difference in women's propensity to claim strategic nontruthful reporting (4% for women versus 7% for men), and a clear tendency for people in relationships and participants in the couples match to claim nontruthful behavior

[6] Given the amount of time and effort typically devoted to the residency preference decision, and the large incentives surrounding this decision, this low rate of conscious mistakes is perhaps to be expected.

[7] The latter definition aligns best with economists' use of the term.

for "other reasons" (consistent with the explanations seen in the free responses discussed above). The observed differences in the response distributions by age are not statistically significant at traditional significance levels (Fisher's exact test p-value = 0.196).

Next are four measures of the academic abilities of the respondent: college GPA, as well as test scores for the MCAT and Medical Licensing Exams (step 1 and step 2). Directionally, all four of these measures show better preforming students to be more likely to tell the truth, and less likely to specifically report strategic nontruthful behavior. However, Fisher's exact tests do not reject the null hypothesis of independence relative to these academic performance measures.

The final row offers a measure of the competitiveness of the respondents' specialty: the number of U.S. applicants applying for positions in that specialty, divided by the number of positions available.[8] Directionally, we see that applicants in specialties with more competition for positions are more likely to report nontruthful behavior in general, and strategic nontruthful behavior in specific. However, as with the academic measures above, a Fisher's exact test does not reject the null hypothesis of independence relative to this subject characteristic.

The analysis of table 2 further explores the association of these different individual characteristics on propensity towards strategic misrepresentation. This table presents the results of a multinomial logit regression predicting self-assessment of reporting behavior based on the subject characteristics just considered.[9] Included predictor variables are dummy variables for gender, relationship status, and dual-match participation, as well as continuous

[8] Information regarding the number of applicants and positions is drawn from the NRMP 2012 match summary (NRMP, 2012).

[9] Due to the small sample size reporting nontruthful behavior due to a mistake, this group is excluded from this analysis.

measures of age and the competition ratio discussed in the previous paragraph. To ease assessment and interpretation, I condense the four available academic ability measures into a single academic ability index. This academic ability index is calculated using principle component analysis on college GPA and the three available test-scores.[10]

Two notable results arise from this analysis. First, the estimates suggest a negative association between academic ability and propensity to strategically misreport preferences. Quantitatively, a 1 standard deviation increase in the academic ability index is associated with a 2 percentage point reduction in the rate of strategic misrepresentation, all else equal. This suggests that better students tend to be more strategically sophisticated. This finding has important implications for assessing the impact of strategic mistakes on the final match; we will further discuss this issue in section 3.

Next, notice that propensity towards strategic misreporting shows some evidence of association with both gender and with the competition for positions in the subject's chosen specialty. The estimated average marginal effects suggest that, ceteris paribus, women are 3 percentage points less likely to strategically misrepresent their preferences ($p = 0.098$), and that an increase in the competition measure of 1 unit is associated with a 1 percentage point increase in probability of strategic misrepresentation ($p = 0.053$). Since this study contains no exogenous variation in either gender or competitiveness, strong causal claims are not possible; however, to the extent that futile pursuit of a strategic advantage is indeed a "competitive" behavior, these results suggest that this setting reflects similar patterns to those in recent studies of gender differences in competition (for a review, see Niederle and

[10] Factor loadings are available in appendix table A1, and demonstrate that all 4 measures are positively associated with this index, and non-response on each item is negatively associated with the index.

Vesterlund, 2011).

2.2 Evidence of Disruption of Utility Maximization

While the results of the previous section demonstrate that a minority of students directly assess their own behavior as nontruthful, these results are vulnerable to a common criticism of survey data: that self-reports might not accurately reflect actual behavior. In this section, I assess this concern by examining the relationship between reported truthtelling status and several proxies for welfare. Under the typical assumption of welfare-maximizing behavior, a respondent's true rank-order of residencies should align with that respondent's rank-order of welfare forecasts. If individuals strategically misrepresent their preferences, this alignment is disrupted. This yields the testable prediction here assessed: if these self-reports are valid, we should expect the proxies for welfare to be more weakly associated with the preferences reported by those individuals reporting nontruthful behavior.

To test this prediction, I turn to more detailed data on respondents' assessments of the residencies in their preference ordering. For each of their top-4 residencies, respondents faced a battery of 12 questions eliciting evaluations of residency attributes.[11] The full text of these questions is available in appendix table A2, and summarized here. Nine of these attributes were included to capture important determinants of residency choices. These elicit, on a scale from 1 to 100, perceptions of the prestige and status associated with the residency; the quality of social life expected during the residency; the desirability of the residency's location; the expected amount of anxiety experienced on a typical day; the extent to which

[11] The four residencies were considered in random order. Additionally, the order of the 12 attributes was randomized for each residency.

life would seem worthwhile; the expected amount of stress on a typical day; expectations of future career prospects; the degree of control over one's life afforded by the residency; and, for respondents in a relationship, the desirability of that matching for the spouse or significant other. Three additional attributes were crafted after subjective well-being (SWB) questions common to large-scale social surveys. These elicit the respondents' predictions of their overall life assessment should they attend this residency, their predicted life satisfaction during the residency, and their predicted happiness on a typical day.

These data are used to create two groups of proxies for welfare. The first group consists of the three SWB questions.[12] The second group consists of the predicted utility values coming from a revealed-preference approach, rationalizing the observed preference orderings with a latent utility function over residency attributes.

In order to assess the association between a given welfare measure and reported preference orderings, I implement the rank-order logit model of Beggs, Cardell, and Hausman (1981). In this model, I assume that each individual's ordinal ranking of residencies is rationalized by a latent, random index: $I_{ir} = \beta X_{ir} + \epsilon_{ir}$ (where subscript i denotes individuals and subscript r denotes the residency considered from the top 4). The coefficient vector β is estimated by maximizing the sum of individual log-likelihoods that $I_{i1} > I_{i2} > I_{i3} > I_{i4}$—i.e., maximizing the likelihood that the estimated model rationalizes the observed choices. The error term is assumed to follow a type-I extreme-value distribution, permitting the evaluation of likelihood

[12] While notions of "happiness" or "satisfaction" do not perfectly map to economists' notions of utility or welfare, these measures have been used to approximate economic utility in a variety of settings. Example applications include pricing noise (van Praag and Baarsma, 2005), informal care (van den Berg and Ferrer-i-Carbonell, 2007), the risk of floods (Luechinger and Raschky, 2009), and air quality (Levinson, 2012), as well as quantifying the impact of relative income comparisons (Luttmer, 2005) and the Moving to Opportunity project (Ludwig et al., 2012). Recent work has shown substantial positive associations between preferences inferred from choice data and happiness data, while simultaneously demonstrating systematic differences between these objects (Benjamin, Heffetz, Kimball, and Rees-Jones, 2012, 2014; Perez-Truglia, 2015).

in closed-form.

Panel A of table 3 estimates rank-order logit models where the ordering is predicted by one of the three SWB measures. Separate coefficients are estimated for those indicating truthful reporting, nontruthful reporting for strategic reasons, and nontruthful reporting for other reasons.[13] Since the magnitude of marginal utilities are measured relative to the error term this framework, the implied predictive power of a given attribute is increasing in the absolute value of its associated coefficient. For example, when comparing two residencies with a 1 standard deviation difference in life satisfaction, a larger β implies a higher probability that the more satisfying residency is chosen.[14] To facilitate quantitative comparisons, appendix table A4 formally calculates these differences in probability. To facilitate assessment of the statistical significance of observed differences, the bottom two rows of each panel in table 3 provide p-values for two-sided Wald tests that $\beta_{truthful} = \beta_{strategic}$ and $\beta_{truthful} = \beta_{other}$.

Across these three measures, the estimated coefficients for truthful and nontruthful reporters show clear and systematic differences. Analysis of all three suggests that these welfare proxies are more predictive of choice for those who indicated truthful preference reporting. The direction of all comparisons is as predicted, with strong statistical significance seen in 4 of the 6 comparisons.

In panel B of table 3, I construct utility estimates from revealed-preference analysis of residency attributes, then assess the differential predictive power of these utility proxies by truthtelling status. This exercise proceeds in two steps. In the first step, I estimate rank-order logit models predicting choice as a function of residency attributes. First-stage

[13] Individuals indicating nontruthful reporting due to making a mistake are excluded, due to the extremely small sample size of this group.

[14] This claim relies on the assumption that satisfaction is a desirable property, and thus the associated coefficients are positive.

regression coefficients are reported in appendix table A3. The estimated models are then used to calculate predicted values of the latent linear-utility index rationalizing reported choices, $\bar{U} = \hat{\beta}X$. In the second step, reported in panel B, I examine the strength of the association between these revealed-preference welfare metrics and truth-telling status in the same manner as in panel A.

Columns 1 through 4 of panel B differ in the attributes and sample used to calculate the revealed-preference welfare metric $\bar{U}$. In the first column of panel B, the first-stage specification predicts choice using the 9 non-SWB attributes, as in the primary specification of Benjamin, Heffetz, Kimball, and Rees-Jones (2014). In the third column of panel B, the first-stage specification includes the 9 non-SWB attributes as well as the three SWB measures. In columns 2 and 4, I conduct the same exercises as in columns 1 and 3, but restrict the first-stage estimation sample to those indicating truthful preference reporting.[15] Examining the differences in coefficients across reporting-status in these four specifications, we see that nontruthful reporters have significantly weaker associations between this utility metric and reported preferences.

In summary, among those indicating nontruthful reporting, we see a systematically weaker link between reported preferences and welfare-relevant metrics—whether taken from revealed-preference approaches, or from direct statements of subjects' predicted well-being.

[15] Conditional on finding evidence in support of the existence of nontruthful reporters, it follows that choices need not reveal preferences for this group. This motivates estimating the revealed-preference weights on residency attributes solely from truthful reporters. Notice that since $\bar{U}$ was estimated based from the truthful sample, the coefficient on the second-stage rank-order logit regression predicting choice with $\bar{U}$ is mechanically 1 for truthful reporters (as seen in both column 2 and 4).

2.3 Robustness Concerns

In this section, I present and consider two important robustness concerns relevant for assessing these results.

Non-representative sample: This survey is conducted among a possibly non-representative sample of medical students. Consequently, these estimates are potentially subject to sample selection bias. While such a bias could not explain the presence of suboptimal behavior if none existed in population, it could affect estimates of the prevalence of this behavior. In the course of preparing this dataset, significant attention was devoted to assessing selection into the survey population (for supporting analysis and tests, see Benjamin, Heffetz, Kimball, and Rees-Jones, 2013). Selection could occur at two stages: first, the medical schools which agreed to participate in this study might not be representative of the full population of medical schools; and second, the students within each school which complete the survey might not represent the schools' student population. I find no evidence of the first category of selection, and limited evidence of the second. Comparing medical schools which agreed to participate in this study with those that did not, no statistically distinguishable differences are detected across total enrollment, MCAT scores, undergraduate GPA scores, acceptance rates, U.S. News Research Rankings, or gender composition. Comparing the demographics and test scores of survey participants to the average characteristics of their school, the only statistically significant difference was slightly higher reported college GPAs (0.04 points higher in the survey sample, $p < 0.001$). Of course, while evidence of selection on observables is limited, selection on unobservables remains possible, and indeed is likely. For example, particularly prosocial students may be more likely to voluntarily respond to a web-survey;

this could lead to an overestimate of the rate of truthful reporting. This concern is reasonable, and inference on the population rate of truthful behavior should be performed with this caveat in mind.

Measurement error in self-reports: The validation exercise presented in section 2.2 demonstrates that the self-reports of reporting behavior analyzed in this paper do meaningfully predict the propensity of reported preferences to be welfare maximizing. While this establishes that these survey measures do have some association with the true behavior we aim to study, it does not rule out the possibility of measurement error. As with any survey elicitation, a confound arises if subjects are not reporting their perceptions entirely accurately or truthfully to the surveyor. Given medical students are repeatedly advised and instructed to report their preferences truthfully in the match process, the most natural concern would be a hesitance to admit nontruthful behavior. The survey was designed to emphasize confidentiality in an effort to alleviate this concern. However, to the extent that this concern persisted for survey respondents, some degree of underestimation of the true rate of non-truthful reporting is expected.

To help assess the rate of measurement error in survey responses, a follow-up survey was administered which asked the same questions, unexpectedly and separated in time. Test-retest correlation was high across key elements of the survey (e.g., 0.87 for a dummy variable indicating truthful preference reporting; $p < 0.001$), offering evidence in support of the reliability of these measures.[16]

[16] Analysis is based on the 129 who respondents answered the multiple-choice question from table 1 in both waves. Of the 22 who indicated nontruthful behavior when first surveyed, only 2 changed their assessment to truthful when recontacted. 3 students who had previously assessed their behavior as truthful reassessed it as nontruthful.

3 Discussion

In this paper I have documented the perceptions of medical students about their own truthful reporting, and validated these measures with two complementary approaches to welfare analysis. Among students surveyed in the residency match, most do indeed perceive their reported preference ordering to be truthful. However, a subpopulation of students appear to be misrepresenting their true preference ordering in an attempt at strategic behavior, in a manner which theoretical considerations suggest is suboptimal.

These results are relevant when assessing the costs and benefits of different matching mechanisms. While strategy-proofness is a desirable property, it does not come for free; for example, Abdulkadiroğlu, Che, and Yasuda (2011) demonstrate that the non-strategy-proof Boston mechanism can yield outcomes which Pareto dominate those of the DAA. This suggests the choice to implement the DAA does involve some welfare cost relative to existing alternatives; this cost has been argued to be justified due to the benefits this mechanism affords to the strategically unsophisticated, among other things. The results of this paper demonstrate that the punishment of the strategically unsophisticated is not eliminated in the DAA as is commonly assumed, implying a reweighing of its benefits relative to its costs. For an experimental investigation studying this comparison in depth, see Featherstone and Niederle (2014).

Considerations such as these led Daniel McFadden (2009) to suggest that "tolerance of behavioral faults be added to the criteria for good mechanism design." While mechanism designers have historically been concerned about students' ability to deduce and implement complex optimal strategies, the results of this paper suggest attention to strategic sophistication is needed in environments with the simplest optimal strategy: truthful reporting. While greater understanding of the theoretical consequences of this behavior are necessary, some immediate results exist. As demonstrated in appendix section B, bounds may be derived on the extent to which nontruthful reporters are harmed by their suboptimal behavior. Furthermore, in environments where strategic sophistication is correlated with ability, and where schools rank students according to an imperfect measurement of their ability, the presence of this suboptimal behavior has the potential to facilitate positive assortative matching, thus providing an interesting channel through which lack of strategic sophstication may improve market efficiency. As we continue to deploy two-sided matching mechanisms to the field, further study and attention to these issues will likely prove fruitful.

4 References

Abdulkadiroğlu, Atila, Yeon-Koo Che, and Yosuke Yasuda. 2011. "Resolving Conflicting Preferences in School Choice: The "Boston" Mechanism Reconsidered." *American Economic Review*, 101(1), 399 - 410.

Abdulkadiroğlu, Atila, Parag Pathak, and Alvin Roth. 2005. "The New York City High School Match." *American Economic Review, Papers and Proceedings*, 95: 364-367.

Abdulkadiroğlu, Atila, Parag Pathak, Alvin Roth, and Tayfun Sonmez. 2005. "The Boston Public School Match." *American Economic Review, Papers and Proceedings*, 95: 368-371.

Agarwal, Nikhil. 2015. "An Empirical Model of the Medical Match." *American Economic Review,* 105(7): 1939-1978.

Azevedo, Eduardo and Eric Budish. 2013. "Strategy-proofness in the Large." Working paper.

Beggs, S., S. Cardell, and J. Hausman. 1981. "Assessing the Potential Demand for Electric Cars." *Journal of Econometrics*, 17 (1): 119.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones. 2012. "What Do You Think Would Make You Happier? What Do You Think You Would Choose?" *American Economic Review*, 102(5): 2083 - 2110.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones. 2013. "Survey Appendix to: Can Marginal Rates of Substitution Be Inferred From Happiness Data? Evidence from Residency Choices."

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones. 2014. "Can Marginal Rates of Substitution Be Inferred From Happiness Data? Evidence from Residency Choices." *American Economic Review*, 104(11): 3498 - 3528.

Calsamiglia, Caterina, Guillaume Haeringer, and Flip Klijn. 2010. "Constrained school choice: an experimental study." *American Economic Review*, 100(4): 1860 - 1874.

Chen, Yan and Tayfun Sonmez. 2006. "School choice: an experimental study." *Journal of Economic Theory*, 127(1): 202 - 231.

Dubins, Lester and David Freedman. 1981. "Machiavelli and the Gale-Shapley Algorithm." *American Mathematical Monthly*, 88(7): 485-494.

Featherstone, Clayton and Muriel Niederle. 2014. "Improving on Strategy-Proof School Choice Mechanisms: An Experimental Investigation." Working paper.

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

Immorlica, Nicole and Mohammad Mahdian. 2005. "Marriage, Honesty, and Stability." *Proceedings of the sixteenth annual ACM-SIAM symposium on discrete algorithms*, 53-62.

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

Kojima, Fuhito and Parag Pathak. 2009. "Incentives and Stability in Large Two-Sided Matching Markets." *American Economic Review*, 99(3): 608 - 627.

Levinson, Arik. 2012. "Valuing Public Goods Using Happiness Data: The Case of Air Quality." *Journal of Public Economics*, 96 (9-10): 869-80.

Li, Shengwu. 2015. "Obviously Strategy-Proof Mechanisms." *SSRN Working Paper No. 2560028.*

Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, and Lisa Sonbonmatsu. 2012. "Neighborhood Effects on the Long-Term Well-Being of Low- Income Adults." *Science*, 337(6101): 1505-10.

Luechinger, Simon, and Paul A. Raschky. 2009. "Valuing Flood Disasters Using the Life Satisfaction Approach." *Journal of Public Economics*, 93 (34): 62033.

Luttmer, Erzo F. P.. 2005. "Neighbors as Negatives: Relative Earnings and Well-Being." *Quarterly Journal of Economics*, 120 (3): 963-1002.

McFadden, Daniel. 2009. "The Human Side of Mechanism Design: A Tribute to Leo Hurwicz and Jean-Jacque Laffont." *Review of Economic Design*, 13:77-100.

National Resident Matching Program. 2012. "National Resident Matching Program, Results and Data: 2012 Main Residency Match." National Resident Matching Program, Washington, DC.

Niederle, Muriel and Lise Vesterlund. 2011. “Gender and Competition.” *Annual Review of Economics,* 3: 601-603.

Pais, Joana and Ágnes Pintér. 2008. “School choice and information: An Experimental Study on Matching Mechanisms.” *Games and Economic Behavior,* 64(1): 303 - 328.

Pathak, Parag and Tayfun Sonmez. 2008. “Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism.” *American Economic Review*, 98: 1636 - 1652.

Perez-Truglia, Ricardo. 2015. “A Samuelsonian Validation Test for Happiness Data.” *Journal of Economic Psychology*, 49: 74 - 83.

Roth, Alvin. 1982. “The Economics of Matching: Stability and Incentives.” *Mathematics of Operations Research*, 7(4): 617 - 628.

Roth, Alvin. 1984. “The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory.” *Journal of Political Economy*, 92(6): 991-1016.

Roth, Alvin. 1990. “New Physicians: A Natural Experiment in Market Organization.” *Science*, 250: 1524-1528.

Roth, Alvin. 1991. “A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians and Surgeons in the U.K.” *American Economic Review*, 81: 415-440.

Roth, Alvin. 1996. “The NRMP as a Labor Market.” *Journal of the American Medical Association*, 275: 1054-1056.

Roth, Alvin. 2008. “Deferred Acceptance Algorithms: History, Theory, Practice, and Open Questions.” *International Journal of Game Theory, Special Issue in Honor of David*

Gale on his 85th birthday, 36: 537 - 569.

Roth, Alvin, and Elliot Perason. 1999. "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design." *American Economic Review*, 89(4): 748 - 780.

Roth, Alvin, and Xiaolin Xing. 1994. "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions." *American Economic Review*, 84, 992 - 1044.

van den Berg, Bernard, and Ada Ferrer-i-Carbonell. 2007. "Monetary Valuation of Informal Care: The Well-Being Valuation Method." *Health Economics*, 16(11): 1227 - 44.

van Praag, Bernard M. S., and Barbara E. Baarsma. 2005. "Using Happiness Surveys to Value Intangibles: The Case of Airport Noise." *Economic Journal*, 115(500): 224 - 46.

Table 1: Alignment between reported preferences and true preferences

	True preferences reported	*True preferences not reported*			
		Chose strategically	Made mistake	Other	
Full sample	83.33%	5.38%	0.36%	10.93%	n = 558
Gender					
Male	84.69%	6.80%	0.00%	8.50%	n = 294
Female	81.82%	3.79%	0.76%	13.64%	n = 264
Relationship status					
Single	86.87%	5.56%	0.51%	7.07%	n = 198
Long-term relationship	76.68%	6.22%	0.52%	16.58%	n = 193
Married	86.75%	4.22%	0.00%	9.04%	n = 166
Dual-match participation					
Regular applicant	84.10%	5.56%	0.38%	9.96%	n = 522
Dual-match applicant	72.22%	2.78%	0.00%	25.00%	n = 36
Age (median = 26)					
Below median	82.75%	7.03%	0.32%	9.90%	n = 313
Above median	84.10%	3.35%	0.42%	12.13%	n = 239
College GPA (median = 3.8)					
Below median	82.65%	6.46%	0.68%	10.20%	n = 294
Above median	84.21%	4.05%	0.00%	11.74%	n = 247
MCAT (median = 32)					
Below median	83.92%	4.90%	0.35%	10.84%	n = 286
Above median	85.71%	3.57%	0.45%	10.27%	n = 224
MLE Step 1 (median = 228)					
Below median	82.85%	6.20%	0.00%	10.95%	n = 274
Above median	85.39%	3.37%	0.75%	10.49%	n = 267
MLE Step 2 (median = 241)					
Below median	83.46%	6.02%	0.00%	10.53%	n = 266
Above median	84.52%	3.57%	0.79%	11.11%	n = 252
U.S. applicants / positions in specialty (median = 0.89)					
Below median	85.63%	4.06%	0.31%	10.00%	n = 320
Above median	80.26%	7.30%	0.43%	12.02%	n = 233

Notes: This table summarizes respondents' self-assessed reporting practices, broken down by demographic groups. Question text: "When forming the ranking of residencies to submit to the NRMP, some candidates submit an ordering that is not the true order of how desirable they find the programs. When forming your list, did you report the exact ordering of your true preferences?" Available multiple-choice responses: "Yes"; "No – I chose my list strategically"; "No – I tried to report my true preferences, but I made a mistake"; "No – Other reason".

Table 2: Predictors of nontruthful reporting behavior

	(1)		(2)	
	Multinomial Logit		Avg. Marginal Effects	
Predicted response	Strategic	Other	Strategic	Other
Female	-0.585	0.677**	-0.032*	0.065**
	(0.4149)	(0.2963)	(0.0191)	(0.0269)
Long-term relationship	0.381	0.927**	0.014	0.088***
	(0.4552)	(0.3614)	(0.0229)	(0.0336)
Married	0.091	0.105	0.004	0.007
	(0.5328)	(0.4243)	(0.0248)	(0.0298)
Dual-match participant	-0.676	0.975**	-0.031	0.124*
	(1.0608)	(0.4456)	(0.0282)	(0.0689)
Age	-0.202*	0.093**	-0.011*	0.010**
	(0.1142)	(0.0441)	(0.0058)	(0.0040)
Academic ability index	-0.432***	-0.117	-0.021***	-0.008
	(0.1307)	(0.1477)	(0.0067)	(0.0132)
Specialty's excess applicants	0.270*	0.045	0.013*	0.003
	(0.1378)	(0.1384)	(0.0069)	(0.0125)
Constant	2.630	-5.458***		
	(2.9699)	(1.2392)		
N	544	544	544	544

Notes: Standard errors in parentheses. Column 1 presents multinomial logit regression coefficients. Column 2 presents the associated average marginal effects, measured relative to the baseline of truthful reporting. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Validating responses on truthful reporting

Panel A

	(1)	(2)	(3)
	Predicted Variable: Preference Ordering		
Predictor:	Life Assessment	Life Satisfaction	Happiness
$\beta_{Truthful}$	9.03***	7.69***	5.32***
	(0.508)	(0.478)	(0.427)
$\beta_{Strategic}$	4.47***	6.16***	4.14***
	(1.099)	(1.433)	(1.289)
β_{Other}	3.20***	4.95***	2.36***
	(0.836)	(1.068)	(0.867)
N	2179	2179	2178
p: $\beta_{Truthful} = \beta_{Strategic}$	0.00	0.31	0.39
p: $\beta_{Truthful} = \beta_{Other}$	0.00	0.02	0.00

Panel B

	(1)	(2)	(3)	(4)
	Predicted Variable: Preference Ordering			
Predictor:	$\bar{U}$	$\bar{U}$	$\bar{U}$	$\bar{U}$
$\bar{U}$ estimation sample:	Full sample	Truthful reporters	Full sample	Truthful reporters
$\bar{U}$ weighted attributes:	9 non-SWB	9 non-SWB	All 12	All 12
$\beta_{Truthful}$	1.08***	1.00***	1.12***	1.00***
	(0.054)	(0.050)	(0.053)	(0.047)
$\beta_{Strategic}$	0.67***	0.57***	0.62***	0.50***
	(0.144)	(0.125)	(0.130)	(0.107)
β_{Other}	0.78***	0.68***	0.65***	0.51***
	(0.118)	(0.105)	(0.098)	(0.081)
N	2153	2153	2150	2150
p: $\beta_{Truthful} = \beta_{Strategic}$	0.01	0.00	0.00	0.00
p: $\beta_{Truthful} = \beta_{Other}$	0.02	0.01	0.00	0.00

Notes: Standard errors in parentheses. Each column presents rank-order logit coefficients from a model predicting residency preference orderings using the variable in the column header, with separate coefficients estimated for the different self-reported truth-telling statuses of table 1. Individuals reporting nontruthful reporting due to a mistake are excluded. The bottom two rows of each panel report p-values for Wald tests of the null hypotheses that $\beta_{Truthful} = \beta_{Strategic}$ and $\beta_{Truthful} = \beta_{Other}$. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A Appendix Tables

Table A1: PCA scoring coefficients for academic ability index

Variable	Scoring coefficient
MLE Step 1 score	0.4045
MLE Step 2 score	0.3418
College GPA	0.1379
MCAT score	0.3506
MLE Step 1 nonresponse	- 0.4324
MLE Step 2 nonresponse	- 0.3493
College GPA nonresponse	- 0.3728
MCAT nonresponse	- 0.3601

Notes: Scoring coefficients from the principle component analysis of academic performance measures. Included were the four measures of academic performance, as well as dummy variables indicating non-response for each of the four measures. The resulting index is standardized before inclusion in regressions.

Table A2: Attribute prompts

Variable label	Question prompt (beginning “On a scale from 1 to 100, . . . ”)
Life Assessment	. . . where 1 is “worst possible life for you” and 100 is “best possible life for you” where do you think the residency would put you?
Life Satisfaction	. . . how satisfied do you think you would be with your life as a whole while attending this residency?
Happiness	. . . how happy do you think you would feel on a typical day during this residency?
Prestige/Status	. . . how would you rate the prestige and status associated with this residency?
Social Life	. . . what would you expect the quality of your social life to be during this residency?
Location	. . . taking into account city quality and access to family and friends, how desirable do you find the location of this residency?
Anxiety	. . . how anxious do you think you would feel on a typical day during this residency?
Worthwhile life	. . . to what extent do you think your life would seem worthwhile during this residency?
Stress	. . . how stressed do you think you would feel on a typical day during this residency?
Career Prospects	. . . how would you rate your future career prospects and future employment opportunities if you get matched with this residency?
Control	. . . how do you expect this residency to affect your control over your life?
Desirable for SO	. . . how desirable is this residency for your spouse or significant other?

Notes: Question prompts for the 12 residency attribute questions assessed in section 2.2. Table reproduced from Benjamin, Heffetz, Kimball, and Rees-Jones (2014).

Table A3: Rank-order logit estimates for revealed-preference utility measures

	(1)	(2)	(3)	(4)
	Predicted Variable: Preference Ordering			
Prestige/Status	2.52***	2.67***	2.51***	2.68***
	(0.337)	(0.385)	(0.346)	(0.398)
Social Life	1.55***	1.99***	0.39	0.40
	(0.311)	(0.364)	(0.338)	(0.399)
Location	1.71***	1.95***	1.07***	1.22***
	(0.230)	(0.270)	(0.242)	(0.288)
Anxiety	-0.26	-0.14	0.22	0.28
	(0.307)	(0.340)	(0.320)	(0.357)
Worthwhile life	4.42***	5.22***	1.88***	2.45***
	(0.520)	(0.617)	(0.585)	(0.704)
Stress	-0.14	-0.40	0.32	0.13
	(0.313)	(0.355)	(0.326)	(0.377)
Career Prospects	3.21***	3.71***	2.80***	3.39***
	(0.513)	(0.592)	(0.529)	(0.621)
Control	0.40	0.60*	0.06	0.18
	(0.303)	(0.352)	(0.320)	(0.377)
Desirable for SO	2.56***	2.27***	2.48***	2.15***
	(0.264)	(0.308)	(0.277)	(0.329)
Life Satisfaction			3.32***	3.18***
			(0.518)	(0.595)
Happiness			1.91***	2.35***
			(0.498)	(0.573)
Life Assessment			3.16***	4.68***
			(0.506)	(0.632)
N	2169	1797	2166	1796

Notes: Standard errors in parentheses. This table presents coefficient estimates from the rank-order logit models used to create the utility proxies assessed in panel B of table 3. Columns 1 and 3 are estimated from the full sample, and columns 2 and 4 are estimated solely from respondents indicating truthful preference reporting behavior. All attribute ratings are divided by 100 before inclusion in the regression. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Validating responses on truthful reporting: quantifying effect size

Panel A

	(1)	(2)	(3)
	Predicted Variable: Preference Ordering		
Predictor:	Life Assessment	Life Satisfaction	Happiness
$\hat{\Pr}(A \succeq B \vert Truthful)$	0.80^{***}	0.76^{***}	0.69^{***}
	(0.013)	(0.013)	(0.014)
$\hat{\Pr}(A \succeq B \vert Strategic)$	0.67^{***}	0.72^{***}	0.65^{***}
	(0.038)	(0.044)	(0.044)
$\hat{\Pr}(A \succeq B \vert Other)$	0.62^{***}	0.68^{***}	0.59^{***}
	(0.031)	(0.035)	(0.031)
N	2179	2179	2178

Panel B

	(1)	(2)	(3)	(4)
	Predicted Variable: Preference Ordering			
Predictor:	$\bar{U}$	$\bar{U}$	$\bar{U}$	$\bar{U}$
$\bar{U}$ estimation sample:	Full sample	Truthful reporters	Full sample	Truthful reporters
$\bar{U}$ weighted attributes:	9 non-SWB	9 non-SWB	All 12	All 12
$\hat{\Pr}(A \succeq B \vert Truthful)$	0.91^{***}	0.91^{***}	0.94^{***}	0.94^{***}
	(0.009)	(0.009)	(0.007)	(0.007)
$\hat{\Pr}(A \succeq B \vert Strategic)$	0.81^{***}	0.79^{***}	0.82^{***}	0.81^{***}
	(0.048)	(0.048)	(0.047)	(0.047)
$\hat{\Pr}(A \succeq B \vert Other)$	0.84^{***}	0.83^{***}	0.84^{***}	0.81^{***}
	(0.034)	(0.035)	(0.034)	(0.036)
N	2153	2153	2150	2150

Notes: The table presents calculations associated with a thought experiment meant to assist in quantifying the effect size implied by table 3. Consider a choice between two residencies, A and B. If residency A is rated 1 standard deviation higher than B according to the given welfare metric, what is the model's implied probability that A will be preferred to B? Given the assumption of a type-I extreme-value error distribution, this can be calculated as $\frac{e^{\beta*SD}}{1+e^{\beta*SD}}$, providing the estimates found above. Standard errors are in parentheses, and are calculated using the delta method. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B Consequences of Suboptimal Behavior

In this appendix, I provide results which assist in bounding the consequences of nontruthful reporting in the DAA.

To begin, we will lay out the basic notation and definitions necessary for analyzing a two-sided matching market. While the notation below is general to other two-sided matching settings, let us refer to the the two groups being matched as students S and residencies R. Each student s_i has a preference ordering over residencies, denoted $\preceq_{s_i}$. Each residency r_i has a preference ordering over students, denoted $\preceq_{r_i}$, as well as a quota for the number of students it could accept, denoted Q_{r_i}. For both students and residencies, these preferences provide a complete ordering of the members of the opposite set, and how each compares to the possibility of being unmatched (denote by $\emptyset$).

Define a *matching* to be a single-valued function $\mathcal{M} : S \rightarrow \{R \cup \emptyset\}$, providing an assignment of each student to either a specific residency or to being unmatched. A matching is *feasible* if it does not assign any residency a number of students exceeding its quota—that is, $|\mathcal{M}^{-1}(r_i)| \leq Q_{r_i}$ for all r_i.

The difficulty of the matching problem is that preferences are not observed by the market organizer; instead, we must rely on reported preferences. Let T denote a vector encoding the reported preferences of all market participants, and let $\mathcal{T}$ denote the space of all possible sets of preferences. A *feasible mechanism* is a single-valued function $\phi : \mathcal{T} \rightarrow \mathcal{M}$, mapping all each vector of all reported types to a feasible matching.

The fundamental goal of the analysis to follow will be to assess the consequences of non-truthful behavior in the student-proposing DAA in particular, or strategy-proof mechanisms in general. Denote the student-proposing DAA as ϕ^{DAA}, with the algorithm implemented as described in Gale and Shapley (1962). Define a *strategy-proof mechanism* to refer to a feasible mechanism where it is a weakly dominant strategy for all students to report their true preferences.

Equipped with these basic definitions and notation, we may begin to explore the consequences of nontruthful play in this setting. A first result, previously referenced in section 1, bears repeating: the student-proposing DAA is strategy-proof for students (Dubins and Freedman, 1981; Roth 1982). It follows immediately that any change in the final matching induced by a falsely reported preference ordering can only make that student worse off.

While this suboptimal behavior does harm those who pursue it, under mild assumptions it is possible to bound the extent of harm done. These bounds are formalized in proposition 1 and its corollary.

Proposition 1. *Consider any strategy-proof mechanism. Let $M_s^{\preceq}$ denote the school to which student s would match if preferences $\preceq$ are reported, taking all other reported preferences as given. If the student has preference ordering $\preceq^T$ and submits preference ordering $\preceq^F$, the resulting school assignment $M_s^{\preceq^F}$ will satisfy i) $M_s^{\preceq^F} \preceq^T M_s^{\preceq^T}$, and ii) $M_s^{\preceq^T} \preceq^F M_s^{\preceq^F}$. That is, the resulting match is weakly less preferred to the truthful match according to true preferences, and weakly more preferred to the truthful match according to reported preferences.*

Proof. Condition i follows immediately from the assumption of a strategy-proof mechanism; if this condition did not hold, there would be scope for benefit from preference misrepresentation. To prove condition ii, assume for the sake of contradiction that $M_s^{\preceq^T} \succ^F M_s^{\preceq^F}$. If $\preceq^F$

were true preferences, reporting preferences $\preceq^T$ would result in a strictly preferred match. This contradicts the assumption that the mechanism is strategy-proof. □

Corollary 1. *Consider any strategy-proof mechanism. If a student would match after reporting preferences truthfully, and if this student's reported preferences rank his truthful match above being unmatched (i.e., $\emptyset \preceq^T M_s^{\preceq^T}$), then this student will not become unmatched due to his reporting pattern.*

Proof. Follows immediately from proposition 1. □

Corollary 1 provides a degree of protection to unsavvy students in many school-choice environments. Often, the primary welfare determinant is not *where* the student matches, but *whether* a student matches. For example, in the residency-choice context, matching to a program several spots lower on one's preference ordering will not seriously jeopardize the student's career path or lifetime income. In contrast, failing to match will severely impede career progress, and have substantial effects on lifetime income. Corollary 1 shows that while nontruthful reporting can harm an unsavvy student, it cannot cause that student to experience the worst possible outcome under many plausible ways in which preferences could be misrepresented. Furthermore, proposition 1 guarantees that the fall in truthful preference rankings that could be experienced is bounded by the largest difference between true and reported rankings, which is small under many of the misrepresentation heuristics considered in section 2.

These results demonstrate that, while suboptimal behavior is of course harmful to a student, the nature of strategy-proof mechanisms provides inherent protections against these consequences. It is worth noting, however, that these protections do not extend to the other participants in this market. In particular, a truth-telling student can be severely harmed by another student's misrepresentation, as is demonstrated in the following example.

Example 1. *Consider a matching problem with three students (denoted A, B, and C) matching to two residencies (denoted 1 and 2). Let preferences be assigned according to appendix table 5 below, and final matches be determined by the student-proposing DAA.*

Table 5: Preferences in example 1

Residency	$\preceq_{r_i}$	Student	$\preceq_{s_i}$
1	B ≺ C ≺ A	A	2 ≺ 1
2	C ≺ B ≺ A	B	1 ≺ 2
		C	2 ≺ 1

In this case, truthful reporting of preferences will result in student A matching with residency 1, student B matching with residency 2, and student C remaining unmatched. If we instead assume that student A misrepresents his preferences by reversing his ordering of the two residencies, the new result of the student-proposing DAA would assign student C to residency 1, student A to residency 2, and would leave student B unmatched. Notice that student C has benefited from A's misrepresentation, going from being unmatched to being assigned his first choice. In contrast, student B was harmed by A's misrepresentation, going from his first choice to being unmatched.

Example 1 demonstrates that truth-telling students may either gain or lose from another student's misrepresentation. Furthermore, the potential losses they might face do not have the same favorable bounds previously derived for the student making the misrepresentation. However, notice the mechanic which permits this outcome to occur: in this example, both students and residencies have meaningful heterogeneity in their preferences. To construct examples where truth-tellers are harmed from another student's misrepresentation, significant idiosyncrasies in preferences are needed. If we instead consider an application of the student-proposing DAA in which all residencies share a common preference ordering over students, and all students share a common preference ordering over residencies, truth-telling students cannot be harmed by another student's misrepresentation. If a student misrepresents his preferences, then the rank distribution of residency assignments for the truth-telling students first order stochastically dominates the rank distribution that would have been achieved under truthful preference reporting.