Games and Economic Behavior ••• (••••) •••–•••

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match

Alex Rees-Jones [1]

The Wharton School, University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 10 January 2016
Available online xxxx

JEL classification:
C78
D03

Keywords:
Matching
Deferred acceptance algorithm
Suboptimal behavior

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 I 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.

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,[2] which have been empirically shown to be detrimental to the success of these markets (Roth, 1990, 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, truth-telling is approximately optimal for all market participants in sufficiently large markets (Immorlica and Mahdian, 2005; Kojima and Pathak, 2009; Azevedo and Budish, 2013). Strategy-proof mechanisms therefore provide a comparatively simple optimal strategy, which has been viewed as

E-mail address: alre@wharton.upenn.edu.

[1] 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 the AEA annual meetings, 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.

[2] 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.

http://dx.doi.org/10.1016/j.geb.2017.04.011
0899-8256/

especially useful in the student-to-school matching setting. If optimal play is pursued, students may entirely avoid 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 et al., 2005a, 2005b).

This paper explores empirically whether the benefits of strategy-proof mechanisms are fully realized. The typically expressed logic suggests that incentivizing the simple truthful reporting strategy 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. Among unsophisticated students, failures of optimal behavior might arise.[3] 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. All else equal, 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 the DAA, and commonly finds a fraction of respondents with nontruthful reporting behavior (see, e.g., Chen and Sönmez, 2006; Pais and Pintér, 2008; Calsamiglia et al., 2010; Klijn et al. 2013; Ding and Schotter, 2015; Featherstone and Niederle, 2016). 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.[4] The validated self-classification approach presented in this paper offers a rare demonstration that failures of truthful reporting persist outside of the lab. These results complement the concurrent work of Hassidim et al. (2016), who study the 2014 roll-out of a DAA matching mechanism in the Israeli psychology match. The authors find that submitted preferences commonly rank an unfunded position higher than the exact same position with funding. Under the reasonable assumption that students prefer more money to less, this finding implies a high lower bound on the rate of suboptimal preference reporting in this nascent matching market. Taking our results together, Hassidim et al. (2016) demonstrate that substantial misunderstanding of optimal play exists when these mechanisms are first deployed, whereas the results presented here demonstrate that this misunderstanding persists even after decades of institutional history and refinement of training interventions. In summary, failures of optimal play persist in perhaps the most well-studied and carefully designed two-sided matching market 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 understanding of the decision-making environment is developed, and when stakes are sufficiently high and outcomes are sufficiently scrutinized.[5] The population 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 to the design of mechanisms that can minimize misunderstanding.

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 the practical deployment of matching mechanisms.

[3] However, a lack of sophistication does not necessarily result in suboptimal reporting. Even absent an understanding of the mechanism, truth-telling could arise from, e.g., moral considerations, reliance on correct advice, or the utilization of truth-telling as a default strategy when the optimal strategy is not understood.

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

[5] For discussions of this line of logic (as it applies to interpreting and contrasting lab and field experiments) see Levitt and List (2006, 2007, 2008).

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

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.

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, 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.[6]

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 (median number of applications: 29).[7] Interested programs invite the student to visit and interview with program representatives (median number of interviews offered/attended: 15/11). Both the students and the residencies use these interviews to gather information about their potential match partners.[8] Following this interview process, students and residency representatives both determine their preferences over possible matches with whom they interviewed. These preferences are submitted to the NRMP at a coordinated time, and a binding match is announced several weeks later.

As a result of the pre-screening process, the student's task when submitting preferences is simply to rank-order the comparatively small set of schools with which they interviewed (median number of programs ranked: 11).[9] In the process of determining that rank-ordering, students receive a substantial amount of advice. The NRMP website explicitly describes the incentive compatibility of reporting true preferences, contains tutorials on the DAA, and advertises that research on its algorithm contributed to the awarding of the 2012 Nobel Prize in Economics. Furthermore, medical schools often offer significant advising to their students as they undergo this process, and previous generations of medical students commonly share their own experience and advice.[10] In short, decisions in this environment are not made in isolation. Substantial explanation of the mechanism, advice, and institutional history may be drawn upon to inform preference reporting.

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 et al., 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

[6] 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.

[7] Application statistics are presented for U.S. seniors that successfully matched, and are drawn from the 2013 NRMP Applicant Survey (NRMP, 2013).

[8] Several papers have criticized the conduct of student interviews, as they open the possibility for programs to pressure students to rank them highly in exchange for a high ranking on their own list (e.g., Fisher, 2009; Nagarkar and Janis, 2012). Behavior of this sort is expressly forbidden by the NRMP, but may persist none-the-less. To help assess the importance of this potential cause of nontruthful reporting, participants in my study were asked "During your interviews, did any school representative offer to rank you higher in exchange for a high position on your ranking?". Only 16 respondents (3%) indicated that this had occurred, and of those only 3 respondents indicated that they nontruthfully reported preferences.

[9] Students are permitted to list up to 700 programs on their rank-order list, although additional fees apply once the students' list exceeds 20 programs.

[10] Of course, the advice received from previous match participants need not be good advice. In an experimental study of the DAA, Ding and Schotter (2015) find that facilitating intergenerational advice reduces the rate of truth-telling over time.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

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.[11] 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 et al. (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 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.[12] 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.[13] 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.[14]

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 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 performing students to

[11] 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).

[12] 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.

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

[14] Appendix Table A.1 presents the relationship between true preference orderings and submitted preference orderings for this group of respondents.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

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".

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.[15] 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.[16] Included predictor variables are dummy variables for gender, relationship status, and dual-match participation, as well as continuous 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 principal component analysis on college GPA and the three available test-scores.[17]

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

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

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

[17] Factor loadings are available in appendix Table A.2, and demonstrate that all 4 measures are positively associated with this index. Non-response on each item is negatively associated with the index.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

Table 2
Predictors of nontruthful reporting behavior.

Predicted response	(1)		(2)	
	Multinomial logit		Avg. marginal effects	
	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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

that better students tend to be more likely to play optimally. 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 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 truth-telling 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.[18] The full text of these questions is available in appendix Table A.3, 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 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 to mimic 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.

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

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

These data are used to create two groups of proxies for welfare, each with different strengths and weaknesses. The first group consists of the three SWB questions described above. While notions of "happiness" or "satisfaction" do not perfectly map to economists' notions of utility, they are often thought to contain some signal of underlying welfare. Consequently, these measures have been used to approximate economic utility in a variety of settings when choice data is unavailable or imperfect.[19] The second group consists of the predicted utility values estimated from a revealed-preference approach, rationalizing the preference orderings submitted to the NRMP with a latent utility function over residency attributes. While a choice-based approach of this sort may be considered the standard in economic evaluations, it does have limitations in this setting: if we believe that self-reports do reflect behavior—i.e., that those who indicated they misrepresented their preferences actually did so—then utility functions estimated from this preference data are suspect. However, as will be discussed below, the results of this analysis are still informative for validating the survey estimates of section 2.1.

I measure the association between a given welfare measure and the preference orderings reported to the NRMP with the rank-order logit model of Beggs et al. (1981). In this approach, 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 in closed-form.

Panel A of Table 3 estimates rank-order logit models where the residency 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.[20] Since the magnitudes of marginal utilities are measured relative to the error term in 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.[21] 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.

Panel B of Table 3 estimates rank-order logit models where the residency ordering is predicted by a constructed revealed-preference utility measure. 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 regression coefficients are reported in appendix Table A.4). These estimated models are used to calculate predicted values of the latent linear-utility index, $\bar{U} = \hat{\beta} X$, then used to predict choices in the second step in a manner analogous to panel A. Under the null hypothesis that all students truthfully report preferences, this approach provides valid estimates of the latent utility model. If we additionally impose the assumption that preferences are homogeneous (or heterogeneous in a manner independent of self-reported truth-telling status), this null hypothesis additionally predicts that the resulting welfare metric $\bar{U}$ would be equally predictive of the choices of those who say they tell the truth and those who do not. Finding that this measure is equally predictive would cast doubts on the validity of self-reported nontruthful behavior by failing to reject the null hypothesis that all reports were truthful. In contrast, a finding that $\bar{U}$ is less predictive of choice for the self-identified nontruthful reporters would support the validity of the self-reports by rejecting that null hypothesis.

Columns 1 and 3 of panel B conduct this test, differing in the attributes used to calculate the first-stage revealed-preference welfare metric $\bar{U}$. Column 1 predicts choice using the 9 non-SWB attributes whereas column 3 predicts choice using the 9 non-SWB attributes as well as the 3 SWB measures. Which model is preferred depends on your position on whether SWB measures are best treated as direct measures of utility or as valued abstract commodities that enter the utility function. However, either approach confirms that the estimated utility model is substantially more predictive for those who indicated truthful preference reporting. Of course, this rejection of the null hypothesis of universal truthful reporting calls into question the first-stage utility estimates used in this exercise. In columns 2 and 4 I conduct analogous exercises to columns 1 and 3, but exclude respondents self-reporting their behavior to be nontruthful. I then apply this estimated model to the full sample (an approach that again requires the assumption that preferences are homogeneous or heterogeneous in

[19] 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 et al., 2012, 2014; Perez-Truglia, 2015).

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

[21] To facilitate quantitative comparisons, appendix Table A.5 formally calculates these differences in probability as implied by my estimated models.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

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 manner independent of self-reported truth-telling status). As before, we see that self-classified nontruthful reporters have significantly weaker associations between these utility metrics and reported preferences.[22]

In summary, among those indicating nontruthful reporting, there is a systematically weaker link between reported preferences and welfare-relevant metrics—whether taken from direct statements of subjects' predicted well-being, or from revealed-preference approaches. While each approach has different weaknesses—i.e., adopting a SWB measure as an approximation to utility in panel A, or making strong assumptions about the nature of preference heterogeneity in panel B—taken together these differing approaches all suggest some failure of truthful preference reporting.

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 et al., 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 that 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

[22] Notice that since $\bar{U}$ was estimated from the self-reported truthful sample, the coefficient on the second-stage rank-order logit regression predicting choice with $\bar{U}$ is mechanically 1 for truthful reporters in these columns.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

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 that 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.[23]

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 that theoretical considerations suggest is suboptimal.

Market organizers often aim to create a decision environment that facilitates optimal decision-making. If employing an incentive compatible matching algorithm is not sufficient to achieve that goal, what other steps must be taken? What features of algorithms or markets lead to the belief that truthful behavior is, in fact, optimal? While our answers to these questions remain incomplete, two strands of recent literature offer guidance on these important questions and inform the interpretation the results presented in this paper.

In a theoretical approach to addressing these questions, Li (2015) defines the concept of "obviously strategy-proof" (OSP) algorithms—that is, algorithms that could be understood to be strategy proof without the application of contingent reasoning. This classification is shown to capture intuitive elements of what it means for an optimal strategy to be easy to understand. Furthermore, Li finds that mechanisms satisfying this property result in higher rates of truth-telling in the lab. These results offer one explanation for the persistence of suboptimal behavior in this setting: as proven in Ashlagi and Gonczarowski (2016), the DAA is not OSP, and indeed there exists no stable matching mechanism that is OSP. While this particular concept therefore cannot prescribe an alternative practical mechanism in this context, continued work in this spirit—formally defining the theoretical properties of "easy to understand" mechanisms, and aiming to develop and study mechanisms which satisfy those constraints—has promise in this setting.

In a more empirically focused approach to addressing these questions, a recent experimental literature has examined the comparative importance of advice for participants in strategy-proof matches. Provision of advice or information on others' play has been shown to significantly influence truth-telling behavior in lab experiments involving the DAA (Braun et al., 2014; Ding and Schotter, 2015) and the closely related "Top Trading Cycle" (TTC) mechanism (Guillen and Hing, 2014; Guillen and Hakimov, 2016b). Taking this line of logic one step further, Guillen and Hakimov (2016a) horserace two interventions meant to encourage optimal play in the TTC mechanism—one providing a detailed explanation of the mechanism, and one providing advice without explanation—and find that the provision of advice is substantially more effective at promoting truth-telling. In summary, this literature may be interpreted as demonstrating that the careful design and explanation of the mechanism alone is insufficient, and the provision of advice is a primary determinant of optimal play. Of course, participants in the medical match do commonly receive advice, and this advice may be responsible for the comparatively high rate of truth-telling observed in this study. The results of this paper demonstrate that room for improvement still remains.

Beyond the practical questions of optimal deployment of strategy-proof mechanisms, these results are relevant when assessing the costs and benefits of the DAA to non-strategy-proof alternatives. While strategy-proofness is a desirable property, it does not come for free; for example, Abdulkadiroğlu et al. (2011) demonstrate that the non-strategy-proof Boston mechanism can yield outcomes which that Pareto dominate those of the DAA. This suggests that the choice to implement

[23] 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.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

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 presence of (and punishment of) suboptimal behavior 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 (2016).

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 greater understanding of the theoretical consequences of these behavioral faults are necessary, some immediate results exist. As demonstrated in appendix section Appendix B, bounds may be derived on the extent to which nontruthful reporters are harmed by their suboptimal behavior. Furthermore, in environments where truth-telling is correlated with ability, and where schools rank students according to an imperfect signal of their ability, the presence of this suboptimal behavior has the potential to facilitate positive assortative matching (Rees-Jones, 2017). Of course, it need not be the case that good students are also good game theorists—however, some correlation between truth-telling status and measures of student ability has been found both in this paper and among the participants of the psychology match studied in Hassidim et al. (2016). Further attention to the perhaps nuanced channels through which this suboptimal behavior influences welfare will prove necessary as we continue to deploy two-sided matching mechanisms to the field.

Appendix A. Appendix tables

Table A.1
Preference modification among strategic respondents.

Submitted preferences	Frequency	Percent
1 > 2 > 4 > 3	5	19%
1 > 3 > 2 > 4	4	15%
1 > 3 > 4 > 2	3	11%
2 > 1 > 3 > 4	5	19%
2 > 1 > 4 > 3	4	15%
2 > 3 > 1 > 4	1	4%
2 > 3 > 4 > 1	1	4%
3 > 1 > 2 > 4	1	4%
3 > 1 > 4 > 2	1	4%
3 > 2 > 1 > 4	1	4%
3 > 4 > 2 > 1	1	4%
Total	27	

Notes: Survey respondents who indicated nontruthful preference reporting were asked to indicate their actual preference ordering over the top 4 choices they submitted to the NRMP. This table presents the true preference ordering over the top four residencies submitted to the NRMP by self-assessed strategic respondents. Residencies are denoted by their position in their true preference ranking over these four options. For example, the first row of this table describes respondents who thought their true preference ordering matched their submission to the NRMP for the first two choices, but who preferred their fourth-ranked residency on the NRMP submission to their third-ranked residency on the NRMP submission. Three students are excluded from this analysis, one with nonresponse on this question and two with invalid orderings.

Table A.2
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 principal 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.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

Table A.3
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 et al. (2014).

Table A.4
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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

Table A.5
Validating responses on truthful reporting: quantifying effect size.

Panel A

	(1)	(2)	(3)
	Predicted variable: preference ordering		
Predictor:	Life assessment	Life satisfaction	Happiness
$\hat{\Pr}(A \succeq B \mid Truthful)$	0.80***	0.76***	0.69***
	(0.013)	(0.013)	(0.014)
$\hat{\Pr}(A \succeq B \mid Strategic)$	0.67***	0.72***	0.65***
	(0.038)	(0.044)	(0.044)
$\hat{\Pr}(A \succeq B \mid Other)$	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:	$\tilde{U}$	$\tilde{U}$	$\tilde{U}$	$\tilde{U}$
$\tilde{U}$ estimation sample:	Full sample	Truthful reporters	Full sample	Truthful reporters
$\tilde{U}$ weighted attributes:	9 non-SWB	9 non-SWB	All 12	All 12
$\hat{\Pr}(A \succeq B \mid Truthful)$	0.91***	0.91***	0.94***	0.94***
	(0.009)	(0.009)	(0.007)	(0.007)
$\hat{\Pr}(A \succeq B \mid Strategic)$	0.81***	0.79***	0.82***	0.81***
	(0.048)	(0.048)	(0.047)	(0.047)
$\hat{\Pr}(A \succeq B \mid Other)$	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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Appendix 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 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 (denoted by $\emptyset$).

Define a *matching* to be a single-valued function $\mathcal{M} : S \to \{R \cup \emptyset\}$, 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, $\left|\mathcal{M}^{-1}(r_i)\right| \leq Q_{r_i}$ 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 $\phi : \mathcal{T} \to \mathcal{M}$, mapping 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 nontruthful 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 the introduction, 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.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

Table B.1
Preferences in Example 1.

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

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 B.1, and final matches be determined by the student-proposing DAA.

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.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011

References

Abdulkadiroğlu, Atila, Che, Yeon-Koo, Yasuda, Yosuke, 2011. Resolving conflicting preferences in school choice: the "Boston" mechanism reconsidered. Amer. Econ. Rev. 101 (1), 399–410.
Abdulkadiroğlu, Atila, Pathak, Parag, Roth, Alvin, 2005a. The New York city high school match. Am. Econ. Rev. Pap. Proc. 95, 364–367.
Abdulkadiroğlu, Atila, Pathak, Parag, Roth, Alvin, Sonmez, Tayfun, 2005b. The Boston public school match. Am. Econ. Rev. Pap. Proc. 95, 368–371.
Agarwal, Nikhil, 2015. An empirical model of the medical match. Amer. Econ. Rev. 105 (7), 1939–1978.
Ashlagi, Itai, Gonczarowski, Yannai, 2016. Stable matching mechanisms are not obviously strategy-proof. arXiv preprint arXiv:1511.00452.
Azevedo, Eduardo, Budish, Eric, 2013. Strategy-proofness in the large. Working paper.
Beggs, S., Cardell, S., Hausman, J., 1981. Assessing the potential demand for electric cars. J. Econometrics 17 (1), 1–19.
Benjamin, Daniel J., Heffetz, Ori, Kimball, Miles S., Rees-Jones, Alex, 2013. Survey appendix to: Can marginal rates of substitution be inferred from happiness data? Evidence from residency choices.
Benjamin, Daniel J., Heffetz, Ori, Kimball, Miles S., Rees-Jones, Alex, 2014. Can marginal rates of substitution be inferred from happiness data? Evidence from residency choices. Amer. Econ. Rev. 104 (11), 3498–3528.
Benjamin, Daniel J., Heffetz, Ori, Kimball, Miles S., Rees-Jones, Alex, 2012. What do you think would make you happier? What do you think you would choose? Amer. Econ. Rev. 102 (5), 2083–2110.
Braun, Sebastian, Dwenger, Nadja, Kübler, Dorothea, Westkamp, Alexander, 2014. Implementing quotas in university admissions: an experimental analysis. Games Econ. Behav. 85, 232–251.
Calsamiglia, Caterina, Haeringer, Guillaume, Klijn, Flip, 2010. Constrained School choice: an experimental study. Amer. Econ. Rev. 100 (4), 1860–1874.
Chen, Yan, Sönmez, Tayfun, 2006. School choice: an experimental study. J. Econ. Theory 127 (1), 202–231.
Ding, Tingting, Schotter, Andrew, 2015. Intergenerational advice and matching: an experimental study. Working paper.
Dubins, Lester, Freedman, David, 1981. Machiavelli and the Gale–Shapley Algorithm. Am. Math. Mon. 88 (7), 485–494.
Featherstone, Clayton, Niederle, Muriel, 2016. Boston versus deferred acceptance in an interim setting: an experimental investigation. Games Econ. Behav. 100, 353–375.
Fisher, Carl, 2009. Manipulation and the match. J. Am. Med. Assoc. 302 (12), 1266–1267.
Gale, David, Shapley, Lloyd, 1962. College admissions and the stability of marriage. Am. Math. Mon. 69, 9–15.
Guillen, Pablo, Hakimov, Rustamdjan, 2016a. How To Get Truthful Reporting in Matching Markets: A Field Experiment. WZB Discussion Paper No. SP II 2015-208.
Guillen, Pablo, Hakimov, Rustamdjan, 2016b. Not quite the best response: truth-telling, strategy-proof matching, and the manipulation of others. Exper. Econ., 1–17.
Guillen, Pablo, Hing, Alexander, 2014. Lying through their teeth: third party advice and truth telling in a strategy proof mechanism. Europ. Econ. Rev. 70, 178–185.
Hassidim, Avinatan, Romm, Assaf, Shorrer, Ran, 2016. 'Strategic' Behavior in a Strategy-Proof Environment. SSRN Working Paper No. 2784659.
Immorlica, Nicole, Mahdian, Mohammad, 2005. Marriage, honesty, and stability. In: Proceedings of the Sixteenth Annual ACM–SIAM Symposium on Discrete Algorithms, pp. 53–62.
Klijn, Flip, Pais, Joana, Vorstaz, Marc, 2013. Preference intensities and risk aversion in school choice: a laboratory experiment. Exper. Econ. 16 (1), 1–22.
Kojima, Fuhito, Pathak, Parag, 2009. Incentives and stability in large two-sided matching markets. Amer. Econ. Rev. 99 (3), 608–627.
Levinson, Arik, 2012. Valuing public goods using happiness data: the case of air quality. J. Public Econ. 96 (9–10), 869–880.
Levitt, Steven, List, John, 2006. What do laboratory experiments tell us about the real world? Working paper.
Levitt, Steven, List, John, 2007. What do laboratory experiments measuring social preferences reveal about the real world? J. Econ. Perspect. 21 (2), 153–174.
Levitt, Steven, List, John, 2008. Homo economicus evolves. Science 319 (5865), 909–910.
Li, Shengwu, 2015. Obviously Strategy-Proof Mechanisms. SSRN Working Paper No. 2560028.
Ludwig, Jens, Duncan, Greg J., Gennetian, Lisa A., Katz, Lawrence F., Kessler, Ronald C., Kling, Jeffrey R., Sonbonmatsu, Lisa, 2012. Neighborhood effects on the long-term well-being of low-income adults. Science 337 (6101), 1505–1510.
Luechinger, Simon, Raschky, Paul A., 2009. Valuing flood disasters using the life satisfaction approach. J. Public Econ. 93 (3–4), 620–633.
Luttmer, Erzo F.P., 2005. Neighbors as negatives: relative earnings and well-being. Quart. J. Econ. 120 (3), 963–1002.
McFadden, Daniel, 2009. The human side of mechanism design: a tribute to Leo Hurwicz and Jean-Jacque Laffont. Rev. Econ. Design 13, 77–100.
Nagarkar, Purushottam, Janis, Jeffrey, 2012. Fixing the "Match": how to play the game. J. Grad. Educ. 4 (2), 142–147.
National Resident Matching Program. 2012. National Resident Matching Program, Results and Data: 2012 Main Residency Match. National Resident Matching Program. Washington, DC.
National Resident Matching Program. 2013. Results of the 2013 NRMP Applicant Survey by Preferred Specialty and Applicant Type. National Resident Matching Program, Washington, DC.
Niederle, Muriel, Vesterlund, Lise, 2011. Gender and competition. Annu. Rev. Econ. 3, 601–603.
Pais, Joana, Pintér, Ágnes, 2008. School choice and information: an experimental study on matching mechanisms. Games Econ. Behav. 64 (1), 303–328.
Pathak, Parag, Sonmez, Tayfun, 2008. Leveling the playing field: sincere and sophisticated players in the Boston mechanism. Amer. Econ. Rev. 98, 1636–1652.
Perez-Truglia, Ricardo, 2015. A Samuelsonian validation test for happiness data. J. Econ. Psych. 49, 74–83.
Rees-Jones, Alex, 2017. Mistaken play in the deferred acceptance algorithm: implications for positive assortative matching. Am. Econ. Rev. 107 (5), 225–229. http://dx.doi.org/10.1257/aer.p20171028.
Roth, Alvin, 1982. The economics of matching: stability and incentives. Math. Oper. Res. 7 (4), 617–628.
Roth, Alvin, 1984. The evolution of the labor market for medical interns and residents: a case study in game theory. J. Polit. 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. Amer. Econ. Rev. 81, 415–440.
Roth, Alvin, 1996. The NRMP as a labor market. J. Am. Med. Assoc. 275, 1054–1056.
Roth, Alvin, 2008. Deferred acceptance algorithms: history, theory, practice, and open questions. In: Special Issue in Honor of David Gale on his 85th birthday. Int. J. Game Theory 36, 537–569.
Roth, Alvin, Peranson, Elliot, 1999. The redesign of the matching market for American physicians: some engineering aspects of economic design. Amer. Econ. Rev. 89 (4), 748–780.
Roth, Alvin, Xing, Xiaolin, 1994. Jumping the gun: imperfections and institutions related to the timing of market transactions. Amer. Econ. Rev. 84, 992–1044.
van den Berg, Bernard, Ferrer-i-Carbonell, Ada, 2007. Monetary valuation of informal care: the well-being valuation method. Health Econ. 16 (11), 1227–1244.
van Praag, Bernard M.S., Baarsma, Barbara E., 2005. Using happiness surveys to value intangibles: the case of airport noise. Econ. J. 115 (500), 224–246.

Please cite this article in press as: Rees-Jones, A. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ. Behav. (2017), http://dx.doi.org/10.1016/j.geb.2017.04.011