



## Management Science

Publication details, including instructions for authors and subscription information:  
<http://pubsonline.informs.org>

### Can Market Participants Report Their Preferences Accurately (Enough)?

Eric Budish, Judd B. Kessler

To cite this article:

Eric Budish, Judd B. Kessler (2022) Can Market Participants Report Their Preferences Accurately (Enough)?. Management Science 68(2):1107-1130. <https://doi.org/10.1287/mnsc.2020.3937>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact [permissions@informs.org](mailto:permissions@informs.org).

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2021, The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

# Can Market Participants Report Their Preferences Accurately (Enough)?

Eric Budish,<sup>a</sup> Judd B. Kessler<sup>b</sup>

<sup>a</sup> Booth School of Business, University of Chicago, Chicago, Illinois 60637; <sup>b</sup> The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Contact: [eric.budish@chicagobooth.edu](mailto:eric.budish@chicagobooth.edu),  <https://orcid.org/0000-0002-1483-8491> (EB); [judd.kessler@wharton.upenn.edu](mailto:judd.kessler@wharton.upenn.edu) (JBK)

Received: May 31, 2019

Revised: July 30, 2020

Accepted: September 22, 2020

Published Online in Articles in Advance:  
June 2, 2021

<https://doi.org/10.1287/mnsc.2020.3937>

Copyright: © 2021 The Author(s)

**Abstract.** In mechanism design theory it is common to assume that agents can perfectly report their preferences, even in complex settings in which this assumption strains reality. We experimentally test whether real market participants can report their real preferences for course schedules “accurately enough” for a novel course allocation mechanism, approximate competitive equilibrium from equal incomes (A-CEEI), to realize its theoretical benefits. To use market participants’ real preferences (i.e., rather than artificial “induced preferences” as is typical in market design experiments), we develop a new experimental method. Our method, the “elicited preferences” approach, generates preference data from subjects through a series of binary choices. These binary choices reveal that subjects prefer their schedules constructed under A-CEEI to their schedules constructed under the incumbent mechanism, a bidding points auction, and that A-CEEI reduces envy, suggesting subjects are able to report their preferences accurately enough to realize the efficiency and fairness benefits of A-CEEI. However, preference-reporting mistakes do meaningfully harm mechanism performance. One identifiable pattern of mistakes was that subjects had relatively more difficulty reporting cardinal as opposed to ordinal preference information. The experiment helped to persuade the Wharton School to adopt the new mechanism and helped guide aspects of its practical implementation, especially around preference reporting.

**History:** Accepted by Yan Chen, decision analysis.



**Open Access Statement:** This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as “*Management Science*.” Copyright © 2021 The Author(s). <https://doi.org/10.1287/mnsc.2020.3937>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>.

**Supplemental Material:** The online appendices are available at <https://doi.org/10.1287/mnsc.2020.3937>.

**Keywords:** market design • experiments • matching theory • course allocation • preference elicitation • combinatorial assignment • combinatorial allocation

## 1. Introduction

One of the exciting features of market design research is that it can help bring mechanisms designed using abstract microeconomic theory into practice to solve real-world resource-allocation problems. This feature has encouraged an explosion of research in matching and auction theory and has led to several well-known market design “success stories,” in which a mechanism has made it all the way from theory to practice. These include auctions for wireless spectrum around the world and matching mechanisms for entry-level medical labor markets, public schools, and organ transplantation.<sup>1</sup> To bring these market design mechanisms to practice often requires innovative academic research to help test the theory and evaluate its suitability for practice. In this spirit, this paper reports on a novel kind of laboratory experiment—based on bringing real market participants’ real preferences into the laboratory as opposed to inducing subject preferences as is typical in the market design experimental

literature—that tested a new market design and helped shepherd it from theory to practice.<sup>2</sup>

The context is the problem of combinatorial assignment—matching bundles of indivisible objects to agents without the use of monetary transfers, for example, matching students to schedules of classes—well known to be a difficult problem in market design. The theory literature on this problem contains mostly impossibility theorems that prove there is no perfect mechanism,<sup>3</sup> and the mechanisms used in practice have been shown to have critical flaws.<sup>4</sup> In an attempt to make progress on this problem, Budish (2011) proposes a new mechanism for combinatorial assignment, called approximate competitive equilibrium from equal incomes (A-CEEI). A-CEEI, unlike prior mechanisms, satisfies attractive properties of efficiency, fairness, and incentives though, as the name implies, only does so approximately.

At around the same time Budish (2011) was published, an opportunity to potentially implement a

new mechanism arose at the Wharton School at the University of Pennsylvania. Wharton's mechanism, a "fake-money" bidding points auction (BPA) used widely at many educational institutions,<sup>5</sup> was having the kinds of efficiency, fairness, and incentives problems one would expect given the theoretical criticisms of the mechanism (Sönmez and Ünver 2010), and the Wharton administration convened a committee to consider alternatives.

Although attractive in theory, however, A-CEEI makes an assumption that raises a serious concern about its suitability for use in practice: market participants can perfectly report their preferences to the mechanism. In Budish (2011), agents have ordinal preferences over all feasible bundles. As is standard in mechanism design theory (Fudenberg and Tirole 1991, Myerson 1991, Bergemann and Morris 2005), agents are assumed to be able to directly report these preferences to the mechanism. But this assumption often strains reality, and A-CEEI is such a case. In a context such as Wharton's, there might be hundreds of millions of feasible schedules in a given semester.

Clearly, in such combinatorial allocation settings, perfect preference reporting is an unrealistic goal, and whether market participants can report perfectly is an uninteresting question. Instead, the relevant question to answer before seriously considering bringing the theory to practice is whether market participants can report their preferences "accurately enough" to realize the benefits of the mechanism. Let us make this question more precise. In any practical implementation of the A-CEEI mechanism, participants cannot be expected to manually rank all schedules. Instead, participants must report a limited set of preference data—via what is known as a preference-reporting language (Milgrom 2009, 2011)—that can be used to construct an ordinal ranking over schedules. The question is whether participants can report such preference data with sufficient accuracy (i.e., whether the ordinal ranking generated by the preference data they report is close enough to their true preferences) that the efficiency and fairness benefits of A-CEEI are realized.

This positive question about A-CEEI's suitability, in turn, raises a deeper methodological question that pertains to market design more broadly. How can a researcher generate data that yields an assessment of preference reporting if agents' true preferences are fundamentally unknown? In the case of A-CEEI, how can we compare the ordinal ranking generated from the data that agents report to the mechanism to agents' true preferences? How can we measure the extent to which inaccurate preference reporting harms mechanism performance?

One potential approach would be to use the induced preferences methodology that is regularly employed in market design experiments. In an induced preferences experiment, the researcher gives subjects artificial

preferences and offers monetary rewards based on how well the subjects perform in the mechanism as evaluated based on these artificial preferences. For example, if, in a multi-object matching experiment, a subject is given an induced value for the bundle  $\{A, B\}$  of \$25 and then obtains the bundle  $\{A, B\}$  in the laboratory matching market, the subject would be compensated with a payment of \$25. Although this technique has been extremely important in the history of market design experiments and is invaluable for answering certain kinds of questions,<sup>6</sup> we found it to be limiting in our setting.

To see why, consider what we could learn about preference reporting to A-CEEI from experiments in which we use the induced preferences approach. If we induced preferences in a format that could be immediately reported to the mechanism (e.g., in the same language as the preference-reporting language used by A-CEEI in the laboratory), we would just be telling subjects their preferences and asking them to report them right back to us. This trivializes our central question about reporting complex preferences and instead tests whether subjects believe the advice in the experimental instructions that it is in their best interest to report their preferences truthfully (which is an interesting question in its own right; see Li 2017, Rees-Jones 2018, Rees-Jones and Skowronek 2018, Hassidim et al. 2021).

If we induce preferences in a format different from what could be reported back to the mechanism, this too misses the central question of interest. This exercise would be testing whether subjects can translate a small amount of information from one language that the researcher created (for conveying preferences to the subject) to another language that the researcher created (for reporting preferences back to the mechanism). For example, we could induce a subject with the preference ranking over bundles  $\{1, 2\} > \{1, 3\} > \{2, 3\} > \{1, 4\} > \{2, 4\} > \{3, 4\}$  and ask them to report item values for objects that express these preferences. Or we could give subjects more general preference goals—for example, a course in marketing is worth 100 points, a course in accounting is worth 80 points, you only want one course per major, early morning classes are worth 50 points less than afternoon classes, etc.—and ask subjects to use the preference-reporting language to express these induced preferences. Such experiments would certainly yield insight as to whether subjects have a basic grasp of a preference-reporting language. But they intrinsically cannot test whether real market participants can translate their own real preferences—however these preferences, over a combinatorially large set of outcomes, are represented in their own minds—into an A-CEEI preference report.

To be able to use real market participants' real preferences, we develop a new experimental design methodology, which we call the "elicited preferences" approach. The elicited preferences approach

collects data on subjects' real preferences by directly asking subjects binary comparison questions. In our setting, these questions were of the following form: Do you prefer Schedule A or Schedule B? When a subject chooses a preference for Schedule A or B, we treat this as the subject's true preference (i.e., just as we assume that a subject who is paid more to obtain Schedule A than Schedule B in an induced preferences experiment prefers Schedule A over Schedule B).<sup>7</sup> Put differently, our elicited preferences approach replaces the (reasonable) assumption of the induced preferences approach—that subjects prefer to earn more money from the experiment—with the (we argue also reasonable) assumption that subjects are able to report which of two schedules they prefer when asked directly.<sup>8</sup>

With the elicited preferences approach, we are able to bring real market participants' real preferences into the laboratory market. Specifically, our experimental subjects were Wharton MBA students who were asked to report their real preferences over schedules of real Wharton courses to A-CEEI using a realistic, professionally designed user interface. As described in detail in Section 2.5, we carefully tailored the binary comparisons we asked subjects in order to generate the necessary preference data to test preference-reporting accuracy and to test whether subjects could report preferences accurately enough to realize the efficiency and fairness benefits of A-CEEI relative to the incumbent bidding points auction. In addition, comparing the performance of the mechanism with regard to efficiency and fairness measures based on binary comparisons to efficiency and fairness measures based on what subjects report to A-CEEI, we can quantify the harm caused by preference-reporting mistakes.

There were two other advantages to using real market participants' real preferences. First, the realism enhanced the demonstration value of the experiment. Demonstration to policy makers who ultimately decide whether to implement a market design is a common goal of market design experiments (Roth 2015a); using real market participants' real preferences yields a more realistic, and thus more persuasive demonstration. Second, the realism facilitated a search for "side effects" of the mechanism; that is, issues left out of the theory that might be important for practice.<sup>9</sup> Issues left out of the theory are especially of concern here because A-CEEI had never been used before; many other market design implementations have had direct precedents that assuage these concerns.<sup>10</sup> Because our experimental subjects were real market participants who were playing in a realistic environment, we could search directly for side effects using surveys. The surveys, both quantitative and free-response, covered topics such as perceived fairness,

satisfaction with received schedule, ease of use, transparency, and overall "liking" of the mechanism.

An important disadvantage of our elicited preferences approach is that subjects' behavior is not incentivized.<sup>11</sup> This lack of incentives likely caused subjects to exert less effort in the laboratory than they would have if they were playing for real stakes, which, in turn, adds noise to subjects' behavior. We took care in the design to ensure that such noise pushes against finding accurate preference reporting and against our finding benefits of the A-CEEI mechanism so that our results on the efficiency and fairness gains of A-CEEI would constitute a lower bound (see Section 2.6 for a discussion).<sup>12</sup>

We briefly summarize the main results. Students reported their preferences accurately enough that A-CEEI outperformed the benchmark, the incumbent Wharton bidding points auction, on each of our quantitative measures of efficiency and fairness with most (though not all) differences statistically significant. The magnitudes were modest but all broadly consistent with the theory. However, we also find that subjects had significant difficulty with preference reporting (although large mistakes were comparatively rare) and that this difficulty meaningfully harmed mechanism performance. The efficiency and fairness improvement of A-CEEI over the bidding points auction would have been substantially larger if not for preference-reporting mistakes. The only negative side effect we found in the surveys was that students found A-CEEI to be something of a "black box," that is, nontransparent.

The experiment persuaded Wharton to adopt A-CEEI—implemented as "Course Match" beginning in Fall 2013—and guided several aspects of its practical implementation.<sup>13</sup> Some limited data from the first year of implementation underscores the external validity of our findings: A-CEEI increased equity in both total expenditure and the distribution of popular courses, and survey data suggest that A-CEEI has increased students' satisfaction with their assigned schedules, their perceptions of fairness, and their overall satisfaction with the course-allocation system. For example, the percentage of students responding that they found the course-allocation mechanism "effective" or "very effective" increased from 24% in the last year of the bidding points auction to 53% in the first year of A-CEEI, and the percentage of students who agreed or strongly agreed that the course allocation mechanism "allows for a fair allocation of classes" increased from 28% to 65%.

### 1.1. Contributions and Related Literature

Our paper makes four contributions to the market design literature. First, and most directly, the paper provides evidence on the efficacy of a specific mechanism

for course allocation: Budish's (2011) A-CEEI mechanism. Other work that discusses novel course-allocation mechanisms includes Sönmez and Ünver (2010), Budish and Cantillon (2012), Budish et al. (2013), Nguyen et al. (2016), Hashimoto (2018), Akbarpour and Nikzad (2020), and Nguyen and Vohra (2020).

Second, the paper contributes to an ongoing dialogue in the literature about the importance of preference reporting and language design (Milgrom 2009, 2011). We add to the burgeoning empirical literature on preference-reporting errors and the harm they can cause to a mechanism's performance (see, e.g., Rees-Jones 2018, Rees-Jones and Skowronek 2018, Hassidim et al. 2021). At the same time, we show that participants can report complex preferences accurately enough to realize the benefits of a mechanism with complex reporting requirements.

Third, the paper introduces a new experimental design methodology, the elicited preferences approach, which allows researchers to evaluate market designs in the laboratory using real market participants' real preferences by designing appropriate binary comparisons. This methodology can be used to evaluate other market designs with nontrivial preference-reporting requirements. This methodology may also be useful for evaluating decision supports for market designs, that is, tools that are designed to help participants more accurately report their preferences. Such decision supports play an important role not only in market designs with complex preference-reporting requirements, such as A-CEEI, but also in settings in which the preference reporting per se is simple but thinking through one's preferences is difficult, for example, school choice (cf. Narita 2016, Kapor et al. 2020). By comparing subjects' ability to report their preferences with and without a particular decision support, the elicited preferences approach can identify the efficacy of that decision support and help optimize the performance of existing market designs.

The elicited preferences methodology is a complement to the induced preferences methodology, which has been at the heart of a rich experimental literature in market design.<sup>14</sup> Within matching, experiments using induced preferences have explored decentralized markets (e.g., Echenique and Yarovitz 2013), including issues such as unraveling and congestion (e.g., Niederle and Roth 2009); the transition to centralized clearinghouses (e.g., Kagel and Roth 2000); and problems in those centralized clearinghouses, such as strategic misreporting (e.g., Castillo and Dianat 2016, Echenique et al. 2016) and clearinghouse collapse (e.g., McKinney et al. 2005). In addition, a rich line of experimental work has used induced preferences to explore school choice mechanisms in the laboratory, including work comparing

the performance of various mechanisms, such as deferred acceptance, the Boston mechanism, and top trading cycles (e.g., Chen and Sönmez 2006, Pais and Pinter 2008, Calsamiglia et al. 2010, Featherstone and Niederle 2016, Ding and Schotter 2017). Finally, induced preference laboratory experiments have been used to explore new matching mechanisms (e.g., Hakimov and Kesten 2018, Fragiadakis and Troyan 2019, Hakimov et al. 2019) and to explore new incentives criteria for market design (e.g., Li 2017, Chen et al. 2018).

Finally, our paper contributes a new theory-to-practice success story to the market design literature. This is valuable for two related reasons. The first reason is that market design implementations beget further market design implementations. The Wharton committee was already familiar with the work done by economists redesigning spectrum auctions and matching markets, and this gave the committee some comfort that economists might have something useful to say about their problem, too. Our specific market design implementation paves some new ground—the mechanism descends from general equilibrium theory as opposed to auction or matching theory, ordinary individuals are asked to report the kinds of complex preferences more commonly associated with high-stakes combinatorial auctions, and a laboratory experiment played a pivotal role in the adoption decision—so we have some hope that one day other researchers seeking to implement new market designs will be able to use our implementation as a helpful precedent, just as we used the spectrum auctions and matching markets as a helpful precedent.

The second reason, as emphasized by Roth (2002), is that academic work on the practical implementation of market design theory is an important complement to the theory itself. This work shows whether a particular theory is robust and raises new questions for theory to consider (e.g., the optimal design of preference-reporting languages). As Roth (2002, p. 1342) writes, "Whether economists will often be in a position to give highly practical advice depends in part on whether we report what we learn, and what we do, in sufficient detail to allow scientific knowledge about design to accumulate... If the literature of design economics does mature in this way, it will also help shape and enrich the underlying economic theory."

## 1.2. Organization of the Paper

The remainder of this paper is organized as follows. Section 2 describes the experimental design. Section 3 provides initial data on subjects' preference-reporting ability and presents our results on fairness and efficiency. Section 4 analyzes preference-reporting mistakes. Section 5 reports on the survey data and the search for unintended consequences of the mechanism.

Section 6 reports on the first year of practical implementation and concludes.

## 2. Experimental Design

### 2.1. Real Market Participants' Real Preferences

Our experimental subjects were Wharton MBA students, recruited by an email sent by the Wharton administration (see Online Appendix A).<sup>15</sup> There were 132 subjects over eight experimental sessions, conducted in a computer laboratory at Wharton during the week of November 28, 2011 (see the full text of the experimental instructions in Online Appendix C).

Subjects were given a list of 25 Wharton course sections for the upcoming Spring 2012 semester. These courses were chosen by the Wharton Course Allocation Redesign Team (the “Wharton committee”) to be representative of course offerings in the upcoming semester with a tilt toward popular courses (see the list of courses and sample descriptions in Online Appendix D). Each course section had a capacity of three to five seats.

Subjects were instructed that they would participate in two course-allocation procedures, Wharton's current system and an alternative system, and that their goal in the study was to use each system to obtain the best course schedule they could given their own true preferences. Here is some of the key text from the experimental instructions:

While using each system, please imagine that it is the spring term of your second year at Wharton, so this will be your last chance to take Wharton classes. Please try to construct your most preferred schedule given the courses that are available.

In real life, we know you take these decisions very seriously. We ask that you take the decisions in this session seriously as well. We will provide you with time to think carefully while using each system.

We then gave subjects five minutes to look over the course offerings and think about their preferences before describing the first mechanism.

### 2.2. Flow of Each Experimental Session

In half of the sessions, we ran the BPA first, and for half of the sessions, we ran A-CEEI first.<sup>16</sup> Details of the mechanisms are in Sections 2.3 and 2.4, respectively. For each mechanism,

- i. We read aloud the instructions for that specific mechanism.
- ii. Subjects participated in that mechanism to assemble a schedule of spring 2012 courses (starting from a blank slate for each mechanism).
- iii. Subjects responded to Likert-scale survey questions about their experience with the mechanism. See Section 5 and Online Appendix J for details of the surveys.

After subjects had participated in both mechanisms,

- i. Subjects performed a series of binary comparisons between pairs of schedules. These binary comparisons were designed to provide measures of efficiency, fairness, and preference-reporting accuracy. See Section 2.5 for details of the binary comparisons.

- ii. Subjects responded to Likert-scale survey questions comparing the two mechanisms.

- iii. Subjects provided free-form response comments.

### 2.3. Wharton BPA

At the time of the experiment, Wharton's bidding points auction, a variant on the bidding points auction mechanism used at a wide variety of educational institutions (Sönmez and Ünver 2010), worked as follows. In the first round of the BPA, students would submit bids for courses with the sum of their bids not to exceed their budget (of an artificial currency called bidding points). If a course had  $k$  seats, the  $k$  highest bidders for that course obtained a seat and paid the  $k + 1^{\text{st}}$  highest bid. After this first bidding round, there were then eight additional rounds, spaced over a period of time lasting from the end of one semester to the beginning of the next, in which students could both buy and sell courses using a double auction.<sup>17</sup> In each round of the double auction, for each course, all offers to buy were aggregated into a demand curve, all offers to sell were aggregated into a supply curve (with empty seats treated as additional supply offered at an ask price of zero), and if demand and supply crossed, trades would be executed at the lowest market-clearing price (i.e., a 0-DA in the terminology of Rustichini et al.(1994)).

Our laboratory implementation of the BPA was as similar as possible to the real Wharton bidding points auction, subject to the constraints of the laboratory. For time considerations, we used four rounds instead of nine.<sup>18</sup> For the first round, subjects were given five minutes to select their bids with an initial budget of 5,000 points. For the remaining three rounds, subjects were given two and a half minutes to select their bids and asks. The experiment used the standard web interface of the real Wharton bidding points auction so that it would be as familiar as possible to subjects. The instructions for the BPA were familiar as well because all subjects had previously used the real Wharton bidding points auction mechanism to pick their courses. (See Online Appendix C, “Instructions for Course Auction.”)

### 2.4. A-CEEI

A-CEEI has four steps: (i) students report their preferences, (ii) each student is assigned an equal budget (5,000 points in the experiment) plus a small random amount (used to break ties),<sup>19</sup> (iii) the computer finds (approximate) market-clearing prices,

and (iv) each student is allocated the student's most preferred affordable schedule—the affordable schedule the student likes best given the student's report in step (i) based on the student's budget set in step (ii) and the prices found in step (iii).<sup>20</sup>

The instructions described the A-CEEI mechanism, which was unfamiliar to the subjects, and explained to subjects that their only responsibility in using the mechanism was to tell the computer their true preferences; the computer would then compute market-clearing prices and buy them the best schedule they could afford at those prices. Because our interest was in whether subjects could report their preferences accurately enough to realize the theoretical benefits of the A-CEEI mechanism—and not in testing whether subjects could infer the strategy-proofness of the mechanism—we explicitly instructed subjects to be as truthful as possible in their preference reporting. The instructions advised students "... you do not need to think about the prices of the courses or the values that other students assign to courses. You get the best schedule possible simply by telling the computer your true values for courses."<sup>21</sup> The instructions used the metaphor of providing instructions to someone shopping on your behalf to explain the rationale for reporting one's true preferences as accurately as possible. (See Online Appendix C, "Instructions for the Course Matching System.")

**2.4.1. Preference-Reporting Language.** As discussed in the introduction, A-CEEI requires an ordinal ranking over all feasible schedules from each agent so that the mechanism can always select the agent's most preferred affordable bundle from any possible choice set. In any practical implementation of A-CEEI, agents cannot be expected to directly report preferences over all possible bundles. Instead, agents need to report a more limited set of information that describes their preferences, using a language provided as part of the mechanism implementation (cf. Milgrom 2011).

The preference-reporting language we implemented in the laboratory, a simplified version of the language proposed in Othman et al. (2010) and similar in spirit to the language proposed in Milgrom (2009), had two components. First, subjects could report cardinal item values, on a scale of 1 to 100, for any course section they were interested in taking; if they did not report a value for a course section, its value was defaulted to zero.<sup>22</sup> Second, subjects could report "adjustments" for any pair of course sections. Adjustments assigned an additional value, either positive or negative, to schedules that had both course sections together. Adjustments are a simple way for students to express certain kinds of substitutabilities and complementarities.<sup>23</sup> Subjects did not need to report schedule constraints, which were already known by the system.

The user interface for this language, designed by Wharton information technology professionals, is displayed as Figure 1.

To calculate a subject's utility for a schedule, the system summed the subject's values for the individual courses in that schedule together with any adjustments (positive or negative) associated with pairs of courses in the schedule. The subject's rank order list over all schedules could, thus, be obtained by ordering schedules from highest to lowest utility.<sup>24</sup> Observe that this means that the cardinal preference information that subjects submit for individual courses and pairs of courses induces an ordinal ranking over all feasible schedules.

We emphasize that, although both we and the Wharton committee believed this preference-reporting language to be reasonable—in particular, the Wharton committee felt strongly that adding more ways to express nonadditive preferences would make the language too complicated—there is no reason to believe that this preference-reporting language is optimal. As we discuss in the conclusion, optimal language design is an interesting open question for future research.

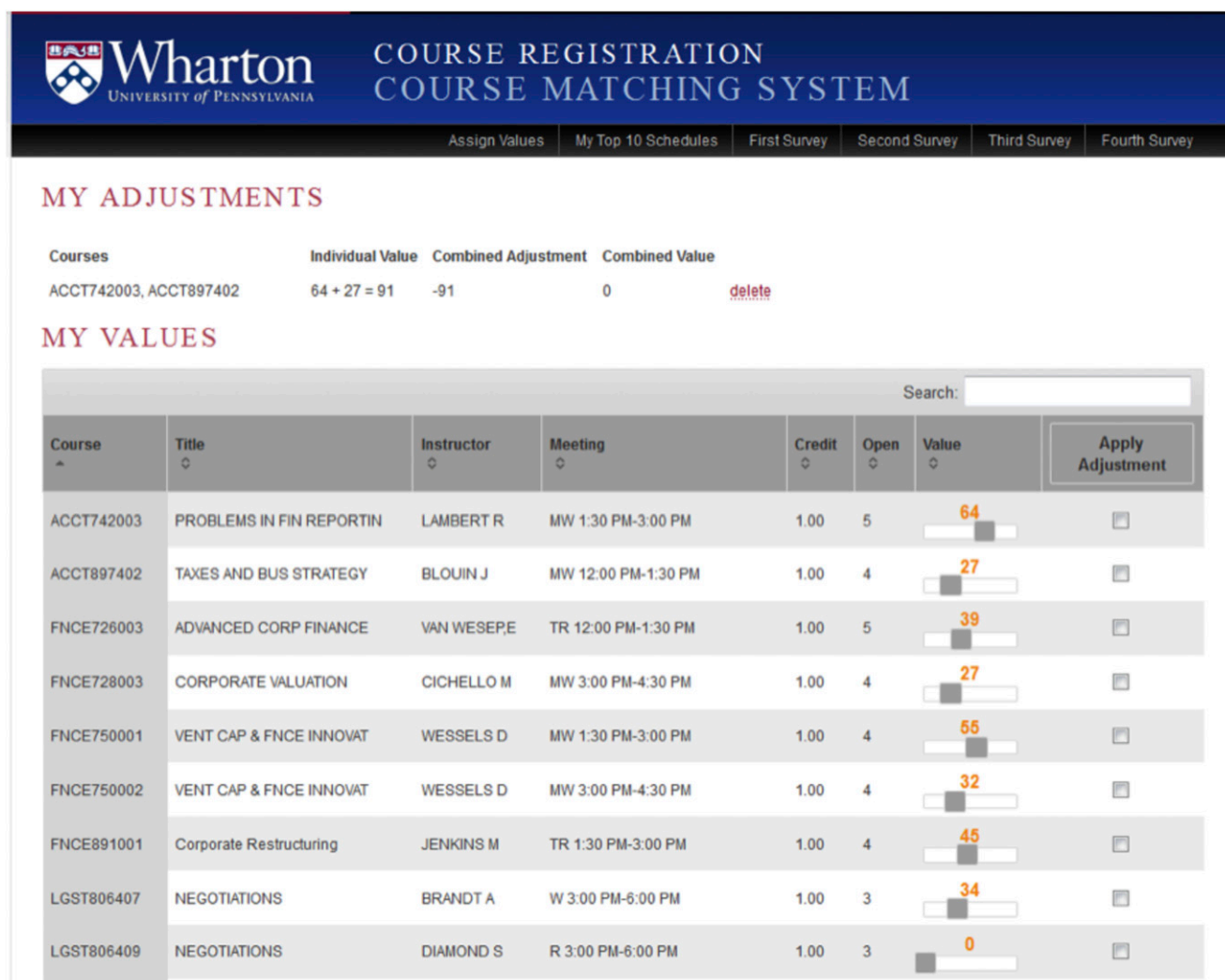
Given the complexity of preference reporting and, in particular, the complexity of translating cardinal item values and adjustments into an ordering over schedules, we provided subjects with a decision support tool, the "top-10 widget," which allowed them to translate the preference information they had provided so far into a list of what the system currently calculated to be their 10 most preferred schedules (displayed in order with the accompanying sum of the cardinal utilities and adjustments next to each schedule). Subjects could use this widget at any time while reporting their values and could go back to make modifications to their values, for example, if they realized the 10 schedules listed were not their favorites or were in the wrong order. Students were given 10 minutes to report their preferences.

## 2.5. Binary Comparisons

A simple methodological innovation, binary comparisons, is what allowed us to elicit preference data that reflects market participants' real preferences. The logic behind the methodology is that, although reporting ordinal preferences over every possible schedule using the preference-reporting language is cognitively complex and all but certain to be somewhat inaccurate, making a binary comparison between two specific schedules is cognitively simple and likely to accurately reflect true preferences.

After using both mechanisms, subjects were shown up to 19 pairs of schedules and asked to report which of the two schedules they preferred on a scale of "Strongly Prefer," "Prefer," and "Slightly Prefer" for each schedule. See Figure 2 for a screenshot.

Figure 1. (Color online) Screenshot of the A-CEEI User Interface



Notes. Figure 1 is a screenshot of the top of the user interface for preference reporting. Of the nine course sections that are visible, the hypothetical subject has reported positive values for the first eight. To make adjustments, subjects clicked two checkboxes in the far right column of the interface and were prompted to enter the adjustment in a dialog box. Any previously entered adjustments were listed at the top of the interface. The hypothetical subject has made one adjustment of  $-91$ , which tells the mechanism that getting the two accounting classes (i.e., the first two courses visible) together in the subject’s schedule together is worth zero, effectively reporting that the subject wants one or the other but not both accounting courses.

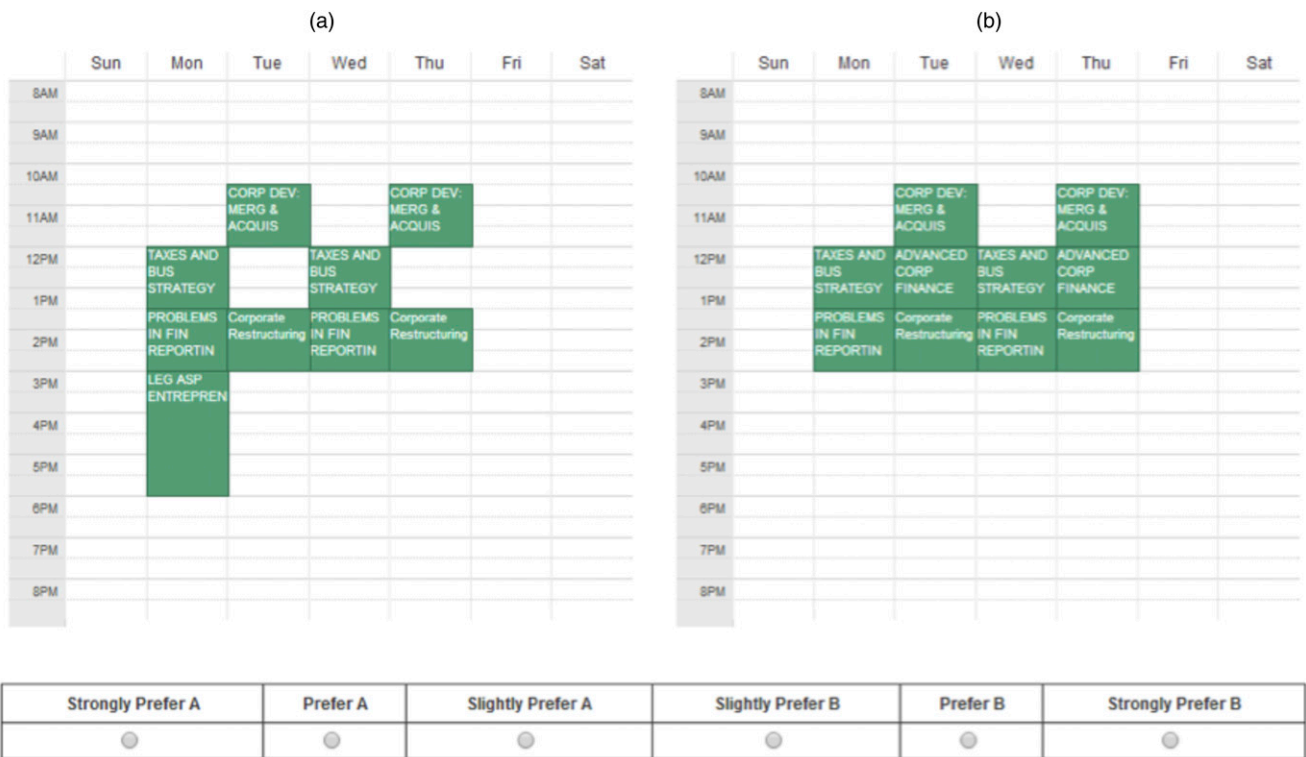
We designed the set of binary comparisons to yield data to test whether agents were able to report preferences accurately enough to realize the efficiency and fairness benefits of A-CEEI relative to the BPA as well as to provide data to directly test agents’ preference-reporting accuracy.

### 2.5.1. Efficiency and Fairness.

**2.5.1.1. Efficiency.** Subjects’ first and last binary comparisons were between the schedule that the subject received under A-CEEI and the schedule received under the BPA. This comparison was asked twice, as the first question and the last question, with the order of the schedules reversed.<sup>25</sup> These binary comparisons

yield a simple social welfare comparison between the two mechanisms. Specifically, if more subjects prefer their A-CEEI schedule to their BPA schedule than vice versa, with similar strength of preference, this suggests that a social planner deciding between the two mechanisms should prefer A-CEEI as should a student choosing between the two mechanisms from behind a veil of ignorance. Note that this comparison can be made at the individual subject level, treating each subject as an independent observation for statistical tests, and also at the market session level, aggregating up preferences to ask which of the two mechanisms generates more social welfare at the session level.<sup>26</sup>

**Figure 2.** (Color online) Screenshot of a Binary Comparison Question



Notes. Figure 2 is a screenshot of a binary comparison. It shows two schedules and asks the subject to pick which of the two the subject prefers.

**2.5.1.2. Fairness.** To measure fairness, each subject completed up to six binary comparisons per mechanism that directly assessed whether the subject envied another subject’s schedule. Envy occurs when an individual prefers someone else’s schedule to the individual’s own schedule; envy-freeness is one of the oldest and most well-established criteria of outcome fairness in economics (Foley 1967, Moulin 1995). To increase the chance of detecting envy, each subject was only shown schedules from the set of others’ schedules that generated at least 50% of the utility of the subject’s own A-CEEI schedule based on the preferences that the subject reported under A-CEEI. Restricting to this set aimed to ensure that subjects would face at least somewhat desirable alternative schedules when answering these binary comparisons. If more than six schedules of other subjects were in this set, six schedules from this set were chosen randomly by the computer to be used in binary comparisons. If six or fewer schedules were in this set, all schedules in the set were used in binary comparisons. This design choice makes the implicit assumption that schedules generating less than 50% of the utility of the subject’s own A-CEEI schedule will not be envied, an assumption that we are able to evaluate ex post (see Online Appendix G).<sup>27</sup>

We use these binary comparisons and the definition of envy-freeness to ask whether subjects experienced

more envy under one mechanism than another. Similar to the analysis for efficiency, we use these binary comparisons to generate a test of fairness at the individual level (i.e., did a subject experience more envy under one mechanism than the other) and a test of fairness at the session level (i.e., did subjects in a market experience more envy under one mechanism than the other).

**2.5.1.3. Remark: Joint Tests.** We emphasize that these binary comparison measures of efficiency and fairness are necessarily joint tests of preference reporting and the mechanisms. That is, these comparisons answer the following question: is preference reporting accurate enough that A-CEEI is able to outperform the BPA on measures of efficiency and fairness? In addition, by comparing efficiency and fairness outcomes based on binary comparisons to the corresponding outcomes if we were to assume reported preferences were accurate, we can assess the extent to which imperfect preference reporting harmed mechanism performance.

**2.5.2. Preference Reporting.** All binary comparisons are tests of the A-CEEI preference-reporting language because we can assess whether the subjects’ preference reports accurately predict their true preference as elicited by the binary comparison between the

two schedules. In addition to the binary comparisons described above, we included five binary comparisons that were aimed specifically at preference-reporting accuracy, which compared the schedule that the subject realized under A-CEEI to the schedule the subject would have received (if distinct) if the subject's budget had been 10% or 30% higher or 10% or 30% lower than it actually was. These binary comparisons provide local tests of preference-reporting accuracy, examining schedules similar to the one the subject received. We investigate why subjects may have had difficulty reporting preferences in Section 4.

## 2.6. Discussion: Incentives

Before we present the results, we want to return to the issue discussed in the introduction: that decisions in our experiment are not incentivized. As described in detail in the introduction, we could not use the induced preferences methodology because that would not allow us to test our fundamental research question of whether market participants could report their real preferences accurately enough to realize the theoretical benefits of A-CEEI. As noted in endnote 12, we were not able to incentivize choices in our experiment because doing so would have required giving subjects some positive probability of receiving—for a real upcoming Wharton spring semester—each of the schedules they constructed in the mechanisms and selected in the binary comparisons. The typical response when researchers are unable to offer desired incentives in a laboratory experiment is to attempt to run a field experiment. In the field, both real market participants' real preferences and incentives for their choices are usually already in place. Such an experiment might have randomly assigned students to use different course allocation mechanisms (e.g., assigned some to use A-CEEI and others to use the Wharton bidding points auction, each for a subset of the available spring semester seats).<sup>28</sup> Given the nature of the problem, however, running a field experiment was just as infeasible as providing incentives for our laboratory study.<sup>29</sup>

We, therefore, faced a design challenge. Although we were able to bring real market participants' real preferences into a controlled laboratory environment, we were not able to incentivize their decisions, and we needed to understand and mitigate any potential risk of the absence of incentives.<sup>30</sup>

The main risk is that subjects might not exert as much effort in an unincentivized experiment as they would in an incentivized one. We, thus, took care to design the experiment so that such lack of effort, if present in our setting, would bias against finding that agents could report their preferences accurately enough for A-CEEI to realize the benefits promised by the theory.

Imagine there are two kinds of experimental subjects: "triers" and "non-triers." Triers exert the same level of effort in the experimental tasks as they would if fully incentivized, and non-triers exert zero effort in the mechanisms, and their binary comparison responses are pure noise, that is, 50/50 coin flips. This noise from the non-triers biases toward less accurate preference reporting under A-CEEI and less ability to detect a difference in efficiency or fairness between A-CEEI and the BPA. This pushes against finding that subjects can report preferences accurately enough for A-CEEI to outperform the BPA: noise from the non-triers biases our results away from finding that subjects can report their preferences accurately and biases our results away from finding that A-CEEI improves efficiency and fairness relative to the BPA.

A subtler case is if the lack of incentives causes subjects to exert effort that is intermediate between full effort and pure noise. To understand what would happen in this case would require an understanding of the function mapping the level of effort to how well subjects perform in the experimental mechanisms and how accurately they reply to binary comparisons. We, of course, do not know this function, but given that the BPA is familiar to subjects and A-CEEI is unfamiliar, we might expect partial effort to harm A-CEEI more than the BPA, which also pushes against finding that A-CEEI outperforms the BPA.

A second potential risk that is distinct from low effort and that would bias some results in our hypothesized direction is that students in the laboratory disliked the Wharton bidding points auction in practice and, thus, attempted to sabotage its performance in the laboratory. Although we cannot rule out this possibility entirely (nor could we even if the experiment were incentivized), a few things give us comfort. First, the subjects in the experiment were representative of the Wharton student body as a whole, on both demographic measures and, crucially, their perception of the Wharton Auction's effectiveness (see Table A1 in Online Appendix B).<sup>31</sup> Second, subjects were recruited to the experimental sessions by an email that came from the Wharton administration that did not mention course allocation, and subjects were explicitly asked in the experimental instructions to take their decisions seriously in the laboratory just like they do in real life. Our impression, given the attentiveness of the subjects and the questions they asked during the sessions, is that the Wharton students in the laboratory took this direction seriously.

## 3. Results on Fairness and Efficiency

In this section, we present summary statistics on the use of the preference-reporting language and provide initial evidence on subjects' preference-reporting

ability in Section 3.1. We then explore whether subjects reported preferences accurately enough to achieve the fairness and equity benefits of A-CEEI in Sections 3.2–3.4.

### 3.1. Preference-Reporting Language Use and Accuracy

Panel A of Table 1 shows summary statistics on how subjects used the preference-reporting language. The first four rows show that subjects assigned positive cardinal utilities to about half of the 25 available courses on average and that they used the cardinal utility range provided by the reporting language (e.g., about half of courses to which they assigned positive values had  $0 < v < 50$  and about half had  $50 \leq v \leq 100$ , where  $v$  indicates cardinal utility level). The last three rows suggest that most subjects chose not to use any adjustments (e.g., the median subject used zero adjustments, and the average number of adjustments across all subjects was slightly more than one).

Panel B of Table 1 provides initial evidence about preference-reporting accuracy. In particular, we take the binary comparison as reflecting the subject’s true preference and ask whether the subject’s reported preferences to A-CEEI correctly ranked the schedules from the binary comparison. If so, we say the preference reports were *consistent* for that binary comparison; otherwise, we say the preference reports generated a *contradiction*. The first observation about these data

is that subjects are usually consistent. As shown in the first row of Panel B, 84.41% of binary comparisons were correctly ranked by reported preferences.<sup>32</sup>

In addition, preference reports are more often consistent when a subject’s binary comparison indicates a strong preference. For example, preference reports are consistent 92.11% of the time when a subject’s binary comparison indicates that the subject “Strongly Prefers” one of the schedules.

Because the preference reports generate a cardinal utility measure for each schedule, we can also ask about the “size” of contradictions. That is, we can calculate the cardinal utility assigned to each schedule in a binary comparison and ask whether the preference reports were “close” to correctly ranking the two schedules when they failed to do so. The smaller the utility difference is, the closer the preference reports were to correctly ordering the schedules (or, put differently, the smaller the changes to preference reports would need to be to eliminate the contradiction). The third row of Panel B reports that, conditional on a contradiction, the median utility difference between schedules is 38 for all comparisons and 30.5 for “Strongly Prefer” comparisons, which is roughly a third of the value of a single highly preferred course or about 10% of the average value of an A-CEEI assigned schedule. Online Appendix Figure A1 shows histograms of the binary comparisons by the utility difference between the schedules. Negative utility

**Table 1.** Use of the Preference-Reporting Language and Contradictions

Panel A: Use of preference-reporting language ( $n = 132$ )						
	Mean	Minimum	25th percentile	Median	75th percentile	Maximum
Number of courses valued $v > 0$	12.45	7	11	12	14	24
Number of courses valued $v = 100$	1.40	0	1	1	1	8
Number of courses valued $50 \leq v \leq 99$	4.87	0	3	5	7	10
Number of courses valued $0 < v < 50$	6.17	0	4	6	8	17
Number of adjustments	1.08	0	0	0	2	10
Number of adjustments $> 0$ (complements)	0.55	0	0	0	1	10
Number of adjustments $< 0$ (substitutes)	0.53	0	0	0	1	6

Panel B: Preference-Reporting consistency			
	All comparisons ( $n = 1, 661$ )	“Prefer” or “Strongly Prefer” ( $n = 1, 400$ )	“Strongly Prefer” ( $n = 735$ )
Consistent comparisons, %	84.41	87.64	92.11
Contradictions, %	15.59	12.36	7.89
If contradiction, median utility difference	–38.00	–35.00	–30.50
If contradiction, median percentage utility difference	–13.35	–12.77	–11.31

*Notes.* Panel A reports on the use of the preference-reporting language for the 132 subjects in the experiment;  $v$  is the cardinal value assigned to a particular course section. Panel B reports summary statistics on the rate at which preference reports are consistent, defined as the ordinal ranking implied by the subject’s preference report correctly predicting the preference as reported in a binary comparison. “If contradiction, median utility difference” reports the median utility difference of the schedule preferred by the binary comparison minus the schedule that the reported preferences ranked higher. “If contradiction, median percentage utility difference” reports the median of the utility difference divided by the utility of the schedule that the reported preferences ranked higher. These data cover the 126 subjects for whom we collected binary choice data (see endnote 37 about the other six subjects).

differences reflect the preference reports failing to predict the preferred schedule (i.e., contradictions). These histograms underscore that “large” contradictions are exceedingly rare. For example, contradictions with a utility difference of more than 100 utils constituted just 2.41% of all binary comparisons and just 1.36% of “Strongly Prefer” binary comparisons.

Taken together, these results suggest that, although there were indeed preference-reporting errors, preference reports to A-CEEI rather accurately reflected subject preferences as elicited by the binary comparisons. The preference reports were consistent with the elicited preferences 84.41% of the time overall and were consistent with the elicited preferences 92.11% of the time when the subject had a strong preference. In addition, when preference reports generated a contradiction, the

utility differences between the schedules were usually quite small. These findings suggest hope that agents are able to report their preferences accurately enough to reap the efficiency and fairness benefits of A-CEEI—the question we turn to next.

### 3.2. Comparing A-CEEI and the BPA

Our main results comparing A-CEEI to the BPA appear in Table 2, which presents results of our efficiency tests (top panel) and fairness tests (bottom panel). The table presents results at the individual subject level (left column) and the market session level (right column).

We provide our main tests of whether subjects can report their preferences accurately enough for A-CEEI to outperform the BPA using binary comparison data (first row of each panel) and give an

**Table 2.** Efficiency and Fairness

Outcome	Data	Aggregation Level	
		Individual Subject	Market Session
Efficiency	Binary Comparison	(A) 56 - Prefer A-CEEI 42 - Prefer Auction 17 - Identical outcomes 17 - Indeterminate preference $p = 0.094$	(B) 6 - Prefer A-CEEI 0 - Prefer Auction 2 - Ties $p = 0.016$
	Reported Preference	(C) 79 - Prefer A-CEEI 35 - Prefer Auction 17 - Identical outcomes 1 - Indeterminate preference $p < 0.001$	(D) 7 - Prefer A-CEEI 0 - Prefer Auction 1 - Tie $p = 0.008$
	Both	Test that Binary Comparison and Reported Preference Classifications are the same: $p < 0.001$ $p = 0.500$	
Fairness	Binary Comparison	(E) 40 - Less Envy A-CEEI 23 - Less Envy Auction 65 - No Envy either 4 - Same Envy both $p = 0.022$	(F) 5 - Less Envy A-CEEI 1 - Less Envy Auction 2 - Tie $p = 0.109$
	Reported Preference	(G) 35 - Less Envy A-CEEI 4 - Less Envy Auction 93 - No Envy either 0 - Same Envy both $p < 0.001$	(H) 8 - Less Envy A-CEEI 0 - Less Envy Auction 0 - Tie $p = 0.004$
	Both	Test that Binary Comparison and Reported Preference Classifications are the same: $p = 0.072$ $p = 0.125$	

Notes. See definitions for the labels listed in the table in the sections of the main text corresponding to each cell (A)–(H). For Efficiency (top panel), we test whether agents are more likely to prefer their A-CEEI schedule to their BPA schedule. For Fairness (bottom panel), we test whether subjects experience less envy in A-CEEI than in the BPA. Individual subject preferences were aggregated to a preference at the market session level using a majority rule social welfare criterion. *P*-values reported in cells (A)–(H) are one-sided sign tests. *P*-values reported in the “Both” rows are matched-pair sign tests that compare a subject’s (or session’s) classification based on binary comparisons to that subject’s (or session’s) classification based on reported preferences with the null hypothesis that the median of these differences is equal to zero.

indication of the extent to which imperfect preference reporting harmed mechanism performance by showing the same tests using reported preference data (second row of each panel). The difference in these tests gives a sense of magnitudes for the harm caused by preference-reporting mistakes (a statistical test of this difference is in the bottom row of each panel). Just as the tests using the binary comparisons generate a lower bound on the benefits of A-CEEI (i.e., because of the lack of incentives and the limited time to report preferences), the tests using reported preference data provide an upper bound on the performance benefits of A-CEEI relative to the BPA if preference reporting could be made more accurate, for example, through education and training of students or by giving students more time to think about and report their preferences than was possible in the laboratory.

We discuss the results from the top panel on efficiency in Section 3.3 and the bottom panel on fairness in Section 3.4. To complement the results presented in Table 2, we present robustness tests of our binary comparison results that utilize the rich nature of the binary comparison data (e.g., including the intensity of preference) and otherwise redefine our outcome variables in Online Appendix Table A2. As discussed throughout Sections 3.3 and 3.4, these robustness tests show that our results are similar under different definitions of our key outcome variables.

We make two remarks regarding methodology. First, we believe it is appropriate to use one-sided statistical tests for the analyses in this section. In the tests based on reported preferences (cells (C), (D), (G), and (H)), we are testing directional predictions based on the theoretical efficiency and fairness benefits of A-CEEI and the theoretical efficiency and fairness problems of the BPA (Sönmez and Ünver 2010). In the tests based on binary comparisons (cells (A), (B), (E), and (F)), a one-sided test is appropriate given the nature of our research question. If we reject the null, we conclude that subjects were indeed able to report their preferences accurately enough to realize the theoretical efficiency and fairness benefits of A-CEEI. If we fail to reject the null, we conclude that subjects had sufficient difficulty with preference reporting that the theoretical benefits failed to manifest.<sup>33</sup> That said, we recognize that some readers may prefer two-sided tests; two-sided tests would double all  $p$ -values in the table and, in particular, would cause the individual subject binary comparison result to go from marginally significant at the 10% level to insignificant.

Second, although we report statistical tests separately for each of the eight cells in the matrix, we consider the gestalt of the results as more informative than any individual test. More specifically, we take comfort that all of the binary comparison results are in

the same direction and that the binary comparison and reported preference results are all consistent with the conclusion that subjects reported accurately enough to realize the theoretical benefits of A-CEEI but that imperfect preference reporting harmed mechanism performance.

### 3.3. Efficiency Tests

**3.3.1. Binary Comparison, Individual Subject (Table 2, Cell (A)).** As described in Section 2.5, our binary comparisons on efficiency provide a measure of social welfare by asking subjects which of the two mechanisms they prefer based on their realized schedules. In particular, we asked subjects who received different schedules from the two mechanisms whether they preferred the schedule they received under A-CEEI or under the BPA. This question was asked twice, once as the first binary comparison and once as the last binary comparison with the order of the schedules reversed between the two.

Consequently, individual subjects can fall into one of four mutually exclusive groups based on their binary comparison data. Subjects can either prefer their A-CEEI schedule in both binary comparisons (which we label “Prefer A-CEEI”), prefer their BPA schedule in both binary comparisons (“Prefer BPA”), not display a consistent preference between the two schedules they received (“Indeterminate preference”), or receive the same schedule from both mechanisms (“Identical outcome”).<sup>34</sup>

As reported in cell (A) of Table 2, 56 subjects Prefer A-CEEI, 42 subjects Prefer BPA, 17 subjects have an Indeterminate preference, and 17 subjects receive Identical outcomes. To test whether A-CEEI outperforms the BPA, we treat each subject as an independent observation, assign subjects with an Indeterminate preference or Identical outcomes as having no preference between the mechanisms, and perform a one-sided sign test.<sup>35</sup> The test yields  $p = 0.094$ . This result suggests that subjects are able to report their preferences accurately enough for A-CEEI to outperform the BPA on this efficiency measure though only at the 10% significance level.

As shown in Online Appendix Table A2, cells (I) and (K), results look similar under robustness specifications that make our definition of preference stricter or more inclusive. In cell (I), subjects are only classified as Prefer A-CEEI if they state that they prefer or Strongly Prefer their A-CEEI schedule to their BPA schedule in both binary comparisons (and likewise for the BPA). Under this stricter definition, fewer subjects are classified as having a preference, but A-CEEI is still preferred to the BPA, at least marginally statistically significantly (see cell (I),  $p = 0.057$ ). In cell (K), subjects are classified as having a preference based on the average intensity of subjects’ preferences across the

two binary comparisons, which allows us to assign a preference to eight additional subjects who previously were classified as having an Indeterminate preference.<sup>36</sup> This yields an overall count of 59 subjects preferring A-CEEI and 47 preferring the BPA (see cell (K),  $p = 0.143$ ).

**3.3.2. Binary Comparison, Market Session (Table 2, Cell (B)).** To conduct our session-level tests, we aggregate these individual preferences up to the session level based on a majority-rule social welfare criterion. We count the number of Prefer A-CEEI and Prefer BPA in each session. If there are more of the former, we classify the session as “Prefer A-CEEI”; if there are more of the latter, we classify the session as “Prefer BPA,” and if there are an equal number, we classify the session as a “Tie.”

As reported in cell (B) of Table 2: six sessions Prefer A-CEEI, zero sessions Prefer BPA, and two sessions are a Tie. To test whether A-CEEI outperforms the BPA, we treat each session as an independent observation and perform a one-sided sign test. The test yields  $p = 0.016$ . Looking at the market session level reaffirms the individual subject level results and indicates that agents are able to report their preferences accurately enough for A-CEEI to outperform the BPA on this efficiency measure.

As shown in Online Appendix Table A2, cells (J) and (L), we get similar market session level results from our other definitions of preference, albeit with slightly less statistical confidence (both cells (J) and (L),  $p = 0.109$ ).

**3.3.3. Reported Preference, Individual Subject (Table 2, Cell (C)).** The second row of Table 2 runs the same tests as the row above but uses reported preference data rather than binary comparison data. Notice that we still have the same four classifications as when analyzing the binary comparison data in cell (A), but definitions have changed slightly because preferences are based on reported preference data. Subjects’ preference reports may imply they receive higher utility from their A-CEEI schedule than their BPA schedule (which we label Prefer A-CEEI), receive higher utility from their BPA schedule than their A-CEEI schedule (Prefer BPA), or receive the same utility from different schedules from each of the two mechanisms (Indeterminate preference). If they receive the same schedule from both mechanisms, we again use the label “Identical outcome.”

As reported in cell (C) of Table 2, 79 subjects Prefer A-CEEI, 35 subjects Prefer BPA, one subject has an Indeterminate preference, and 17 subjects receive Identical outcomes. A one-sided sign test yields  $p < 0.001$ .

**3.3.4. Reported Preference, Market Session (Table 2, Cell (D)).** Applying the same majority-rule social

welfare criterion to the individual preferences based on reported preferences yields a test of whether A-CEEI outperforms the BPA at the market session level based on reported preferences. As reported in cell (D) of Table 2, seven sessions Prefer A-CEEI, zero sessions Prefer the BPA, and one session is a Tie. A one-sided sign test yields  $p = 0.008$ .

**3.3.5. Discussion.** Results from the top row of Table 2 demonstrate that subjects are able to report preferences accurately enough to realize the efficiency benefits of A-CEEI. At both the individual level ( $p = 0.094$ ) and the session level ( $p = 0.016$ ), A-CEEI schedules are preferred to BPA schedules. In addition, reported preference data suggests that, absent preference-reporting mistakes, A-CEEI would dramatically outperform the BPA.

Comparing results in the top row and the second row allows us to test whether A-CEEI outperforms the BPA to a statistically significantly greater extent in reported preferences data than in binary comparison data. For the individual subject data, we run a matched-pair sign test that compares a subject’s classification based on binary comparisons to that subject’s classification based on reported preferences with the null hypothesis that the median of these differences is equal to zero. As shown in the bottom row of the efficiency panel, this test yields  $p < 0.001$ . This suggests that, although subjects report their preferences accurately enough for A-CEEI to outperform the BPA by a slim margin using individual subject data, preference-reporting mistakes significantly harmed mechanism performance. Note that we do not see a significant difference when we run a similar matched-pair sign test on the session-level data.

## 3.4. Fairness Tests

**3.4.1. Binary Comparison, Individual Subject (Table 2, Cell (E)).** As described in Section 2.5, our binary comparisons on fairness allow us to investigate whether subjects experience less envy in A-CEEI than in the BPA. In particular, for each mechanism, each subject was asked to compare the subject’s realized schedule from that mechanism with (up to six) desirable schedules that other subjects in the subject’s session received from that mechanism. This generates a measure of how many schedules each subject envies in each mechanism.

Consequently, individual subjects can again be classified into one of four mutually exclusive groups based on their binary comparison data. Subjects can either experience less envy under A-CEEI than the BPA (which we label “Less Envy A-CEEI”), experience less envy under the BPA than A-CEEI (“Less Envy BPA”), experience no envy under either mechanism (“No Envy either”), or experience the same amount of envy

(i.e., envy the same positive number of others' schedules) in both mechanisms ("Same Envy both").

As reported in cell (E) of Table 2, 40 subjects are classified as Less Envy A-CEEI, 23 subjects are Less Envy BPA, 65 subjects are No Envy either, and four subjects are Same Envy both.<sup>37</sup> To test whether A-CEEI outperforms the BPA, we treat each subject as an independent observation and perform a one-sided sign test. The test yields  $p = 0.022$ , demonstrating that subjects experience less envy under A-CEEI than under the BPA.

As shown in Online Appendix Table A2, cells (M), (O), and (Q), results are robust to three different measures of envy. Our first approach uses a stricter definition of preference, treating subjects as envying another subject's schedule only if they Prefer or Strongly Prefer the other subject's schedule to their own. Fewer subjects are classified as Less Envy A-CEEI and Less Envy BPA, but our results remain strong (cell (M),  $p = 0.001$ ). Our second approach considers envy-freeness as a 0-1 criterion and asks whether the subject was envy-free in one mechanism but not the other. We now classify subjects as having Less Envy A-CEEI if they do not experience any envy in A-CEEI but do experience envy in the BPA and as Less Envy BPA if they do experience envy under A-CEEI but do not experience any envy under the BPA.<sup>38</sup> Subjects are still less likely to experience envy under A-CEEI than the BPA (cell (O),  $p = 0.030$ ). Our third approach combines the two previous approaches, using the binary measure of envy-freeness but using the stricter envy definition of Prefer or Strongly Prefer. Our results remain significant (cell (Q),  $p = 0.005$ ).

**3.4.2. Binary Comparison, Market Session (Table 2, Cell (F)).** As earlier, we compute our session-level results by aggregating up the individual classification as described with regard to cell (E) to the session level. We count the number of Less Envy A-CEEI and Less Envy BPA in each session. If there are more of the former, we classify the session as "Less Envy A-CEEI"; if there are more of the latter, we classify the session as "Less Envy BPA," and if there are an equal number, we classify the session as a "Tie."

As reported in cell (F) of Table 2, five sessions are Less Envy A-CEEI, one is Less Envy BPA, and two are a Tie. To test whether A-CEEI outperforms the BPA, we treat each session as an independent observation and perform a one-sided sign test. The test yields  $p = 0.109$ . Consequently, although we find statistically significant results with regard to fairness at the individual level, we have only directional evidence in support of A-CEEI outperforming the BPA at the session level.

As shown in Online Appendix Table A2, cells (N), (P), and (R), we get similar results at the market session

level from our other definitions of envy with results significant for our stricter definition of envy (cell (N),  $p = 0.016$ ) and directional when considering envy-freeness or envy-freeness with our stricter definition of envy (both cells (P) and (R),  $p = 0.227$ ).

**3.4.3. Reported Preference, Individual Subject (Table 2, Cell (G)).** Again, the second row of the fairness panel runs the same tests as the row above but uses reported preference data rather than binary comparison data. We focus on the same subjects and the same comparison schedules but measure envy based on whether subjects' reported preferences suggest they get more utility from another subject's schedule than their own schedule. We then generate the same four classifications as when analyzing the binary comparison data.

As reported in cell (G) of Table 2, 35 subjects are Less Envy A-CEEI, four subjects are Less Envy BPA, 93 subjects are No Envy either, and zero subjects are Same Envy both. A one-sided sign test yields  $p < 0.001$ .

**3.4.4. Reported Preference, Market Session (Table 2, Cell (H)).** As earlier, we classify sessions based on the number of subjects in each session with the individual classifications in cell (G). As reported in cell (H), this exercise finds that all eight of the sessions are classified as Less Envy A-CEEI. A one-sided sign test yields  $p = 0.004$ .

### 3.5. Discussion

Results from the binary choice data show that subjects are able to report the preferences accurately enough to realize the fairness benefits of A-CEEI. At the individual level, we find that subjects are less likely to experience envy under A-CEEI than under the BPA ( $p = 0.022$ ). At the session level, the pattern of results is directionally consistent but not statistically significant ( $p = 0.109$ ). Again, as expected, A-CEEI dramatically outperforms the BPA when abstracting away from preference-reporting mistakes.

Comparing results in the first and second rows of the panel allows us to test whether A-CEEI outperforms the BPA to a statistically significantly greater extent in reported preferences data than in binary comparison data. Again, we run a matched-pair sign test that compares a subject's classification under binary comparison data to that subject's classification under the reported preference data. As reported in the bottom row of the fairness panel, this test yields  $p = 0.072$  for the individual subject level. This suggests that, although subjects report their preferences accurately enough for A-CEEI to outperform the BPA, preference-reporting mistakes marginally statistically significantly harmed mechanism performance. We find a similar, directional result at the market session level (one-sided sign test,  $p = 0.125$ ).

## 4. Difficulty with Preference Reporting

The efficiency and fairness results in Section 3 show that difficulty with preference reporting meaningfully harmed mechanism performance. Although A-CEEI outperformed the BPA in our efficiency and fairness tests based on the binary comparisons data, A-CEEI outperformed the BPA to a greater extent in our measures based on the reported preference data, which assume that preference reporting is perfect. In this section, we aim to understand the causes of preference-reporting difficulty and to identify ways that preference-reporting accuracy might be improved.

Conceptually, we distinguish between two possible reasons why subjects' preference reports might not reflect their underlying true preferences. First, subjects may have had difficulty using the preference-reporting language we provided in the laboratory to express their underlying true preferences even though, in principle, it was mathematically feasible for them to do so with the language. We evaluate this concern in Section 4.1. Second, there are some kinds of preferences that mathematically cannot be expressed using the language we provided. If such preferences were present in our subject pool, this would necessarily create a discrepancy between subjects' reported preferences and their true preferences. We evaluate this concern in Section 4.2.<sup>39</sup> Section 4.3 discusses the results from this section.

### 4.1. Difficulty Using the Preference-Reporting Language

To assess whether agents had difficulty using the preference-reporting language we provided, we first explore whether they were able to effectively use each of its components: cardinal values to express preferences for individual courses and pairwise adjustments to express certain kinds of complementarities and substitutabilities for pairs of courses. We explore subjects' ability to use each of these components of the language in turn.

To examine subjects' ability to report cardinal item values, we differentiate between the ordinal and cardinal component of a subject's reported preferences for individual courses. For this analysis, we drop the 87 binary comparisons in which one or both schedules triggered an adjustment. For the remaining 1,574 binary comparisons, we differentiate between binary comparisons in which the reported preferences can rank the two schedules based only on the *ordinal* information in the preference reports or whether the *cardinal* information is needed as well. For example, if we rank course sections by a subject's assigned cardinal item values and find that Schedule A consists of the subject's {1st, 3rd, 5th, 7th, 9th} highest value course sections and Schedule B consists of the

subject's {2nd, 4th, 6th, 8th, 10th} highest value course sections, then we know the preference reports rank Schedule A over Schedule B based on ordinal information alone (i.e., we do not need to know the specific cardinal utilities the student assigned to each course).

If, instead, Schedule A consists of a subject's {1st, 2nd, 8th, 9th, 10th} highest value course sections and Schedule B consists of a subject's {3rd, 4th, 5th, 6th, 7th} highest value course sections, ordinal information alone is insufficient for the reported preferences to rank the schedules. Comparisons that rely on cardinal information are those for which the subject's ability to report cardinal preference information accurately is put to the test.

Table 3 reports probit regressions with a dependent variable equal to one if the preference reports contradict the binary comparison choice.<sup>40</sup> Column (1) of Table 3 shows that comparisons that rely on cardinal information are more likely to be associated with a contradiction than those that rely on ordinal information alone (the excluded group). The interpretation of the coefficient is that preference reports are 21 percentage points more likely to generate a contradiction when they rely on cardinal information than when they rely on ordinal information only. This difference is both economically large, relative to an average rate of contradiction of 15.8% among comparisons of schedules without an adjustment, and highly statistically significant, with a z-statistic of 7.00. Although part of this sizeable effect is no doubt because binary comparisons that are revealed ex post to have relied on ordinal information were likely easier for the preference reports to rank ex ante, the result also suggests that subjects had meaningful difficulty reporting cardinal utilities to the preference-reporting language.<sup>41</sup>

To examine subjects' ability to report complementarities and substitutabilities, we explore subjects' use of adjustments. Pairwise adjustments were not used as widely as one might have expected—just 1.08 per subject on average as shown in Table 1. Because of the relatively limited use of adjustments, only 87 binary comparisons involved a schedule in which an adjustment was activated. For this analysis, we compare these binary comparisons, which we say relied on *combinatorial* information, with the other 1,574 comparisons. Column (2) of Table 3 finds that preference reports are directionally but not significantly less likely to generate a contradiction when they rely on combinatorial information. Although it is hard to draw conclusions with this data, the result suggests that adjustments did not detract from preference-reporting accuracy.

Finally, ability to report preferences might be driven by some combination of cognitive ability and effort. Although we cannot directly measure these variables,

**Table 3.** Causes of Contradictions

	Dependent variable: <i>Contradiction</i>			
	(1)	(2)	(3)	(4)
Cardinal (369 comparisons)	0.208 (0.033)***			
Combinatorial (87 comparisons)		−0.043 (0.036)		
High GPA (835 comparisons)			−0.041 (0.024)*	
Lower utility schedule has “elegant” feature (241 comparisons)				0.060 (0.037)*
Predicted probability at mean values	0.158	0.156	0.156	0.156
Observations	1,574	1,661	1,661	1,661
Clusters (subjects)	122	126	126	126
R <sup>2</sup>	0.059	0.001	0.004	0.004

*Notes.* Probit regressions have a dependent variable equal to one if the preference reports generate a contradiction and equal to zero if the preference reports are consistent. Marginal effects are reported. We analyze the 126 subjects for whom we have binary comparison data and exclude comparisons in which both schedules generate equal cardinal utility (such that the reported preference data does not generate a strict preference between schedules). Standard errors are robust and clustered at the subject level. Significance (of two-sided tests) is denoted with stars.

\*\*\* $p < 0.01$ , \*\* $p < 0.05$ , \* $p < 0.1$ .

we have a proxy for them in whether students have a high or low grade point average. We define *High GPA* as being above the median grade point average among our subjects. Column (3) of Table 3 reports that preferences reported by subjects with higher grade point averages are marginally statistically significantly less likely to generate contradictions.<sup>42</sup>

#### 4.2. Limitations of the Preference-Reporting Language

The preference-reporting language we used in the experiment was not fully expressive (as defined, e.g., in Nisan 2006), meaning that there exist ordinal preferences over schedules that subjects would be mathematically unable to express using the language that was provided. The issue is that many kinds of nonadditive preferences cannot be expressed using pairwise adjustments.<sup>43</sup> Additionally, there are many kinds of nonadditive preferences that, in principle, could be expressed using the language but for which the language does not seem especially natural.<sup>44</sup>

The set of potential nonexpressible preferences is vast, and we do not have a disciplined way of exploring all such possibilities as a source of preference-reporting contradictions.<sup>45</sup> Instead, we explored two specific sources of nonadditive preferences that the Wharton committee suggested to us would be the most important, both of which arise from scheduling considerations per se rather than the contents of the classes within the schedule.

The first is whether a student’s schedule is *balanced*: at least one class on each day Monday through Thursday (none of the course sections in our experiment met on

Friday as is typical at Wharton). The second is whether the schedule is *contiguous*: every day on which a student has class the student has at most one 1.5-hour gap between the start of the first class and the end of that last one. According to the Wharton committee, these characteristics make a schedule “elegant” and are highly valued by at least some students. However, subjects are not able to express a value for either characteristic using the preference-reporting language in the experiment. We, therefore, investigate whether the preference reports generated a higher probability of a contradiction when the schedule in a binary comparison that receives a lower utility based on reported preferences is elegant in at least one of these two ways (and so may generate utility that the subject was unable to report using the preference-reporting language) and the other schedule in the binary comparison is not elegant in that way.

Column (4) of Table 3 shows that comparisons in which the schedule with the lower utility is elegant in a way that the schedule with the higher utility is not are marginally statistically significantly more likely to yield a contradiction ( $z = 1.78$ ). That preference reports are more likely to make a contradiction when they give a lower value to an elegant schedule than to a schedule that is not elegant in that way suggests that at least some of the contradictions are a result of the preference-reporting language failing to provide a way for agents to report important features of their preferences. A caveat is that each of these types of nonexpressible preferences (i.e., being balanced and being contiguous) account for only a small number of contradictions each.<sup>46</sup> That said, there are likely many

other nonexpressible preferences that we do not quantify here.

### 4.3. Discussion

Results from this section provide some evidence as to the sources of inaccurate preference reporting. First, subjects had particular difficulty with reporting cardinal preference intensity information: preference reports were meaningfully more likely to generate a contradiction when they relied on cardinal information to determine which schedule was preferred. Second, when subjects expressed nonadditive preferences, they did so with reasonable accuracy, but they did so rarely. Third, subjects with higher grades reported preferences that were less likely to generate contradictions, suggesting that higher-performing students were better able to report preferences. Fourth, evidence suggests that there were some nonadditive preferences that were important to subjects but that subjects were unable to express with the language provided.

These results provide empirical support for some common intuitions in the market design literature, such as the ease of reporting ordinal information relative to cardinal information (Bogomolnaia and Moulin 2001), the importance of nonadditive preferences (Cantillon and Pesendorfer 2007, Reguant 2014), and the overall importance of language design (Milgrom 2009, 2011). We also hope that the overall logic of the results in this section gives the reader additional comfort as to the validity of the experimental methodology.

It is worth noting that our results on preference reporting also guided practical implementation at Wharton in a few ways. First, Wharton opted to use the same language in practical implementation as was used in the laboratory based on the overall level of accuracy of the reports and taking into consideration that subjects had only 10 minutes to report their preferences and had only minimal training. Second, Wharton provided students with extensive training on how to use the reporting language with significant training focused specifically on how to think about cardinal preference intensity because this was such an important source of difficulty in the laboratory. Third, Wharton enhanced the top-10 widget in the preference reporting user interface to allow students to see substantially more than 10 schedules, allowing students to assess whether they had reported their preferences accurately not just for their very most preferred schedules (which may be unattainable if the student likes mostly popular courses) but further down their overall ranking as well.<sup>47</sup> To date, Wharton has opted not to incorporate other ways to report nonadditive preferences beyond the pairwise adjustment tool, fearing excessive complexity.

Developing a conceptual understanding of the trade-off between expressiveness and complexity is an interesting open area for future research.

## 5. Analysis of Survey Data

As noted in the introduction, an additional advantage of using real market participants as experimental subjects and asking them to use a mechanism with their real preferences is that we could search for side effects: issues not captured by the theory that could undermine the potential benefits of a new market design. For example, a mechanism might have attractive theoretical fairness properties, but market participants might, nevertheless, subjectively find it to be unfair. A mechanism might have attractive incentive properties, but participants might not understand that they should report truthfully (see Li 2017, Rees-Jones 2018, Rees-Jones and Skowronek 2018, Hassidim et al. 2021). Market participants might find a mechanism to be frustrating or confusing, properties that would undermine the practical appeal of a mechanism but that seem difficult to capture in a theoretical model.

Concern about side effects was especially pronounced in our setting because of the nature of both the mechanism being considered and the allocation problem. Regarding the mechanism, A-CEEI had never been used before, so it lacked reassuring precedent, and it is complex in several ways that intuitively raise concerns about side effects. Regarding the setting, fear that a new market design might lead to unexpectedly dissatisfied market participants was high at Wharton, where student satisfaction is a top priority; the Wharton committee was concerned about student satisfaction with regard to both the final allocation and the process that led to that allocation.

To address these concerns, we collected a wide variety of survey data to search for issues missed by the theory. After subjects used each mechanism, we asked them a number of survey questions about that mechanism (see questions and results in Online Appendix Table A8). In this section, we highlight the main results from this search for side effects. A full discussion of our survey results can be found in Online Appendix J.

One set of survey questions covered “liking” of the mechanism, and we found very little difference in subjects’ answers to these questions between A-CEEI and the BPA. Our main takeaway from this set of questions was that there was not some important unmeasured side effect that caused subjects to dramatically prefer either A-CEEI or the BPA that our main efficiency and fairness analyses would have missed. These results also seemed to give comfort to the Wharton committee that there was nothing unexpected about the A-CEEI mechanism that led the Wharton student subjects to dislike the system.

Another set of questions asked about strategic play, asking for a level of agreement with the following statements: “I had to think strategically about what other students would do in this course allocation system” and “Someone with perfect knowledge of the historical supply and demand for courses could have had an advantage over me in this system.” Subjects reported much higher agreement with these two for the BPA than for A-CEEI, suggesting that they at least partially understood that A-CEEI was strategy-proof. However, some subjects still stated agreement with these statements for A-CEEI. A lesson for implementation that came out of these survey responses was to do a more thorough job of explaining A-CEEI’s strategy-proofness to students because understanding that historical information and strategizing was not necessary for A-CEEI was positively correlated with other measures of satisfaction with A-CEEI.<sup>48</sup>

A final set of questions asked about transparency, asking for a level of agreement with the following statements: “I understand how this course allocation system works” and “I felt like I had control over my schedule in this course allocation system.” Subjects reported significantly less agreement with these statements for A-CEEI than for the BPA, suggesting they felt that A-CEEI was a “black box,” that is, non-transparent. The transparency issue constitutes a side effect in that it negatively impacted market participants’ evaluation of the mechanism and was not anticipated by the theory.

Wharton acted on this finding in their practical implementation of A-CEEI in two ways. First, Wharton administrators did student-wide presentations about the new mechanism to explain in detail how it works, the theory behind it, and the experimental evidence—all in an effort to enhance transparency. Second, Wharton made a simple but important change to the mechanism’s user interface. In the user interface implemented in the laboratory, subjects were shown the schedule they received under A-CEEI but were not shown market-clearing prices. This prevented subjects from understanding why they received their specific schedule and why, for example, they failed to get some particular course they valued highly. In the practical implementation, Wharton modified the user interface so that students are shown the market-clearing prices.<sup>49</sup>

Finally, the survey data revealed a positive side effect with regard to gender. At the time of our experiment, the Wharton administration was facing evidence that women at Wharton disproportionately disliked the Wharton bidding points auction. A Wharton survey of all second-year students in the year of our experiment found that women reported lower ratings for the effectiveness of the real Wharton bidding points auction than men did (seven-point scale of effectiveness, 4.95 for men versus 4.28 for women,

*t*-test, two-sided,  $p < 0.001$ ).<sup>50</sup> The administration was, therefore, interested in whether A-CEEI would also display a gender disparity. Our survey questions about the BPA generated the same pattern as the Wharton administration had seen in their data. However, there was no gender gap in liking of or satisfaction with A-CEEI. Eliminating the gender difference that was present in attitudes toward the BPA was a positive side effect of A-CEEI not anticipated by the theory.<sup>51</sup>

## 6. Conclusion

Wharton formally decided to adopt A-CEEI for use in practice after a series of administrative meetings in the few months following our experiment. This could not have been an easy decision given the complexity of the A-CEEI mechanism and the lack of direct precedent. Based on our conversations with the committee, our sense is that what ultimately was pivotal in Wharton’s decision to adopt A-CEEI was not any one experimental result but rather the full set of experimental results: the efficiency and fairness gains relative to the BPA; the finding that preference reports were, on the whole, reasonably accurate with large mistakes comparatively rare; the finding that the efficiency and fairness gains would be meaningfully larger if preference reporting accuracy could be improved; the strategic simplicity gains identified in the survey; and the lack of any unexpected side effects beyond the transparency issue, which the committee felt could be addressed in practice with better communication and some modest changes to the user interface.

Unfortunately, it was not possible to obtain the data that would have been necessary to do a full empirical before-and-after comparison of the two mechanisms.<sup>52</sup> However, the limited data that are available are all consistent with the claims made by the theory and the experiment. One simple way to measure outcome fairness is to look at the distribution of the most popular courses; for any one student, we cannot tell if their failure to get popular courses reflects unfairness of the mechanism or their preferences, but the aggregate distribution suggests that A-CEEI improved equity. In the last fall of the bidding points auction, 32% of students got zero of the top 20 most popular courses and 5% got three or more versus 13% and 0%, respectively, under A-CEEI. That is, under A-CEEI, fewer students got none of the most popular courses and fewer (i.e., none) got three or more. Another way to measure outcome fairness is to look at the distribution of the cost of students’ final schedules. The cost of a student’s schedule is defined as the cost of buying all the classes in that student’s schedule at the prices determined by the market design (i.e., the market clearing prices determined by A-CEEI or the prices in the BPA). The Gini index of this distribution went from 0.54 in the last fall of the

bidding points auction to 0.32 in the first fall of A-CEEL.<sup>53</sup> In addition, we used school-wide surveys to investigate the change in mechanisms. At our urging, the annual administration survey of the student body added a few questions about course allocation in the last year of the bidding points auction's use, written in such a way that they could be used again in the first year of A-CEEL (which was implemented as Course Match) with minimal change to language. The percentage of students responding either Agree or Strongly Agree to the statement "I was satisfied with my schedule from (the Course Auction system/Course Match)" increased from 45% in 2013 (the last year of the bidding points auction) to 64% in 2014 (the first year of A-CEEL). The percentage responding either Agree or Strongly Agree for the statement "(The Course Auction, Course Match) allows for a fair allocation of classes" increased from 28% to 65%. The percentage of students responding either Effective or Very Effective to the question "Please rate the effectiveness of the (Course Auction, Course Match) system" increased from 24% to 53%.

An interesting open question for future research is how to design a better preference-reporting language both in this specific setting and in general. The results of the experiment show that the language used in the laboratory and adopted for implementation allowed for preference reports that were accurate enough to yield the efficiency and fairness benefits of A-CEEL, but the results do not at all suggest that the language is optimal. One specific direction to consider based on the experimental results would be to allow students to report richer kinds of nonadditive preferences. A more difficult conceptual question is how to think about the overall trade-off between a language's expressiveness and its efficacy. Too simple of a language may actually complicate the mechanism for participants, who must struggle with how to translate their real preferences into too simplistic of a language.<sup>54</sup> Too complicated of a language would also be suboptimal if participants are unable to effectively "speak" the language. How to design a language that is optimal for a specific setting is a fascinating question in need of a conceptual breakthrough.

A perhaps related question is whether and how to incorporate prior information about the structure of preference heterogeneity in the relevant population into preference reporting. Typically in market design, a mechanism does not assume anything about the agent's preferences that the agent does not explicitly report to the mechanism via the supplied language. Contrast this with, for example, common practice at e-commerce companies such as Amazon or Netflix, which interact whatever data they gather about any particular user's preferences with their prior on the structure of preferences in the population to form a

posterior of that user's type and make recommendations accordingly. That the Wharton committee was able to identify preferences (e.g., about the temporal structure of schedules as described in Section 4.2) that students had difficulty reporting suggests the potential for advancement on this front.

The induced preferences methodology has been critically important in the history of market design experiments, tracing all the way to the early double auction experiments of Chamberlin (1948) and Smith (1962). But, to study whether real market participants can report complex preference information accurately enough to realize the benefits of a new mechanism, we needed a new approach. We, therefore, developed the elicited preferences methodology, in which real market participants play based on their real preferences with tailored binary comparisons used to assess reporting accuracy and mechanism performance. We suspect that, as market design continues to grow as a field—and as computers continue to become more powerful, data more plentiful, and decision supports more sophisticated—market designs leveraging complicated preference information will become increasingly feasible and increasingly common. Other market design researchers can utilize the methodology developed here as part of the toolkit to evaluate new mechanisms, assess and refine preference-reporting languages, and ultimately help bring other useful market designs from theory to practice.

### Acknowledgments

The authors thank Gérard Cachon, without whom this study would never have been possible, and the Wharton School, in particular the Course Allocation Redesign Team and Wharton computing. The authors thank Natalia Drozdoff, Adriaan Ten Kate, Xan Vongsathorn, and Zizhe Xia for excellent research assistance. The authors thank the five anonymous referees and associate editor for outstanding feedback that greatly improved the paper. The authors also thank Mohammad Akbarpour, Eduardo Azevedo, Peter Cramton, Stefano DellaVigna, Clayton Featherstone, Alex Frankel, Emir Kamenica, Scott Kominers, Robin Lee, Stephen Leider, John List, Paul Milgrom, Joshua Mollner, Muriel Niederle, Canice Prendergast, Jesse Shapiro, Alvin Roth, and Glen Weyl as well as seminar participants at Boston University, the Stony Brook Workshop on Experimental Game Theory, ESA Santa Cruz, University of Michigan, Stanford, Wharton, NBER Market Design, Chicago, AMMA 2015, MSR Designing the Digital Economy, Boston College, Princeton, and the University of Virginia. Disclosure: the mechanism design in Budish (2011) and the computational procedure in Othman et al. (2010) are in the public domain. Wharton funded the implementation of the A-CEEL mechanism at Wharton, creating intellectual property (e.g., code) owned by Wharton. In 2017, Wharton spun out this intellectual property into a startup, Cognomos, of which Budish was made a director at its founding and has an equity stake. Wharton had no right of prior review of the present study.

## Endnotes

<sup>1</sup> On spectrum auctions, see Milgrom's (2004) and Klemperer's (2004) fittingly named books, *Putting Auction Theory to Work* and *Auctions: Theory and Practice*, as well as Cramton et al. (2006), Ausubel et al. (2006), Levin and Skrzypacz (2016), and Milgrom and Segal (2020). On matching markets, see Roth's (2015b) book as well as Roth (2002, 2008), Roth and Peranson (1999), Abdulkadiroğlu and Sönmez (2003), Abdulkadiroğlu et al. (2005a, b; 2006, 2017), and Roth et al. (2004, 2005, 2007). For recent surveys of market design that discuss various research directions beyond auctions and matching, see Kominers et al. (2017), Roth (2018), and Milgrom and Tadelis (2019).

<sup>2</sup> See Roth (2015a) for a survey of the literature on market design experiments as well as a detailed discussion of the present paper in Section 6.

<sup>3</sup> See Pápai (2001), Ehlers and Klaus (2003), Hatfield (2009), and Kojima (2009).

<sup>4</sup> See Sönmez and Ünver (2010), Krishna and Ünver (2008), and Budish and Cantillon (2012).

<sup>5</sup> See Sönmez and Ünver (2010) for a list of schools using this mechanism and a description of the (minor) design variations across institutions. See Section 2 for more details on Wharton's variant, which uses a fake-money Vickrey auction in an initial allocation round and then uses double auctions in subsequent rounds.

<sup>6</sup> In Roth's (2015a) recent survey of the literature on market design experiments, every laboratory experiment discussed, except the present paper, uses the induced preferences methodology.

<sup>7</sup> Our use of the term "elicited preferences" is inspired by computer science literature on the problem of eliciting preference information in complex allocation environments (e.g., combinatorial auctions) by using a relatively small number of carefully chosen preference queries. See Sandholm and Boutilier (2006) and Chen and Pu (2004) for surveys of this literature, and see Boutilier (2002) and Parkes (2005) for early contributions.

<sup>8</sup> Outside of market design, it is common to design laboratory experiments around participants' real preferences; famous examples include Kahneman et al. (1990) and Roth et al. (1991). Note that, in these latter settings, theory testing is possible without inducing preferences: in dictator and ultimatum games, subjects' preferences are assumed to be known a priori (favoring more money to less), and in endowment-effect experiments, the quantity of trade is sufficient to establish the effect without knowing subjects' precise values for the objects. In market design experiments, in contrast, theory testing often requires the researcher to have precise knowledge of subjects' heterogeneous preferences, which the induced preferences methodology directly produces.

<sup>9</sup> Our use of the term "side effects" is meant to analogize the FDA drug-approval process. The first step in that process is not to test the efficacy of the drug (that is the last step), but rather to ensure that the drug is not harmful to humans for some unforeseen reason.

<sup>10</sup> In many other practical market design implementations, there were close precedents that could be used to convince practitioners that the theory worked as intended in practice; these precedents lessen the concern about unintended consequences of the theory. For example, the Gale–Shapley deferred acceptance algorithm was independently discovered and implemented by the medical profession in the 1940s—about 15 years before the publication of Gale and Shapley (1962). Roth and Peranson (1999) report on the successful modification of the Gale–Shapley algorithm to accommodate married couples. When the Gale–Shapley algorithm was implemented for school choice, the economists involved in the implementation could point to the algorithm's decades of success in the medical labor market. Doctors discovered the idea of pairwise kidney exchange in the late 1990s; the economists who became involved helped to optimize what had been an ad hoc process to increase the number of potential matches.

<sup>11</sup> A similar lack of incentives arises in market design studies that utilize other types of surveys to ask about preferences and/or beliefs, such as Budish and Cantillon (2012), Kapor et al. (2020) and Rees-Jones (2018). See Bertrand and Mullainathan (2001) for a general discussion of the benefits and costs of survey data.

<sup>12</sup> Also note that the lack of incentives is not intrinsically a feature of the elicited preferences approach we propose. If we could have offered, with some probability, that students would obtain in real life the schedule they obtained in the laboratory version of the mechanism or a schedule they chose in a binary comparison, then all behavior would have been incentivized. However, we were unable to get the Wharton administration to provide such stakes in the laboratory experiment for obvious reasons.

<sup>13</sup> After Wharton elected to adopt the new mechanism in spring 2012, the work of practical implementation began in earnest. The engineering component of this work is reported in Budish et al. (2017).

<sup>14</sup> Some of the earliest examples of experiments using the induced preferences methodology include the early double auction experiments of Chamberlin (1948) and Smith (1962) and combinatorial auction experiments, such as Rassenti et al. (1982) and Goeree and Holt (2010). See Kagel et al. (2010) for an interesting twist on the methodology that uses theory and simulations to guide which induced preferences to explore.

<sup>15</sup> The email indicated that the study was voluntary but that participation was appreciated by the dean's office and, as a further inducement, offered \$250 to two randomly selected subjects per session. The email did not mention that the study was about course assignment. We wanted to attract student subjects who were generally representative of the Wharton MBA student body and to avoid attracting students who were disproportionately happy or unhappy with the current course auction. Subjects were statistically representative of the Wharton student population on every dimension except race and, importantly, were representative with regard to attitudes toward the Wharton bidding points auction (see discussion in Section 2.6 and Table A1 in Online Appendix B).

<sup>16</sup> We did not find any significant differences in the results based on which mechanism was used first. See Online Appendix F for details of this analysis.

<sup>17</sup> Although the first round of the auction closely resembles a real-money Vickrey auction, the attractive properties of the Vickrey auction do not translate to the fake-money setting. The mathematical difference is that preferences are not quasi-linear over objects and money because the money is fake and the game is finite. Intuitively, someone who bids \$10,000 in a real-money auction and loses to someone who bids \$10,001 may be disappointed, but at least, the person can put the money to some alternative use, whereas a student who bids 10,000 points in a fake-money auction and loses to someone who bids 10,001 may end up graduating with a large budget of useless Course Auction currency. As a result, unlike the Vickrey auction, the bidding points auction is not strategy-proof, and equilibrium outcomes can be highly unfair and inefficient. Note, however, that, if the game were infinitely repeated, then unspent fake-money would always have a future use, and so the quasi-linearity assumption would be valid. See Prendergast (2017a, b) for an implementation of a mechanism in this spirit in the context of allocating donated food to food banks across the United States.

<sup>18</sup> In practice, the final allocation of popular courses (i.e., courses with a positive price) is mostly determined by the outcome of the first round. This gave the Wharton committee confidence that there would not be much lost by using four rounds instead of nine. In the laboratory, too, most of the action took place in the first round.

<sup>19</sup> Budish's (2011) result that prices exist for A-CEEI that (approximately) clear the market requires that students have nonidentical budgets. See also Reny (2017) for a recent generalization of this result. The budgets can be arbitrarily close to equal but cannot be

exactly equal. The intuition is that the budget inequality helps break ties. For example, suppose students A and B both place an extremely high value on course X, which has one available seat. If A's budget is 5,000 and B's budget is 5,001, then setting the price of course X to 5,001 clears the market because B can afford it and A cannot. The auction breaks ties in the auction itself rather than in the budgets. If both A and B bid 5,000 points for course X, then the computer randomly selects one student to transact.

<sup>20</sup> See Budish (2011) for a more complete description of how A-CEEI works. See Othman et al. (2010) and Budish et al. (2017) for the computer science behind how to calculate the market-clearing prices in step (iii).

<sup>21</sup> We thought seriously about whether to caveat our instructions by more specifically explaining that A-CEEI is only approximately and not exactly strategy-proof, and therefore, there theoretically are conditions under which an agent could benefit from misreporting. For reasons outlined in detail here, we decided that the best advice we could provide subjects was to report their preferences truthfully and that dwelling on the difference between approximate and exact strategy-proofness would be confusing. At any realized prices, truthful reporting is best because it ensures the student receives the student's most-preferred affordable bundle at those prices. For it to be profitable for a student to benefit from misreporting preferences, it must be the case that the misreport advantageously influences prices while, at the same time, the misreport does not cause the student to get the wrong bundle at the influenced prices. Formally, by reporting preferences as  $u'$  instead of  $u$ , this changes prices from  $p$  to  $p'$ , and the student gets more utility from the bundle the mechanism thinks the student likes best at  $p'$  (based on the misreport  $u'$ ) than from the bundle the student likes best at  $p$  (based on the student's true preferences  $u$ ). The main reason why such misreports are hard to find, even in small markets, is that students require at most one unit of any particular course. Therefore, the "demand-reduction" strategies that are typically used to profitably manipulate prices in multi-object allocation mechanisms do not work here: if a student reduces demand for a course, this can indeed reduce the price for that course, but because reducing demand means pretending to want zero units instead of one unit, this does not do the student any good. A second reason why such misreports are likely to be hard to find is the black box nature of the approximate Kakutani fixed-point computation. Footnote 31 of the 2010 working paper version of Budish (2011) gives an example of the kinds of profitable manipulations that were found in extensive computational exploration in small markets, and they are nonintuitive. Because there is a risk to misreporting—one is no longer guaranteed one's most-preferred affordable schedule at the realized prices—and the benefits of misreporting are difficult, if not impossible, to realize, we decided the best advice we could give was to advise subjects to report truthfully.

<sup>22</sup> We recommended reporting a positive value for at least 12 course sections to ensure receipt of a complete schedule of five courses.

<sup>23</sup> If subjects could report adjustments over arbitrary sets of courses rather than just pairs of courses, then, in principle, the language would allow students to express any possible ordinal ranking over schedules, making the language expressive as defined, for example, in Nisan (2006). We explore limitations of the language in further detail in Section 4.

<sup>24</sup> Computationally, it is not necessary to ever formulate a student's complete rank order list over schedules. Instead, the question of what is a student's most preferred affordable schedule at a given price vector can be translated into a mixed-integer program. This is an important computational advantage because integer programming, though NP-hard, is speedy in practice for problems of this size. The practical implementation of A-CEEI solves billions of integer programs in the process of finding approximate market clearing prices. See Budish et al. (2017) for more details on the computational procedure.

<sup>25</sup> The schedule shown on the left in the first question was shown on the right in the last question. These binary comparisons were only asked if the schedules received under the two mechanisms were different.

<sup>26</sup> We are interested in both individual- and session-level results, and it is worth noting that there are inherent trade-offs between the two. Looking at individual-level data reflects the fact that we care about individual agents being made better off by a mechanism and gives us more data to run our statistical tests but ignores the session structure of our data. Looking at session-level data respects the fact that mechanisms are, by definition, implemented at the market level but gives us only eight sessions to run our statistical tests.

<sup>27</sup> Results in Online Appendix G show that, although this assumption is unlikely to hold perfectly, its failure to hold works against us finding that A-CEEI generates Less Envy than the BPA.

<sup>28</sup> To evaluate whether one mechanism outperforms the other, such a field experiment would presumably also need an incentivized elicitation procedure, for example, testing for envy by giving students the option to trade their realized schedule for the realized schedules of other students, with some positive probability.

<sup>29</sup> A field experiment was a nonstarter at Wharton, presumably both for logistical reasons and because of concerns about students' perceptions of fairness. The prospect of such a field experiment also raises a Catch-22 because even if the Wharton administration had considered such a field experiment, they would likely have wanted to see initial evidence that the mechanism could be successful—evidence of the kind generated by a laboratory experiment like ours.

<sup>30</sup> Although subjects' decisions were not incentivized, subjects were compensated for their time in the form of two \$250 prizes per session to randomly chosen subjects. The Wharton committee thought that two \$250 prizes per session would be more appropriate and attractive compensation than paying each student the expected value of roughly \$30. Suffice it to say that MBA students are different from the typical undergraduate subject pool.

<sup>31</sup> We used anonymous Wharton IDs to match experimental subjects to data from an administration survey conducted at the end of each school year. Our laboratory subjects rated the Wharton auction's "effectiveness" an average of 4.69 on a scale of 0 to 7, essentially identical to the overall Wharton average of 4.68.

<sup>32</sup> Although subjects are consistent for the majority of comparisons, if we instead look at data by subject, we see that 75.4% of subjects' preference reports generated at least one contradiction. This number reflects the fact that a larger fraction of our subjects exhibit preference reporting errors than subjects in prior experimental work with induced preferences. For example, Rees-Jones and Skowronek (2018) find that 23.3% of subjects with induced preferences over five objects fail to report these to a strategy-proof mechanism in the correct rank order. That our number is substantially higher may reflect the more complex data reporting demands of A-CEEI than the rank order list in Rees-Jones and Skowronek (2018).

<sup>33</sup> Note that even if the BPA were to perform much better than A-CEEI, we would *not* conclude that the BPA is a better mechanism on efficiency or fairness grounds. Rather, we would go back to the drawing board regarding the preference-reporting language.

<sup>34</sup> As shown in Figure 2, subjects were not given an option to report that they were indifferent between two schedules, and so seeming preference reversals among subjects with an Indeterminate preference may be a reflection that some subjects felt indifferent between the two schedules. It could also be an indication of subject errors or random choices. As discussed in Section 2.6, the extent to which subjects respond randomly works against us finding any differences between the mechanisms.

<sup>35</sup> We treat Prefer A-CEEI (and, later, Less Envy A-CEEI) as A-CEEI Outperforming the BPA, Prefer BPA (and, later, Less Envy BPA) as the BPA outperforming A-CEEI, and all other classifications as A-CEEI and the BPA performing equally well. The sign test assigns a positive value to an observation in which A-CEEI outperforms the BPA and a negative value to an observation in which the BPA outperforms A-CEEI. It then tests whether the median of these values is equal to zero. Note that, with data of this form, the sign test is equivalent to a binomial probability test whether our data could have come from a data-generating process in which A-CEEI outperforms the BPA and the BPA outperforms A-CEEI are equally likely to arise.

<sup>36</sup> Under this more inclusive definition, we assign subjects a preference for A-CEEI if they indicated a stronger preference when they said they preferred their A-CEEI schedule than when they said they preferred their BPA schedule (three subjects) and assign them a preference for the BPA if the opposite (five subjects).

<sup>37</sup> These 65 subjects classified as No Envy either include six subjects for whom we did not collect data on envy because of a bug in our survey code: in the first three sessions, we did not collect binary comparison data from subjects who received the same schedule under both A-CEEI and the BPA. Although this bug was unfortunate, we believe, if anything, it is likely to work against us finding Less Envy under A-CEEI than the BPA. We come to this conclusion by looking at the other 11 subjects with identical A-CEEI and BPA schedules. Among this group, nine are No Envy either and two are Less Envy A-CEEI. Consequently, if the missing six subjects were similar to these 11, their data would have made our results somewhat stronger.

<sup>38</sup> We classify fewer subjects as Less Envy A-CEEI and Less Envy BPA because now we treat anyone who experiences envy under both mechanisms as Same Envy both even if the number of schedules they envy is different across the two mechanisms.

<sup>39</sup> In addition, subjects using A-CEEI in the laboratory environment might have failed to put in sufficient effort to fully conceptualize their preferences or to reflect their preferences using the reporting language. Although we designed our experiment so that noise in subjects' preference reports (and binary comparison choices) would work against us finding evidence that subjects reported their preferences accurately enough for A-CEEI to outperform the bidding points auction, failure to report preferences carefully could certainly contribute to preference-reporting errors.

<sup>40</sup> We report marginal effects so that the coefficients can be interpreted as the change in probability of a contradiction and cluster our standard errors at the subject level to account for correlations in the errors for each subject. Online Appendix I demonstrates the robustness of the results presented in Table 3.

<sup>41</sup> If we rerun the regression in Column (1) of Table 3 with the reported preference utility difference between schedules as an additional control variable, the coefficient on *Cardinal* is 0.150 with a z-statistic of 5.41. This is lower than the coefficient without such a control (0.208), which suggests that comparisons that rely on cardinal information are more likely to be a close call than comparisons that rely on ordinal information only, but the coefficient remains economically large, which suggests that reporting cardinal preference information is per se difficult. Similarly, if we rerun the regression in Column (1) of Table 3 controlling for a dummy for whether the binary comparison reflected a slight preference, preference, or strong preference—which may be correlated with the difficulty of reporting preferences consistent with the binary comparison—the coefficient on *Cardinal* becomes 0.176 with a z-statistic of 6.19.

<sup>42</sup> These results are consistent with evidence from Rees-Jones (2018), Rees-Jones and Skowronek (2018), and Hassidim et al. (2021) that market participants who perform poorly on academic measures are more likely to misreport their preferences in strategy-proof environments.

<sup>43</sup> We discussed with the Wharton committee whether to allow subjects to express adjustments over arbitrary sets of courses rather than just pairs, which, in principle, would make the language fully expressive. In these discussions, the committee concluded that arbitrary set-wise adjustments would be too complicated for students.

<sup>44</sup> For example, suppose a student wants to express that they want at most one out of a set of  $k$  classes. They could express this in principle using just pairwise adjustments, but it would take  $k \binom{k-1}{2}$  such adjustments (reporting that any two of the  $k$  courses together have negative total value). A simpler way to convey the same preferences would be to report a constraint of the form “at most one out of these  $k$ ” were the ability to do so provided. See Milgrom (2009) for an example of a preference-reporting language that allows agents to express preferences of this form: at most  $k$  out of set  $S$ . There are numerous analogous examples.

<sup>45</sup> With roughly 50,000 possible schedules in the laboratory, there are 50,000! possible ordinal preferences over schedules or roughly  $10^{12,499}$ . As such, the up to 19 binary comparisons we ask of subjects do not provide enough data to identify patterns in such a large set without prior guidance on where to look.

<sup>46</sup> There are 15 contradictions in which the lower utility schedule is balanced and the higher utility schedule is not, and 35 contradictions in which the lower utility schedule is contiguous and the higher utility schedule is not. If we run the regression reported in Table 3, column (4), separately on *balanced* and *contiguous*, the coefficient on *balanced* is 0.15 and significant at the 1% level, and the coefficient on *contiguous* is 0.034 and not significant. The large magnitude on *balanced* suggests this feature may be important to a meaningful proportion of students, but the number of observations is small.

<sup>47</sup> In the free-response component of our survey, several subjects specifically mentioned the top-10 widget as a helpful feature of the user interface.

<sup>48</sup> For more discussion of the benefits of strategy-proofness in market design see, for example, Pathak and Sönmez (2008, 2013), Roth (2008), Azevedo and Budish (2019), and Li (2017).

<sup>49</sup> Gérard Cachon, the chair of Wharton's course allocation redesign team, wrote to us in personal correspondence: “I have heard that this makes a difference—some students say ‘when I saw the prices, I understood why I got what I got.’”

<sup>50</sup> Although the survey question asked about “effectiveness” broadly, it was the only question asked about the BPA, and so responses are likely to be driven by feelings about the BPA on multiple dimensions.

<sup>51</sup> If we interpret the BPA as “competitive” because it is highly strategic and A-CEEI as “noncompetitive” because it is approximately strategy-proof, the finding echoes a famous finding in the gender literature (Niederle and Vesterlund 2007).

<sup>52</sup> Ideally, we would have used a school-wide survey to obtain true preferences from students during the last year of the Wharton bidding points auction; this would have allowed us to compare student outcomes from actual play of the bidding points auction to counterfactual play of A-CEEI, analogously to other studies that have used survey data, such as Budish and Cantillon (2012), de Haan et al. (2015), and Kapor et al. (2020). Unfortunately, the Wharton administration did not want to conduct such a survey, fearing that a survey of students' “true preferences” at the time they were participating in the bidding points auction would have been confusing, especially given that a school-wide announcement had been made concerning the adoption of the new, truthful mechanism. Because of the complexity of the equilibrium of the bidding points auction, it is an open question whether it is possible to infer true preferences from strategic play in the absence of such a survey (e.g., as He (2017), Agarwal and Somaini (2018), and Calsamiglia et al. (2020) are able to do in the school-choice setting).

<sup>53</sup> For further details on these data and the engineering details of the practical implementation, see Budish et al. (2017).

<sup>54</sup> A practical example of using a too-simple reporting language is the restriction on the ability of military cadets to trade off years of service against their desired military branch in cadet-branch matching (Sönmez and Switzer 2013). Also related are limitations on the length of preference lists in school choice (cf. Pathak and Sönmez 2013). See also Hatfield and Kominers (2017), who study theoretically how the design of the contract language in many-to-many matching affects whether preferences, as expressed through the language, are guaranteed to be substitutable and to yield a stable match.

## References

- Abdulkadiroğlu A, Sönmez T (2003) School choice: A mechanism design approach. *Amer. Econom. Rev.* 93(3):729–747.
- Abdulkadiroğlu A, Agarwal N, Pathak P (2017) The welfare effects of coordinated assignment: Evidence from the New York City high school match. *Amer. Econom. Rev.* 107(12):3635–3689.
- Abdulkadiroğlu A, Pathak P, Roth AE (2005a) The New York City high school match. *Amer. Econom. Rev.* 95(2):364–367.
- Abdulkadiroğlu A, Pathak P, Roth AE, Sönmez T (2005b) The Boston public school match. *Amer. Econom. Rev.* 95(2):368–371.
- Abdulkadiroğlu A, Pathak P, Roth AE, Sönmez T (2006) Changing the Boston school choice mechanism: Strategy-proofness as equal access. Working paper, Duke University, Durham, NC.
- Agarwal N, Somaini P (2018) Demand analysis using strategic reports: An application to a school choice mechanism. *Econometrica* 86(2):391–444.
- Akbarpour M, Nikzad A (2020) Approximate random allocation mechanisms. *Rev. Econom. Stud.* 87(6):2473–2510.
- Ausubel LM, Cramton P, Milgrom P (2006) The clock-proxy auction: A practical combinatorial auction design. Cramton P, Shoham Y, Steinberg R, eds. *Combinatorial Auctions* (MIT Press, Cambridge, MA), 212–259.
- Azevedo EM, Budish E (2019) Strategy-proofness in the large. *Rev. Econom. Stud.* 86(1):81–116.
- Bergemann D, Morris S (2005) Robust mechanism design. *Econometrica* 73(6):1771–1813.
- Bertrand M, Mullainathan S (2001) Do people mean what they say? Implications for subjective survey data. *Amer. Econom. Rev.* 91(2):67–72.
- Bogomolnaia A, Moulin H (2001) A new solution to the random assignment problem. *J. Econom. Theory* 100(2):295–328.
- Boutilier C (2002) A POMDP formulation of preference elicitation problems. Dechter R, Kearns M, Sutton R, eds. *AAAI/IAAI* (AAAI Press, Edmonton, AB), 239–246.
- Budish E (2011) The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *J. Political Econom.* 119(6):1061–1103.
- Budish E, Cantillon E (2012) The multi-unit assignment problem: Theory and evidence from course allocation at Harvard. *Amer. Econom. Rev.* 102(5):2237–2271.
- Budish E, Cachon G, Kessler J, Othman A (2017) Course match: A large-scale implementation of approximate competitive equilibrium from equal incomes for combinatorial allocation. *Oper. Res.* 65(2):314–336.
- Budish E, Che Y-K, Kojima F, Milgrom P (2013) Designing random allocation mechanisms: Theory and applications. *Amer. Econom. Rev.* 103(2):585–623.
- Calsamiglia C, Fu C, Güell M (2020) Structural estimation of a model of school choices: The Boston mechanism versus its alternatives. *J. Political Econom.* 128(2):642–680.
- Calsamiglia C, Haeringer G, Klijn F (2010) Constrained school choice: An experimental study. *Amer. Econom. Rev.* 100(4):1860–1874.
- Cantillon E, Pesendorfer M (2007) Combination bidding in multi-unit auctions. Center for Economic and Policy Research Discussion Paper, Washington DC.
- Castillo M, Dianat A (2016) Truncation strategies in two-sided matching markets: Theory and experiment. *Games Econom. Behav.* 98:180–196.
- Chamberlin EH (1948) An experimental imperfect market. *J. Political Econom.* 56(2):95–108.
- Chen L, Pu P (2004) Survey of preference elicitation methods. École Polytechnique Fédérale de Lausanne Technical Report IC/2004/67, Switzerland.
- Chen Y, Sönmez T (2006) School choice: An experimental study. *J. Econom. Theory* 127(1):202–231.
- Chen Y, Jiang M, Kesten O, Robin S, Zhu M (2018) Matching in the large: An experimental study. *Games Econom. Behav.* 110:295–317.
- Cramton P, Shoham Y, Steinberg R, eds. (2006) *Combinatorial Auctions* (MIT Press, Cambridge, MA).
- de Haan M, Gautier PA, Hessel O, van der Klaauw B (2015) The performance of school assignment mechanisms in practice. Institute of Labor Economics Discussion Paper No. 9118, Bonn, Germany.
- Ding T, Schotter A (2017) Matching and chatting: An experimental study of the impact of network communication on school-matching mechanisms. *Games Econom. Behav.* 103:94–115.
- Echenique F, Yariv L (2013) An experimental study of decentralized matching. Working paper, California Institute of Technology, Pasadena, CA.
- Echenique F, Wilson AJ, Yariv L (2016) Clearinghouses for two-sided matching: An experimental study. *Quant. Econom.* 7(2):449–482.
- Ehlers L, Klaus B (2003) Coalitional strategy-proof and resource-monotonic solutions for multiple assignment problems. *Soc. Choice Welfare* 21(2):265–280.
- Featherstone C, Niederle M (2016) Boston versus deferred acceptance in an interim setting: An experimental investigation. *Games Econom. Behav.* 100:353–375.
- Foley D (1967) Resource allocation and the public sector. *Yale Econom. Essays* 7(1):45–98.
- Fragiadakis DE, Troyan P (2019) Designing mechanisms to focalize welfare-improving strategies. *Games Econom. Behav.* 114:232–252.
- Fudenberg D, Tirole J (1991) *Game Theory* (MIT Press, Cambridge, MA).
- Gale D, Shapley L (1962) College admissions and the stability of marriage. *Amer. Math. Monthly* 69(1):9–15.
- Goeree JK, Holt CA (2010) Hierarchical package bidding: A paper & pencil combinatorial auction. *Games Econom. Behav.* 70(1):146–169.
- Hakimov R, Kesten O (2018) The equitable top trading cycles mechanism for school choice. *Internat. Econom. Rev.* 59(4):2219–2258.
- Hakimov R, Heller CP, Kübler D, Kurino M (2019) How to avoid black markets for appointments with online booking systems. Working paper, University of Lausanne, Lausanne, Switzerland.
- Hashimoto T (2018) The generalized random priority mechanism with budgets. *J. Econom. Theory* 177:708–733.
- Hassidim A, Romm A, Shorrer RI (2021) The limits of incentives in economic matching procedures. *Management Sci.* Forthcoming.
- Hatfield JW (2009) Strategy-proof, efficient, and nonbossy quota allocations. *Soc. Choice Welfare* 33(3):505–515.
- Hatfield JW, Kominers SD (2017) Contract design and stability in many-to-many matching. *Games Econom. Behav.* 101:78–97.
- He Y (2017) Gaming the Boston school choice mechanism in Beijing. Working Paper No. 15-551, Toulouse School of Economics, Rice University, Houston, TX.
- Kagel JH, Roth AE (2000) The dynamics of reorganization in matching markets: A laboratory experiment motivated by a natural experiment. *Quart. J. Econom.* 115(1):201–235.
- Kagel JH, Lien Y, Milgrom P (2010) Ascending prices and package bidding: A theoretical and experimental analysis. *Amer. Econom. J. Microeconomics* 2(3):160–185.

- Kahneman D, Knetsch JL, Thaler RH (1990) Experimental tests of the endowment effect and the Coase theorem. *J. Political Econom.* 98(6):1325–1348.
- Kapor AJ, Neilson CA, Zimmerman SD (2020) Heterogeneous beliefs and school choice mechanisms. *Amer. Econom. Rev.* 110(5):1274–1315.
- Klemperer P (2004) *Auctions: Theory and Practice* (Princeton University Press, Princeton, NJ).
- Kojima F (2009) Random assignment of multiple indivisible objects. *Math. Social Sci.* 57(1):134–142.
- Kominers SD, Teytelboym A, Crawford VP (2017) An invitation to market design. *Oxford Rev. Econom. Policy* 33(4):541–571.
- Krishna A, Ünver U (2008) Improving the efficiency of course bidding at business schools: Field and laboratory studies. *Management Sci.* 27(2):262–282.
- Levin J, Skrzypacz A (2016) Properties of the combinatorial clock auction. *Amer. Econom. Rev.* 106(9):2528–2551.
- Li S (2017) Obviously strategy-proof mechanisms. *Amer. Econom. Rev.* 107(11):3257–3287.
- McKinney CN, Niederle M, Roth AE (2005) The collapse of a medical labor clearinghouse (and why such failures are rare). *Amer. Econom. Rev.* 95(3):878–889.
- Milgrom P (2004) *Putting Auction Theory to Work* (Cambridge University Press, Cambridge, UK).
- Milgrom P (2009) Assignment messages and exchanges. *Amer. Econom. J. Microeconomics* 1(2):95–113.
- Milgrom P (2011) Critical issues in the practice of market design. *Econom. Inquiry* 49(2):311–320.
- Milgrom P, Segal I (2020) Clock auctions and radio spectrum reallocation. *J. Political Econom.* 128(1):1–31.
- Milgrom PR, Tadelis S (2019) How artificial intelligence and machine learning can impact market design. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 567–585.
- Moulin H (1995) *Cooperative Microeconomics* (Prentice Hall, London).
- Myerson R (1991) *Game Theory: Analysis of Conflict* (Harvard University Press, Cambridge, MA).
- Narita Y (2016) Match or mismatch: Learning and inertia in school choice. Working paper, Yale University, New Haven, CT.
- Nguyen T, Vohra R (2020) Improvement properties in preferences and equilibria with indivisibilities. Working paper, Purdue University, West Lafayette, IN.
- Nguyen T, Peivandi A, Vohra R (2016) Assignment problems with complementarities. *J. Econom. Theory* 165:209–241.
- Niederle M, Roth AE (2009) Market culture: How rules governing exploding offers affect market performance. *Amer. Econom. J. Microeconomics* 1(2):199–219.
- Niederle M, Vesterlund L (2007) Do women shy away from competition? Do men compete too much? *Quart. J. Econom.* 122(3):1067–1101.
- Nisan N (2006) Bidding languages for combinatorial auctions. Cramton P, Shoham Y, Steinberg R, eds. *Combinatorial Auctions* (MIT Press, Cambridge, MA), 215–232.
- Othman A, Budish E, Sandholm T (2010) Finding approximate competitive equilibria: Efficient and fair course allocation. Wiebe van der Hoek W, Kaminka GA, Luck M, Sen S, eds. *Proc. Ninth Internat. Conf. Autonomous Agents Multiagent Systems (AAMAS, Toronto)*, 873–880.
- Pais J, Pintér Á (2008) School choice and information: An experimental study on matching mechanisms. *Games Econom. Behav.* 64(1):303–328.
- Pápai S (2001) Strategyproof and nonbossy multiple assignments. *J. Public Econom. Theory* 3(3):257–271.
- Parkes DC (2005) Auction design with costly preference elicitation. *Ann. Math. Artificial Intelligence* 44(3):269–302.
- Pathak PA, Sönmez T (2008) Leveling the playing field: Sincere and sophisticated players in the Boston mechanism. *Amer. Econom. Rev.* 98(4):1636–1652.
- Pathak PA, Sönmez T (2013) School admissions reform in Chicago and England: Comparing mechanisms by their vulnerability to manipulation. *Amer. Econom. Rev.* 103(1):80–106.
- Prendergast C (2017a) The allocation of food to food banks. Working paper.
- Prendergast C (2017b) How foodbanks use markets to feed the poor. *J. Econom. Perspect.* 31(4):145–162.
- Rassenti SJ, Smith VL, Bulfin RL (1982) A combinatorial auction mechanism for airport time slot allocation. *Bell J. Econom.* 13(2):402–417.
- Rees-Jones A (2018) Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. *Games Econom. Behav.* 108:317–330.
- Rees-Jones A, Skowronek S (2018) An experimental investigation of preference misrepresentation in the residency match. *Proc. Natl. Acad. Sci. USA.* 115(45):11471–11476.
- Reguant M (2014) Complementary bidding mechanisms and startup costs in electricity markets. *Rev. Econom. Stud.* 81(4):1708–1742.
- Reny PJ (2017) Assignment problems. *J. Political Econom.* 125(6):1903–1914.
- Roth AE (2002) The economist as engineer. *Econometrica* 70(4):1341–1378.
- Roth AE (2008) What have we learned from market design? *Econom. J.* 118(527):285–310.
- Roth AE (2015a) Experiments in market design. Kagel JH, Roth AE, eds. *The Handbook of Experimental Economics*, vol. 2 (Princeton University Press, Princeton, NJ), 290–346.
- Roth AE (2015b) *Who Gets What—and Why: The New Economics of Matchmaking and Market Design* (Houghton Mifflin Harcourt, Boston).
- Roth AE (2018) Marketplaces, markets, and market design. *Amer. Econom. Rev.* 108(7):1609–1658.
- Roth AE, Peranson E (1999) The redesign of the matching market for American physicians: Some engineering aspects of economic design. *Amer. Econom. Rev.* 89(4):748–782.
- Roth AE, Sönmez T, Ünver U (2004) Kidney exchange. *Quart. J. Econom.* 119(2):457–488.
- Roth AE, Sönmez T, Ünver U (2005) Pairwise kidney exchange. *J. Econom. Theory* 125(2):151–188.
- Roth AE, Sönmez T, Ünver U (2007) Efficient kidney exchange: Coincidence of wants in markets with compatibility-based preferences. *Amer. Econom. Rev.* 97(3):828–851.
- Roth AE, Prasnikar V, Okuno-Fujiwara M, Zamir S (1991) Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *Amer. Econom. Rev.* 81(5):1068–1095.
- Rustichini A, Satterthwaite MA, Williams SR (1994) Convergence to efficiency in a simple market with incomplete information. *Econometrica* 62(5):1041–1063.
- Sandholm T, Boutilier C (2006) Preference elicitation in combinatorial auctions. Cramton P, Shoham Y, Steinberg R, eds. *Combinatorial Auctions* (MIT Press, Cambridge, MA), 233–263.
- Smith V (1962) An experimental study of competitive market behavior. *J. Political Econom.* 70(2):111–137.
- Sönmez T, Switzer T (2013) Matching with (branch-of-choice) contracts at the United States Military Academy. *Econometrica* 81(2):451–488.
- Sönmez T, Ünver U (2010) Course bidding at business schools. *Internat. Econom. Rev.* 51(1):99–123.