NBER WORKING PAPER SERIES

BRINGING REAL MARKET PARTICIPANTS' REAL PREFERENCES INTO THE LAB: AN EXPERIMENT THAT CHANGED THE COURSE ALLOCATION MECHANISM AT WHARTON

Eric Budish
Judd B. Kessler

Working Paper 22448
http://www.nber.org/papers/w22448

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2016

The authors thank Gerard Cachon, without whom this study would never have been possible, and The Wharton School, in particular the Course Allocation Redesign Team and Wharton Computing. We thank Adriaan Ten Kate and Xan Vongsathorn for excellent research assistance. We 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, and Princeton. Disclosure: the market design theory in Budish (2011) and the computational procedure in Othman, Budish and Sandholm (2010) are in the public domain. The software implementation of the CEEI mechanism at Wharton was funded by and is owned by Wharton. If Wharton commercializes its software then royalties could accrue to a market design research lab overseen by Budish. Wharton had no right of prior review of the present study. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Bringing Real Market Participants' Real Preferences into the Lab: An Experiment that Changed the Course Allocation Mechanism at Wharton
Eric Budish and Judd B. Kessler
NBER Working Paper No. 22448
July 2016
JEL No. C78,C9,D47

ABSTRACT

This paper reports on an experimental test of a new market design that is attractive in theory but makes the common and potentially unrealistic assumption that “agents report their type”; that is, that market participants can perfectly report their preferences to the mechanism. Concerns about preference reporting led to a novel experimental design that brought real market participants’ real preferences into the lab, as opposed to endowing experimental subjects with artificial preferences as is typical in market design. The experiment found that market participants were able to report their preferences “accurately enough” to realize efficiency and fairness benefits of the mechanism even while preference reporting mistakes meaningfully harmed mechanism performance. The experimental results persuaded the Wharton School to adopt the new mechanism and helped guide its practical implementation. It is hoped that the experimental design methodology may be of use to other market design researchers, either for evaluating or improving preference reporting for existing mechanisms or for bringing other new mechanisms that utilize rich preference information from theory to practice.

Eric Budish
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
eric.budish@chicagobooth.edu

Judd B. Kessler
The Wharton School
University of Pennsylvania
3620 Locust Walk
Philadelphia, PA 19104
and NBER
judd.kessler@wharton.upenn.edu

I. Introduction

The promise of market design research is that mechanisms designed using abstract microeconomic theory can be implemented in practice to solve real-world resource allocation problems. This promise has led to 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.[1] 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 endowing experimental subjects with artificial preferences as is typical in the market design experimental literature — that tested a new market design theory and helped shepherd it from theory to practice.[2]

The context is the problem of combinatorial assignment — matching bundles of indivisible objects to agents without the use of monetary transfers, e.g., 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,[3] while the mechanisms used in practice have been shown to have critical flaws.[4] A recent paper of Budish (2011) proposes a new mechanism for combinatorial assignment, called approximate competitive equilibrium from equal incomes (CEEI), which, unlike prior mechanisms, satisfies attractive properties of efficiency, fairness and incentives. 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 auction used widely at many educational institutions,[5] was having the kinds of efficiency, fairness, and incentives problems one would expect given the theoretical criticisms of the

---

[1] 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, Cramton and Milgrom (2006) and Levin and Skrzypacz (forthcoming). On matching markets, see Roth's (2015) book "Who Gets What and Why: The New Economics of Matching and Market Design", as well as Roth and Peranson (1999), Abdulkadiroğlu and Sönmez (2003), Abdulkadiroğlu et al. (2005a, 2005b, 2006), Roth, Sönmez and Ünver (2004, 2005, 2007) and Roth (2002, 2008).

[2] See Roth (forthcoming) for a survey of the literature on market design experiments as well as a detailed discussion of the present paper in Section 6.

[3] See Pápai (2001), Ehlers and Klaus (2003), Hatfield (2009) and Kojima (2009).

[4] See Sönmez and Ünver (2003, 2010), Krishna and Ünver (2008) and Budish and Cantillon (2012).

[5] 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.

mechanism (Sönmez and Ünver 2003, 2010) and the Wharton administration convened a committee to consider alternatives.

While attractive in theory, however, the CEEI mechanism makes an assumption that raises serious concern about its suitability for use in practice: "agents report their type". In Budish (2011), a student's "type" is an ordinal preference relation over all possible schedules of courses, much as in general-equilibrium theory a household's type is an ordinal preference relation over all possible consumption bundles. As is completely standard in mechanism design theory (Fudenberg and Tirole 1991, Myerson 1991, Bergemann and Morris 2005), agents are assumed to be able to simply "report their type" to the mechanism. But this assumption often strains reality, and CEEI is clearly such a case. In a context such as Wharton's, there might be hundreds of millions of schedules in a given semester.

Clearly, in such settings, "perfect" preference reporting is an unrealistic goal. 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 reap the benefits of the mechanism. Let us make this question more precise. In any practical implementation of the 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 — that is, whether the ordinal ranking generated by the preference data they report is close enough to the true preferences in their minds — that the efficiency and fairness benefits of CEEI are realized.

This positive question about 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 CEEI, how can we compare the ordinal ranking generated from the data agents report to the mechanism to agents' true preferences? How can we measure the extent to which inaccurate preference reporting harms mechanism performance?

We designed a novel kind of experiment to answer these questions. Before describing our experimental design, we first explain why a new kind of experimental design is needed. The traditional method used in market design experiments is to endow subjects with artificial preferences and offer monetary rewards based on how well the subjects perform in the mechanism as evaluated based on these preferences. For example, if in a multi-object matching experiment a subject is endowed with a value of $25 for the bundle {A, B}, and then obtains the bundle {A, B} in the laboratory matching market, the subject would be compensated with a payment of $25. While this technique has been extremely important in the history of market design experiments and is invaluable for answering certain kinds of questions,[6] it is a non-starter for our setting because it assumes away the central issue of the difficulty of reporting one's preferences. If we endowed subjects with artificial preferences in a format that could be immediately reported to the mechanism, we would just be telling subjects their preferences and asking them to tell them right back to us, trivializing the reporting task.[7] If we endowed subjects with artificial preferences in a format different from what could be reported to the mechanism, this too misses the central question of interest. This latter exercise tests whether subjects can translate between preferences in one language the researcher created (for conveying preferences to the subject) and another language the researcher created (for reporting preferences to the mechanism). This does not test whether real market participants can translate their own real preferences — however these preferences are represented in their own minds — into data the mechanism can use.[8]

Instead of endowing experimental subjects with artificial preferences, our experimental design brought real market participants' real preferences into the lab. Specifically, our experimental subjects were Wharton MBA students who were asked to report their real

---

[6] In Roth's (forthcoming) recent survey of the literature on market design experiments, every laboratory experiment discussed uses the endowed preferences methodology with the exception of the present paper, which is discussed in detail in Section 6. Some recent examples of market design experiments using the endowed preferences methodology include Kagel and Roth (2000); Chen and Sönmez (2006); Pais and Pinter (2008); Calsamiglia, Haeringer and Klijn (2010); Goeree and Holt (2010); Kessler and Roth (2012); Featherstone and Niederle (2014); Featherstone and Mayefsky (2014); Fragiadakis and Troyan (2014); Kessler and Roth (2014); and Li (2015). See especially Kagel, Lien and Milgrom (2010) for an interesting twist on the methodology that uses theory and simulations to guide which endowed preferences to explore. Outside of market design it is common to design laboratory experiments around participants' real preferences; famous examples include Kahneman, Knetsch and Thaler (1990) and Roth et al. (1991). What makes market design lab experiments different is that theory testing has, outside of the present paper, relied on precise knowledge of subjects' heterogeneous preferences, whereas, e.g., in dictator or 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.

[7] Given that the CEEI mechanism is approximately strategy-proof and we informed subjects as such (as Wharton did in practical implementation), this would simply be testing whether subjects believed the claim in the instructions that it is in their best interest to report their preferences truthfully. This is an interesting question in its own right (cf. Hassidim, Romm and Shorrer 2016; Li 2015) but not the question of interest here. See further discussion in Section 2.2.

[8] By way of analogy, turning one's mental representation of preferences into data that can be reported to the mechanism can be thought of as a translation exercise. Imagine a market participant's mental representation of preferences is in English and data must be entered into the mechanism in Latin. If we endowed preferences in Latin and asked the subject to report preferences in Latin, this would be trivial (it would simply be a transcription exercise). If, instead, we endowed preferences in some other language, say Greek, we could test whether the subject could translate Greek into Latin, but this is fundamentally different than the test of whether the subject can translate English to Latin. Consequently, we cannot test our fundamental question with endowed preferences unless we could somehow endow preferences in English, but this would require us to know the structure of agents' mental representations of their preferences, which is fundamentally unknowable to the researcher.

preferences over schedules of real Wharton courses using a realistic, professionally designed user interface. We then generated our primary data on preference reporting accuracy and mechanism performance using binary comparisons — questions of the form: "Do you prefer Schedule A or Schedule B?" The rationale behind the binary comparisons method is that while reporting preferences over all possible schedules via a preference reporting language is cognitively hard and likely to be inaccurate, reporting which one prefers between two specific schedules is cognitively simple and likely to be accurate. As will be described in detail in Section 2.4, our experiment used carefully tailored binary comparisons to generate data on preference reporting accuracy as well as the efficiency and fairness of the mechanism. Additionally, by comparing the performance of the mechanism on efficiency and fairness measures based on binary comparisons to measures based on the reported preferences, we can quantify the harm caused by preference reporting mistakes.

In addition to allowing us to test the "agents report their type" assumption, there were two other advantages to using real market participants' real preferences. First, it 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 (cf. Roth, forthcoming); using real market participants' real preferences yields a more realistic demonstration. Second, it enabled a search for "side effects" of the mechanism; that is, issues left out of the theory that might be important for practice.[9] Issues left out of the theory are especially of concern here because CEEI had never been used before; most other market design implementations have had direct precedents that assuage this concern.[10] 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.

A disadvantage of our experimental approach is that subjects' behavior is not incentivized. This lack of incentives likely caused subjects to exert less effort in the

[9] 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.

[10] 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; this lessens 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.

laboratory than they would have if 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 mechanism. That is, the lack of incentives should make it harder for us to find results, but should not undermine any results that we do find (see Section 2.5 for a discussion).[11]

We briefly summarize the main results. Students reported their preferences "accurately enough" that CEEI outperformed the benchmark, the incumbent Wharton Auction, on each of our quantitative measures of efficiency and fairness, with most (though not all) differences statistically significant at the 5% level. The magnitudes were modest but all broadly consistent with the theory. However, we also found 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 CEEI over the Wharton Auction would have been substantially larger if not for preference reporting mistakes. The only negative side effect we found in the survey was that students found CEEI to be somewhat of a "black box", i.e., non-transparent. We also found an unanticipated positive side effect, which is that CEEI eliminated a gender disparity in liking of the mechanism that was present for the Wharton Auction in both the laboratory survey and in a student-wide administration survey.

The experiment persuaded Wharton to adopt CEEI — implemented as "Course Match" beginning in Fall 2013 — and guided several aspects of its practical implementation.[12] Some limited data from the first year of implementation demonstrates that CEEI has increased equity in both total expenditure and the distribution of popular courses, and survey data suggest that 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 Auction to 53% in the first year of 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%.

Beyond providing empirical evidence that CEEI is a good solution to the combinatorial assignment problem, our paper makes three contributions to the broader market design

[11] Also note that the lack of incentives is not intrinsically a feature of the experimental design methodology we propose (i.e. using real agents' real preferences plus binary comparisons). If we could have offered with some probability that students would obtain in real life the schedule they obtained in the lab version of the mechanism, or a schedule they chose in a binary comparison, then all behavior would be incentivized. However, we were unable to get the Wharton administration to provide such stakes in the laboratory experiment, for the obvious reasons.

[12] 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, Cachon, Kessler and Othman (forthcoming).

literature. First, the paper contributes to an ongoing dialogue in the market design literature about the importance of preference reporting and language design (cf. Milgrom 2009, 2011). We provide some of the first documented empirical evidence on the prevalence of preference reporting errors and the harm they can cause to a mechanism's performance (see also Hassidim, Romm and Shorrer 2016), while at the same time showing that participants can report complex preferences "accurately enough" to realize the benefits of a mechanism with complex reporting requirements.

Second, the paper introduces a new experimental design methodology that allows researchers to evaluate market designs in the laboratory using real market participants' real preferences and appropriate binary comparisons. This methodology can be used to evaluate other market designs with non-trivial preference reporting requirements. This methodology may also be useful for evaluating decision supports for market designs, i.e., 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 CEEI, but also in settings where the preference reporting per se is simple but thinking through one's preferences is difficult, e.g., school choice (cf. Narita 2016). By comparing subjects' ability to report their preferences with and without a particular decision support, our methodology can identify the efficacy of that decision support and help optimize the performance of existing market designs.

Last, our paper contributes a new theory-to-practice "success story" to the market design literature. This is valuable for two related reasons. First, market design implementations beget further market design implementations. The Wharton Committee was already familiar with the work done by economists re-designing 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 lab 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.

Second, as emphasized by Roth (2002), academic work on the practical implementation of market design theory is an important complement to the theory itself. It 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) 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."

The remainder of this paper is organized as follows. Section 2 describes the experimental design. Section 3 presents the 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.

II. Experimental Design

132 Wharton MBA students participated in 8 experimental sessions, in groups of 14 to 19 subjects per session, conducted in a computer lab at Wharton during the week of November 28, 2011.[13] These subjects were recruited with an email sent by the Wharton Dean's office, which stressed that the study was voluntary but also indicated that participation was appreciated by the Dean's office and as a further inducement offered $250 to two randomly selected subjects per session. The recruitment email did not mention that the study was about course assignment, and we asked subjects not to discuss the study with other students after they participated.[14]

Each study session began with general instructions that gave an overview of the experimental procedure. (For the full text of the instructions see Appendix A.) Subjects were given a list of 25 Wharton course sections for the upcoming spring semester, chosen by the Wharton Course Allocation Redesign Team (the "Wharton Committee") to be representative of course offerings in the upcoming semester with a tilt towards popular courses (see the list of courses and sample descriptions in Appendix B). Each course section had a capacity of 3 to 5 seats and subjects were informed that they needed a schedule of 5 courses.

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

[13] Three pilot sessions were run with MBA students in the week preceding the experiment. During these sessions, a number of bugs in the CEEI code were identified and fixed such that the experiment would successfully solve for CEEI prices (which did not happen in the first pilot session and which took too long in the other two sessions for the experiment to be completed). After these pilot sessions, the experimental instructions were finalized (i.e., changed for clarity and length) for the eight sessions conducted the week of November 28, 2011.

[14] See Appendix C for the text of the recruitment email. The reason the recruitment email was vague and did not mention the purpose of the study is that we wanted to attract student subjects who were generally representative of the Wharton MBA student body and to avoid attracting students who were disproportionally 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 Auction (see Table A1 in Appendix D).

to use each system to obtain the best schedule they could given their own preferences, imagining it was their last semester at Wharton.[15] We then gave subjects five minutes to look over the course offerings and think about their preferences before describing the first mechanism.

In half of the sessions we ran the Auction first, and for half of the sessions we ran CEEI first.[16] For each mechanism we: read instructions for that specific mechanism (see details of the mechanisms in Sections 2.2-2.3), had subjects participate in that mechanism, and then asked survey questions about their experience with the mechanism.

After subjects had participated in both mechanisms, we asked them to make a series of "binary comparisons" between pairs of schedules. These binary comparisons, described in detail in Section 2.4, were designed to provide tests of efficiency, fairness, and preference reporting accuracy. Subjects then completed another set of survey questions and provided free-form response comments.

2.1 Wharton Bidding Points Auction

At the time of the experiment, Wharton's 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 Auction 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+1st 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.[17]

---

[15] Here is some of the key relevant text from the experimental instructions: "Please try to construct your most preferred schedule given the courses that are available." "Think about how interested you are in each of the courses and what would be your ideal schedule or schedules." "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."

[16] We did not find any significant differences in the results based on which mechanism was used first. See Appendix E for details of this analysis.

[17] While 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 dollars in a real-money auction and loses to someone who bids 10,001 may be disappointed, but at least they can put their 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 (2015) for an implementation of a mechanism in this spirit in the context of allocating donated food to food banks across the US.

Our laboratory implementation of the Wharton Auction was as similar as possible to the real Wharton Auction subject to the constraints of the laboratory. For time considerations, we used four rounds instead of nine.[18] 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 auction so that it would be as familiar as possible to subjects. The instructions for the Auction (see Appendix A) were familiar as well, since all subjects had previously used the real-world version of the mechanism to pick their courses.

2.2 Approximate Competitive Equilibrium from Equal Incomes (CEEI)

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),[19] (iii) the computer finds (approximate) market-clearing prices, (iv) each student is allocated her most preferred affordable schedule — the affordable schedule she likes best given her report in step (i) based on her budget set in step (ii) and the prices found in step (iii).[20]

The instructions (see Appendix A) explained the CEEI mechanism, which was unfamiliar to the subjects, and explained to the subjects that their only responsibility in using the mechanism was to tell the computer their preferences over schedules; the computer would then compute market-clearing prices and buy them the best schedule they could afford at those prices. Because the mechanism is approximately strategy-proof, and moreover it is highly unclear how one could profitably manipulate the mechanism even in small markets, the instructions advised students that it is in their best interest to report their preferences truthfully.[21] The instructions used the metaphor of providing

---

[18] 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 lab, too, most of the action took place in the first round.

[19] Budish's (2011) result that prices exist for CEEI that (approximately) clear the market requires that students have non-identical budgets. 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 1 available seat. If A's budget is 5000 and B's budget is 5001, then setting the price of course X to 5001 clears the market because B can afford it while A cannot. The Auction breaks ties in the auction itself rather than in the budgets. If both A and B bid 5000 points for course X, then the computer randomly selects one student to transact.

[20] See Budish (2011) for a more complete description of how CEEI works. See Othman, Budish, and Sandholm (2010) and Budish, Cachon, Kessler and Othman (forthcoming) for how to calculate the market-clearing prices in step (iii).

[21] At any realized prices, truthful reporting is best because it ensures the student receives her most-preferred affordable bundle at those prices. For it to be profitable for a student to benefit from misreporting her 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

instructions to someone shopping on your behalf to explain the rationale for reporting one's true preferences as accurately as possible.

2.3 Preference Reporting Language of CEEI

As emphasized in the Introduction, the theory behind CEEI makes the unrealistic assumption that agents can "report their type" — that is, an ordinal ranking over all feasible schedules — so that the mechanism can always select the agent's most-preferred affordable bundle from any possible choice set. In any practical implementation of CEEI, agents cannot be expected to directly report preferences over all possible bundles. Instead, agents will need to supply a more limited set of information that describes their preferences, using what is called a preference reporting language (cf. Milgrom 2009, 2011).

The preference reporting language we implemented in the lab, 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 0.[22] 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.[23] 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.

---

from the bundle the mechanism thinks she likes best at *p'* (based on her misreport *u'*) than from the bundle she likes best at *p* (based on her true preferences *u*). The main reason why such misreports are likely to be hard to find even in small markets is that students require at most 1 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 since reducing demand means pretending to want 0 units instead of 1 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 non-intuitive. Since 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 felt comfortable advising students to report truthfully. If either of the authors of this paper were participating in this market design, even in a small economy like the ones used in the laboratory, we would report truthfully.

[22] We recommended reporting a positive value for at least twelve course sections to ensure receipt of a complete schedule of five courses.

[23] 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, e.g., in Nisan (2006). We will explore limitations of the language in further detail below in Section 4.

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.[24] Observe that the cardinal preference information subjects submit for individual courses and pairs of courses induces an ordinal ranking over all feasible schedules, i.e., the language allows subjects to report a "type".

We emphasize that while 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 non-additive 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-ten widget" (see Figure 2), 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 ten 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, e.g., if they realized the ten schedules listed were not their favorites or were in the wrong order. Students were given 10 minutes to report their preferences.

[24] 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 CEEI solves billions of integer programs in the process of finding approximate market clearing prices. See Budish, Cachon, Kessler and Othman (forthcoming) for more details on the computational procedure.

Figure 1: Screenshot of CEEI User Interface

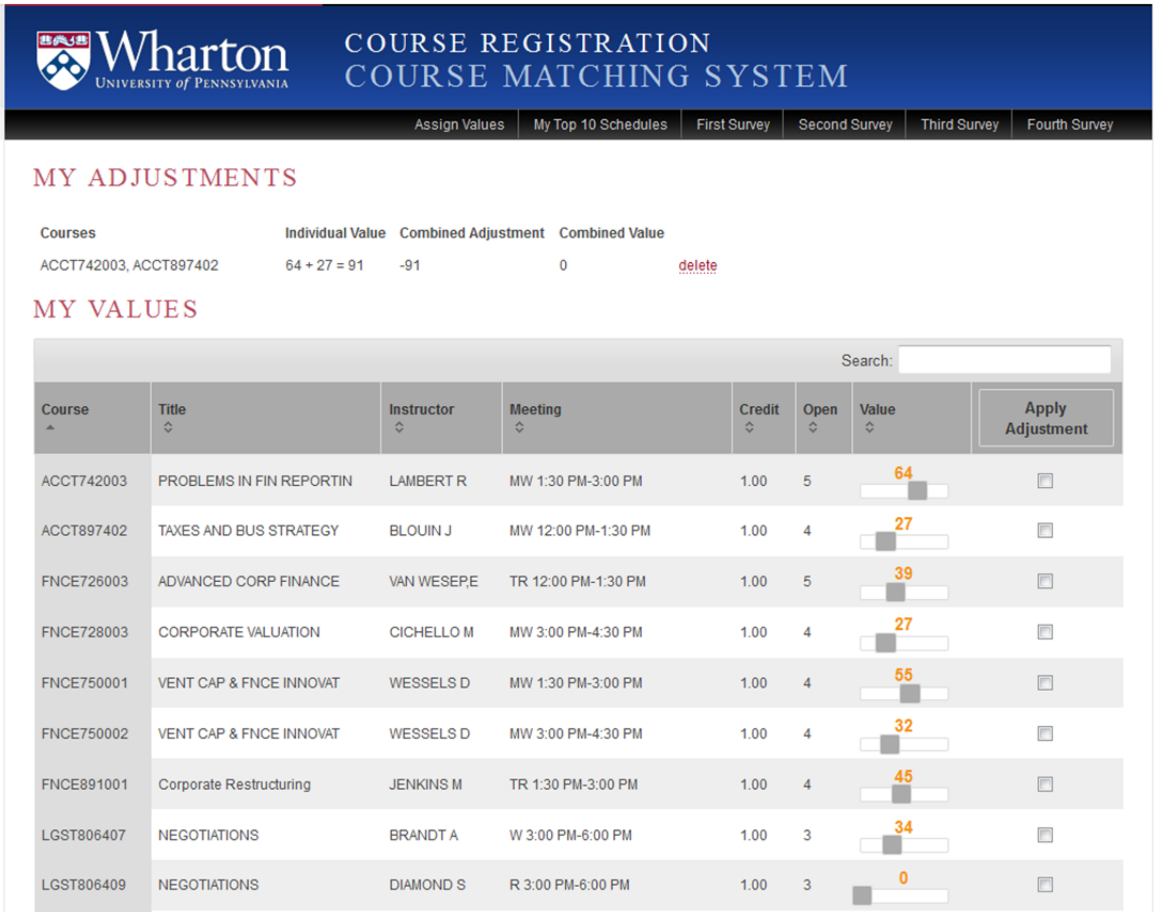

Figure 1 is a screenshot of the top of the user interface for preference reporting. Of the nine course sections that are visible, the subject has reported positive values for the first eight. To make adjustments, subjects clicked two checkboxes in the far right column of the interface and were prompted to enter the adjustment in a dialog box. Any previously entered adjustments were listed at the top of the interface. The 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 his schedule together is worth 0, effectively reporting that the subject wants one or the other, but not both, accounting courses.

Figure 2: Screenshot of Top Ten Widget

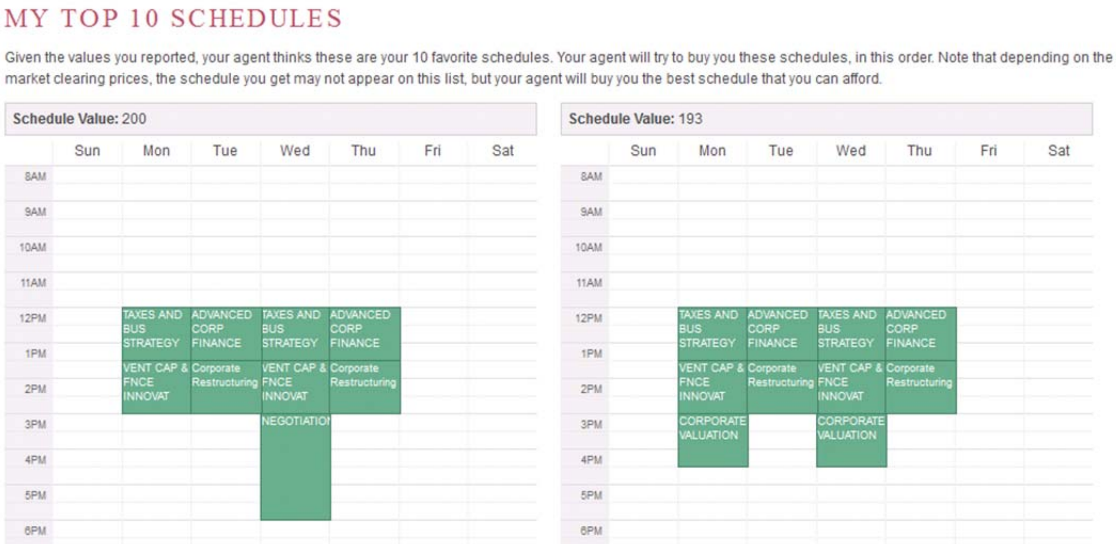

Figure 2 is a screenshot of the top of the "top ten widget". It shows two feasible schedules of 5 courses each (e.g. "Taxes and Business Strategy meets from 12:00-1:30 on Monday and Wednesday in both schedules) and the sum of the cardinal reports, listed as "Schedule Value." The rest of the top ten schedules were shown below these, and subjects could scroll down the screen to see all ten.

Figure 3: Screenshot of Binary Comparison Question

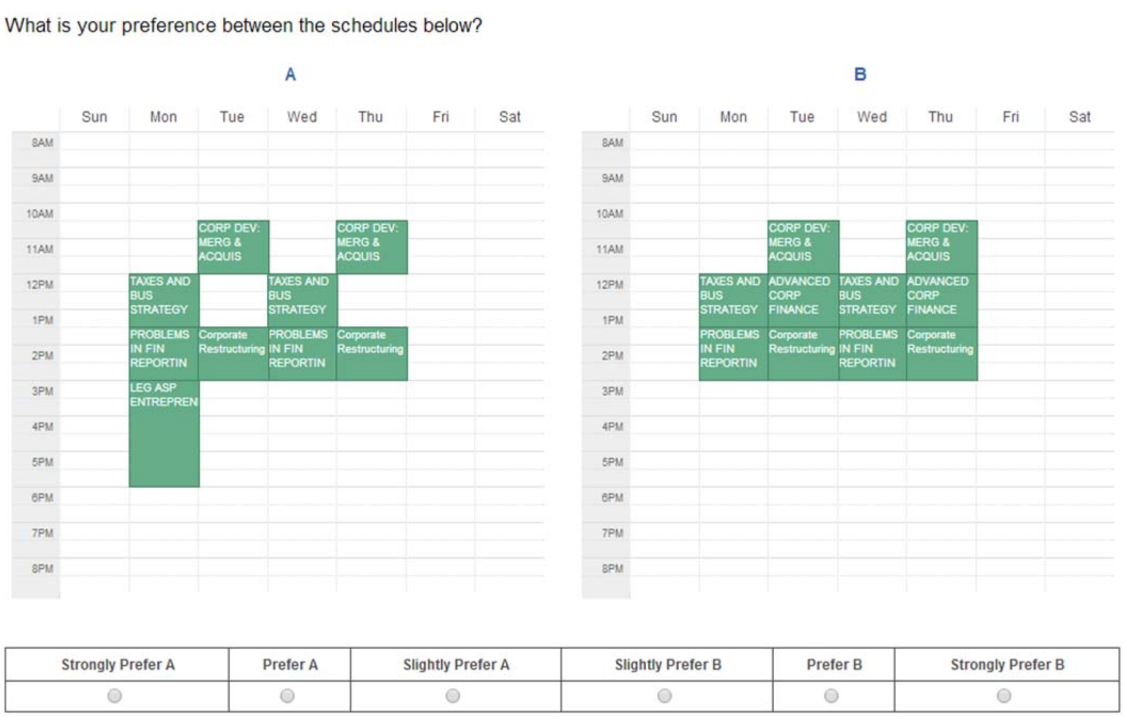

Figure 3 is a screenshot of a binary comparison. It shows two schedules and asks the subject to pick which of the two she prefers.

2.4 Binary comparisons

After using both mechanisms, subjects were asked to make a series of up to 19 binary comparisons between schedules, reporting which of two schedules they preferred and whether they "slightly prefer," "prefer," or "strongly prefer" the schedule they prefer (see Figure 3). The logic behind this methodology is that while reporting ordinal preferences over every possible schedule using the reporting language in the experiment is complex, making a binary comparison between two schedules is simple.

We designed the set of binary comparisons to allow us to perform a number of tests within each mechanism and between mechanisms. Subjects' first and last binary comparisons were between the schedule the subject received under CEEI and the schedule she received under the Auction. This comparison was asked twice, as the first question and the last question, with the order of the schedules reversed.[25] These binary comparisons were used to construct a simple social welfare comparison between the two systems as a measure of the efficiency of the two mechanisms.

Up to twelve binary comparisons per subject were asked to measure envy under the two mechanisms. Envy occurs when an individual prefers someone else's schedule to his own schedule. For each mechanism, each subject was asked to compare his schedule from that mechanism to up to six schedules that other subjects in his session received from the mechanism (e.g., he compared his CEEI schedule to others' CEEI schedules and his Auction schedule to others' Auction schedules).[26] The envy observed in CEEI and the Auction were compared to each other to identify the relative fairness of the mechanisms.

Note that using these binary comparisons to test whether the CEEI realized schedules are preferred to the Auction realized schedules, and to test whether the CEEI outcome achieved less envy than the Auction outcome, is necessarily a joint test of preference reporting and the mechanisms. That is, these comparisons answer the question: is preference reporting accurate enough that CEEI is able to outperform the Auction on measures of efficiency and fairness.

To further explore subjects' ability to accurately report their preferences, we included five binary comparisons involving the realized CEEI schedule and schedules subjects would have received under CEEI if their budget had been 10% or 30% higher or 10% or 30% lower than it actually was. These comparisons are natural tests of preference reporting accuracy, since subjects should always prefer the schedule they obtain with a

[25] 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.
[26] The others' schedules were chosen to be somewhat desirable to the subject so that the comparisons would be non-trivial. In particular, each subject was randomly shown up to six schedules from the set of others' realized schedules that generated at least 50% of the utility as the subject's realized schedule under CEEI, based on the subject's reported preferences.

larger budget. In fact, however, all binary comparisons are tests of the reporting language, because for each binary comparison we can analyze whether the subject's binary choice between schedules is consistent with the preference information supplied to the CEEI preference reporting language.

2.5 Incentives

Before we begin to discuss the results, we want to return to the issue discussed in the Introduction that decisions in our experiment are not incentivized. Since we did not endow subjects' preferences, we do not know which schedules they actually prefer and so cannot pay them more for getting more-preferred schedules or making more accurate binary comparison choices. This raises the concern that subjects might not work as hard in the experiment as if they were being paid. We took care in the design to ensure that such lack of effort, if present in the lab, would bias against our main hypothesized results.

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, while non-triers exert zero effort in the mechanisms and their binary comparison responses are pure noise, i.e., 50/50 coin flips. This noise from the non-triers biases towards less accurate preference reporting under CEEI and less ability to detect a difference in efficiency or fairness between CEEI and the Auction. This pushes against our main results: 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 CEEI improves efficiency and fairness relative to the Auction.

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 Auction is familiar to subjects while CEEI is unfamiliar, we might expect partial effort to harm CEEI more than the Auction, which also pushes against our main results.

A concern that is distinct from low effort, and which would bias some results in our hypothesized direction, is if the students in the lab disliked the Wharton Auction in practice and thus attempted to sabotage its performance in the lab. While 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, both on demographic measures and, crucially, on their perception of the Wharton Auction's effectiveness (see Appendix Table A1). We used anonymous University IDs to match experimental subjects to data from an administration survey conducted at the end of each school year. The subjects in the lab rated the Wharton Auction's "effectiveness" an average of 4.69 on a scale of 0 to 7, essentially identical to the average for the whole student body of 4.68. Second, our results are unaffected by which mechanism was conducted first, i.e., whether subjects first used the Auction or first used CEEI (see Appendix Tables A2 and A3). Subjects who used the Auction first did so without knowing any details of the other mechanism, while subjects who used the Auction second did so after first using CEEI. Third, subjects were recruited to the experimental sessions by an email that came from the Wharton administration[27] and were explicitly asked in the experimental instructions to take their decisions seriously in the lab just like they do in real life. Our impression is that the Wharton students in the laboratory took this direction seriously. It is also worth noting that to the extent that subjects care about helping future Wharton students get the best course allocation mechanism possible, there was a non-monetary incentive for subjects to take the task seriously and answer questions honestly.

III. Results on Fairness and Efficiency

In this section, we present our results on fairness and allocative efficiency. We present two sets of results. The first set of results is based on the binary comparisons and so is a joint test of the CEEI mechanism with the preference reporting language as compared to the Auction. The second set of results is based on the reported preferences under CEEI, and represents an isolated test of CEEI versus the Auction under the assumption that subjects report their preferences perfectly to CEEI. The first set of results quantifies the actual performance of CEEI given potentially imperfect preference reporting, while the second set of results provides an upper bound for how much performance might improve if subjects were able to report their preferences perfectly.

3.1 Results on Fairness

We begin with our results on fairness, which provide the most direct test of the theory in Budish (2011).

Student A is said to *envy* Student B if A prefers B's schedule to her own. An allocation is called *envy-free* if no student envies another. Envy-freeness, introduced by Foley (1967), is arguably the most important criterion of outcome fairness in the economics literature on distributive justice (Moulin 1995, Arnsperger 1994). Unfortunately, in an indivisible goods problem such as course allocation, it is impossible to eliminate envy altogether. If there is some star professor whose course all students want to take (and whose course they value over any other bundle of courses), the students who do not get that course will

[27] The recruitment email (Appendix C) did not mention course allocation to help ensure we did not attract people with particularly strong views on the existing auction.

envy the students who do. Budish (2011) shows, however, that CEEI approximately eliminates envy.[28] In the Auction, by contrast, there is no such guarantee.

Our binary comparisons directly tested for the presence of envy and its magnitude. Specifically, each subject was presented with a set of binary comparisons for each mechanism asking which schedule they preferred between their own schedule from that mechanism and another randomly chosen subject's schedule from that mechanism, as well as the intensity of that preference ("slightly prefer", "prefer", or "strongly prefer"). In total, 117 students completed binary comparisons looking for envy in CEEI and 119 completed binary comparisons looking for envy in the Auction.[29] Table 1 shows that CEEI generated less envy than the Auction, measured either by the percentage of subjects who display any envy in Panel A or by the percentage of binary comparisons across the entire experiment that resulted in envy in Panel B.[30] The difference is especially significant when we restrict attention to what we call "large" envy, which excludes cases of envy caused by only a slight preference for the other schedule.

Next, we look at the envy comparison under the assumption that preference reporting under CEEI is perfectly accurate. Under this assumption, we can look for envy by directly comparing a subject's utility from their own schedule to their utility from another subject's schedule. While in principle we can do this for all pairs of subjects in a session, we restrict attention to the pairs for which there were binary comparison tests as well to facilitate comparison with the results in Table 1. Table 2 displays the results.

[28] More precisely, envy only occurs because of the small randomness in students' budgets. A student with a budget of 5001 might envy a student with a budget of 5002, if there is some schedule that costs 5002 that the first student wants but cannot afford, but the student with a budget of 5001 will never envy a student with a budget of 5000. In addition, any envy that does occur is bounded in magnitude.

[29] We do not have data from all 132 subjects for two reasons. First, a bug in the code for the first three sessions prevented getting binary comparison data from the six subjects who received the same schedule under both CEEI and the Auction. Of the remaining 126 subjects, nine had no other subject in their session with a CEEI schedule that had at least 50% of their own CEEI utility and seven had no other subject in their session with an Auction schedule that had at least 50% of their own CEEI utility.

[30] We use one-sided tests for the analysis of fairness and efficiency in Section 3 since the prior based on the theory is that CEEI is more fair and more efficient than the Auction. We use two-sided tests for the analysis of preference reporting mistakes in Section 4 and the qualitative analysis in Section 5 since we do not have a theory-informed prior.

Table 1: Envy Under CEEI and Auction — Joint Test of the Mechanism and the Reporting Language, using Binary Comparisons

	Auction	CEEI	Probability Ratio Test (one-sided)
Panel A: By Subject (Auction: N = 119, CEEI: N = 117)			
% of subjects who display any envy of another subject's schedule	42%	31%	$p = 0.036$
% of subjects who display any large envy ("prefer" or "strongly prefer") of another subject's schedule	34%	21%	$p = 0.008$
Panel B: By Comparison (Auction: N = 499, CEEI: N = 475)			
% of comparisons in which the subject displays any envy of the other subject's schedule	19%	12%	$p = 0.002$
% of comparisons in which the subject displays any large envy ("prefer" or "strongly prefer") of the other subject's schedule	14%	8%	$p = 0.002$

Table 1 reports envy results based on binary comparisons. Panel A reports the percentage of subjects who displayed envy in one of the binary comparisons designed to test for envy. Panel B reports the percentage of the binary comparisons designed to test for envy in which subjects displayed envy.

Table 2: Envy Under CEEI and Auction — Isolated Test of the Mechanism, Assuming Perfect Preference Reporting

	Auction	CEEI	Probability Ratio Test (one-sided)
Panel A: By Subject (Auction: N = 119; CEEI: N = 117)			
% of subjects who envy another subject's schedule according to CEEI	29%	4%	$p < 0.001$
Panel B: By Comparisons (Auction: N = 499; CEEI: N = 475)			
% of comparisons in which one subject envies the other subject's schedule according to CEEI	15%	2%	$p < 0.001$

Table 2 reports the envy results based on the utility measure constructed from the CEEI preference reports. We analyze the pairs for which there was a binary comparison. When necessary, we made non-legal Auction schedules (i.e. schedules that had too many courses or courses that failed to satisfy scheduling constraints) legal by maximizing the subject's utility among courses in the Auction schedule subject to scheduling constraints.

Table 2 shows that, under the assumption of perfect preference reporting, just 4% of subjects exhibit envy under CEEI. This gap between the 4% in Table 2 and the 31% in Table 1 is the first indication that subjects had difficulty reporting their preferences to the CEEI mechanism: any subject who had no envy under the assumption of perfect reporting but had some envy based on binary comparisons must have failed to report their preferences accurately, since their preference based on reported utility contradicted their subsequent binary choice.

3.2 Results on Allocative Efficiency

We now turn to our results on allocative efficiency. The theory in Budish (2011) focuses on ex-post Pareto efficiency, that is, whether or not an allocation leaves Pareto-improving trades on the table (trades that make all students weakly better off with at least one strictly better off). Unfortunately, the binary comparisons data cannot be used to measure ex-post Pareto inefficiency, since they only contain data on subjects' preferences between a small number of pairs of *schedules*, rather than data about preferences over individual courses that could help us determine, e.g., if subjects A and B should swap course X for course Y. Instead, we use a simple measure of allocative efficiency that was important to the Wharton administration, namely, how many students prefer their CEEI schedule to their Auction schedule, both in aggregate and in each economy (i.e., each session of subjects who were competing for the same seats in courses).

Recall that we asked subjects to compare their CEEI schedule to their Auction schedule twice, with many other binary comparisons in between, and with the position of CEEI and Auction schedules flipped in the second binary comparison. We say that a subject prefers CEEI (Auction) if they report that they prefer their CEEI (Auction) schedule both times they were asked; we say that the subject is indifferent if their preference between the two reverses between the two binary comparisons. Subjects who had identical schedules did not see these binary comparison questions. Table 3 reports the allocative efficiency results.

As can be seen from Table 3, subjects preferred CEEI to the Auction by a margin of 56 to 42, or 57.1% to 42.9% (one-sided binomial probability test against the hypothesis that the ratio is 50%, $p = 0.094$), with seventeen students indifferent between the two schedules and seventeen students receiving exactly the same schedule under each.[31] At the session

[31] We get qualitatively similar results when we look at intensity-weighted preferences between CEEI and Auction schedules (see also Figure 4). If we use the entire response scale and code 1=strongly prefer Auction schedule, 2=prefer Auction schedule, 3=slightly prefer Auction schedule, 4=slightly prefer CEEI schedule, 5=prefer CEEI schedule and 6=strongly prefer CEEI schedule, we can compare average responses to 3.5, the midpoint of the response scale. If we drop subjects who got identical schedules, the mean response is 3.65 indicating a directional preference for CEEI (one-sided t-test that mean>3.5 yields p=0.187). If we code subjects who got identical schedules as 3.5, the mean is 3.63 indicating a similar directional preference (one-sided t-test that mean>3.5 also yields p=0.187).

level, the majority of students preferred CEEI to the Auction in six sessions and the majority never preferred the auction (one-sided binomial probability test that 6 out of 6 is 50%, $p = 0.016$), with 2 ties. The session-level aggregation makes sense to the extent that we think of each session as its own market and view majority rule as a social welfare criterion. Both the individual-level and session-level measures suggest that CEEI outperforms the Auction on this measure of allocative efficiency.

Table 3: Allocative Efficiency — Results from Binary Comparisons of CEEI Schedule and Auction Schedule

Session	Subjects in the Session	Prefer CEEI	Prefer Auction	Indifferent	Identical	Majority Voting Result
1	18	8	5	4	1	CEEI
2	15	5	5	2	3	Tie
3	19	10	4	3	2	CEEI
4	16	5	4	3	4	CEEI
5	18	8	8	2	0	Tie
6	14	6	5	1	2	CEEI
7	18	8	6	2	2	CEEI
8	14	6	5	0	3	CEEI
All	132	56	42	17	17	6-0-2

Table 3 shows allocative efficiency results from each session, reported in the order in which they were conducted. *Prefer CEEI* reports the number of subjects in the session who reported they preferred the CEEI schedule in both binary comparisons. *Prefer Auction* reports the number of subjects in the session who reported they preferred the Auction schedule in both binary comparisons. *Indifferent* reports the number of subjects whose preference for schedules reverses in the two binary comparison questions. *Identical* reports the number of subjects who received identical schedules under the two mechanisms and so did not see the binary comparison questions. *Majority Voting Result* asks which of the two mechanisms would be preferred by the majority of subjects if they were to vote for the mechanisms ex-post based on their preference over the schedules they received from each mechanism.

As we did for our fairness results, we now look for efficiency results based on the assumption that preference reporting under CEEI is perfectly accurate. We perform three different analyses.

First, we repeat the allocative efficiency and majority voting exercises from above but using the reported preferences instead of the binary comparisons to give a sense of the upper bound of CEEI performance under perfect reporting. At the individual level, 69% of students prefer their CEEI schedule to their Auction schedule based on the reported preferences (one-sided binomial probability test, $p<0.001$). At the session level, the majority of students prefer CEEI to Auction in seven sessions and there is one tie (one-sided binomial probability test, $p<0.01$).

Second, with reported preferences, we can compute the possible reduction of ex-post Pareto inefficiency that might arise if subjects were able to perfectly report their preferences. Specifically, we formulate an integer program that solves for the maximum number of Pareto-improving trades in each session given subjects' reported preferences and the initial allocation arrived at in the experiment.[32] CEEI is approximately Pareto efficient but there may be Pareto-improving trades because of the small amount of market-clearing error that is sometimes necessary to run CEEI;[33] the Auction is not Pareto efficient even approximately (cf. Sönmez and Ünver 2010). Table 4 reports the results of this exercise. As predicted by the theory, there is substantially less scope for Pareto-improving trades under CEEI than under the Auction.

Table 4: Results on Pareto Efficiency: Reported Preferences

	Auction	CEEI	Probability Ratio Test (one-sided)
# of Pareto-improving trades detected (% of course seats)	260 (32.8%)	44 (5.6%)	$p < 0.001$
# of students involved in at least one trade (% of students)	98 (74.2%)	22 (16.7%)	$p < 0.001$

Table 4 reports the results of an integer program that solves for the maximum number of Pareto-improving trades in each session based on subjects' reported preferences.

Third, if we assume interpersonal comparability of utilities, we can look directly at the magnitudes of subjects' utility changes between mechanisms under the assumption of perfect preference reporting. We do this in two ways. First, we look at each subject and calculate the percentage difference in utility between their realized schedule from the

[32] We restrict attention to trades in which each subject in the trade gives and gets a single course seat. A subject may engage in an unlimited number of trades, and a trade may involve arbitrarily many subjects. An additional fictitious played called the "registrar" holds all unused capacity and has zero utility from each course.

[33] Budish (2011) shows that there need not exist prices that exactly clear the market, but guarantees existence of prices that clear the market to within a small amount of approximation error. In Budish (2011), error is defined as the square root of the sum of squares of excess demand errors (too many students assigned to a class) and excess supply errors (empty seats in a positively priced class). The Wharton Committee viewed excess demand errors as more costly than excess supply errors, and tuned the CEEI software accordingly for the experiment. Over the eight sessions, there were ten total seats of excess supply (median: one seat per session) and two total seats of excess demand (median: zero seats per session). The Pareto-improving trades exercise reported in the text treats the registrar as owning the ten seats of excess supply and ignores the two seats of excess demand. In the practical implementation of CEEI at Wharton, we modified the mechanism in a small way to entirely prevent excess demand errors that cause violations of strict capacity constraints (e.g., due to fire codes). See Budish, Cachon, Kessler and Othman (forthcoming).

Auction and their realized schedule from CEEI. This measure is depicted below as Figure 4.

Figure 4: Distribution of Change in Utility Going from Auction to CEEI

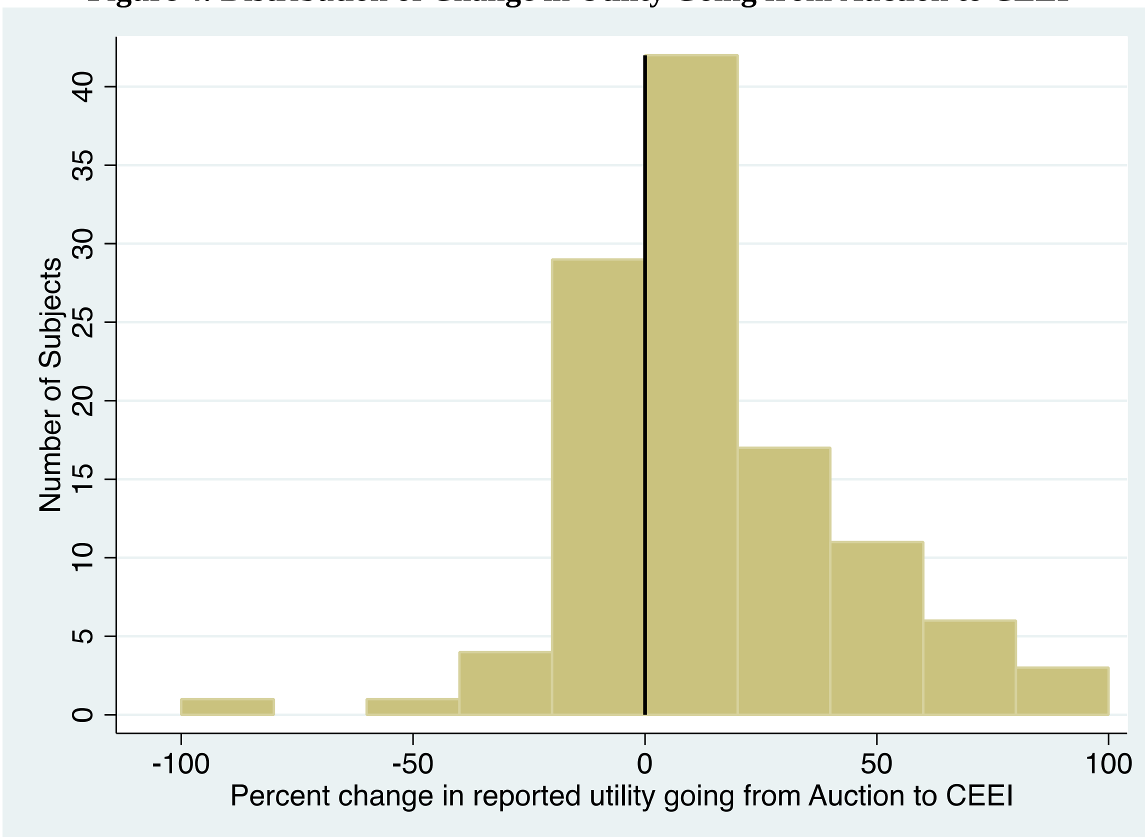

Figure 4 shows a histogram of the percentage change in utility going from the schedule received in the Auction to the schedule received in CEEI. When necessary, we made non-legal Auction schedules (i.e. schedules that had too many courses or courses that failed to satisfy scheduling constraints) legal by maximizing the subject's utility among courses in the Auction schedule subject to scheduling constraints. Bins are 20 percentage points wide and the graph excludes the 18 subjects who got the same utility from both schedules. One observation had a utility increase of over 100% and is included in the highest percentage increase bar.

The majority of mass is to the right of 0 in Figure 4, a visual confirmation of the fact noted above that 69% of students (79 out of 114) prefer their CEEI schedule to their Auction schedule based on reported preferences. Moreover, the winners win more than the losers lose: thirty-seven students have at least a 20% utility improvement when switching from the Auction to CEEI, whereas only six students have at least 20% utility harm from switching from the Auction to CEEI.

Second, in Figure 5 we plot the distribution of utilities from schedules coming from the Auction and coming from CEEI. The distribution of utilities under CEEI second-order stochastically dominates the distribution under the Auction. This implies that under perfect preference reporting a utilitarian social planner prefers the distribution of outcomes under CEEI to that under the Auction, so long as the planner has a weak preference for equality (the social welfare analogue of risk-aversion). However, the right tail of outcomes under the Auction is better than that under CEEI, so we do not obtain first-order stochastic dominance.

Figure 5: Distribution of Utility Under CEEI and the Auction, Based on Reported Preferences

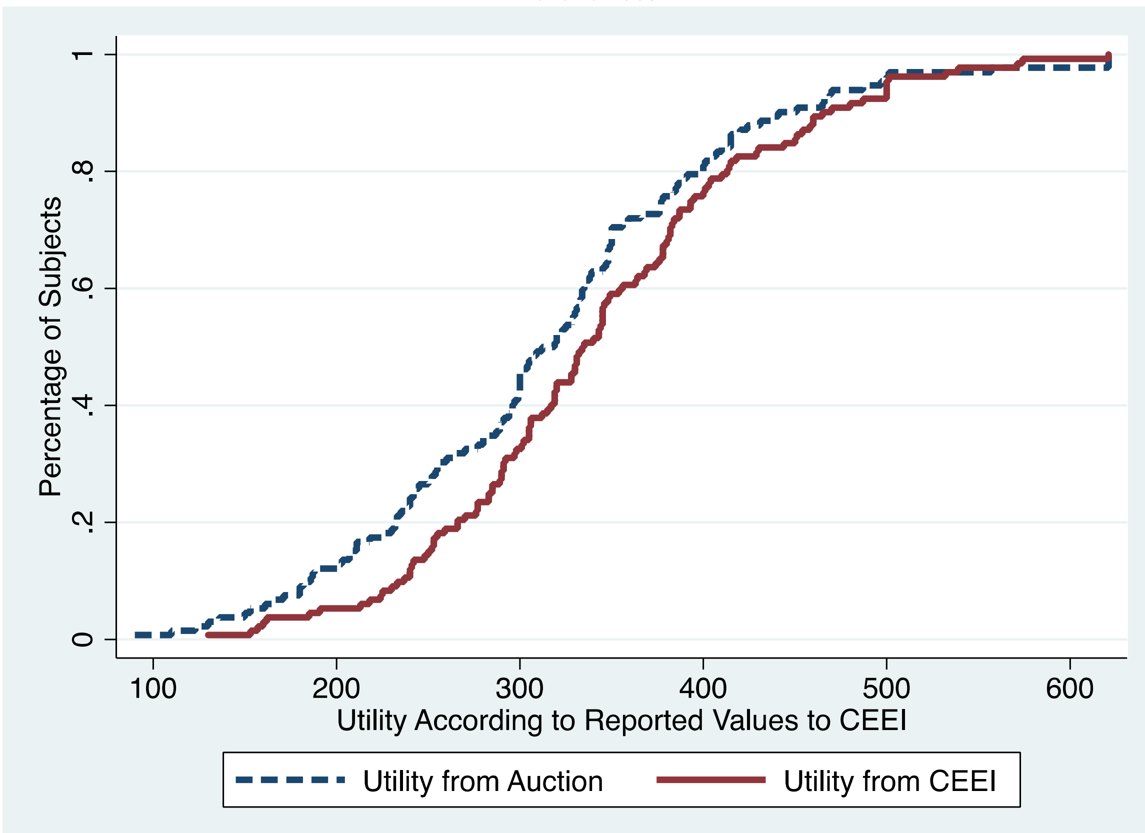

Figure 5 plots the CDF of utility according to reported values to CEEI for both the Auction and CEEI. Three utilities (two in the Auction and one in CEEI) are above 2,000 and have been Winsorized at 621, the next-highest utility value.

3.3 Slack Analysis

In consultation with the Wharton Committee, we recruited students with the aim of obtaining 20 subjects per session. Our turnout was worse than forecast, especially in the evening sessions. As a result, the number of subjects per session ranged from 14 to 19.

This variation in attendance inadvertently generated variation in what we term "slack," defined as $100\ \times \left(\frac{\#seats\ supplied}{\#seats\ demanded} - 1\right)$. "Slack" is the percentage of excess capacity in the available courses times 100, and it ranged from 12.6 to 29.3 in our sessions.[34]

If there is too much slack in a market, then the allocation problem is trivial: under virtually any course allocation system all students will get exactly what they want (or something close to it). Thus, in sessions with a relatively large amount of slack, we might expect that CEEI and the Auction would do equally well, whereas in sessions with a relatively small amount of slack, we may expect the relative advantages of CEEI to be more pronounced.

Table 5 presents regression results that analyze how the level of slack in our sessions affects whether subjects were more likely to display envy under CEEI or the Auction in their binary comparisons and whether subjects preferred their schedule from CEEI to their schedule from the Auction in the binary comparisons. This analysis was meaningful to the decision makers at Wharton because the slack in their real-world problem is in the range of 15 to 20, slightly less than the average amount of slack in our experiment. The constant in the regressions shows the percentage of subjects who displayed less envy under CEEI than under the Auction and the percentage of subjects who prefer their CEEI schedule to their Auction schedule under the average amount of slack in the experiment. The coefficient of *Slack – Mean(Slack)* shows the effect on these measures of increasing slack relative to the average amount of slack.

The significant effect of slack on the likelihood that subjects show less envy under CEEI and on the preference of the CEEI schedule demonstrates that as slack decreases (i.e., the allocation problems get harder), the efficiency and fairness benefits of CEEI over the Auction become more pronounced. The results suggest that for each fewer point of slack (i.e., percentage point of excess seats), the percentage of subjects who experience less envy under CEEI than the Auction increases by 0.696 percentage points. Similarly, for each fewer point of slack, the percentage of subjects who prefer their CEEI schedule to their Auction schedule increases by 0.918 points.

[34] We adjusted the number of seats available in the experimental session based on the number of subjects who participated, but in a coarse way. For example, if fewer than eighteen students showed up to a session, then all course sections with five seats would be adjusted to have four seats. This allowed us to implement and communicate changes to the number of seats even though subjects had printed instructions with the number of seats for each course. This also means that our measure of slack is not perfectly correlated with the number of subjects in the session.

Table 5: Effect of Slack on Relative Performance of CEEI vs. Auction

	Fairness		*Allocative Efficiency*
	Less Envy CEEI	Less Strong Envy CEEI	Prefer CEEI
	(1)	(2)	(3)
Constant	55.47 (3.54)***	56.50 (2.99)***	55.30 (1.21)***
Slack – Mean(Slack)	-0.696 (0.235)**	-1.35 (0.218)***	-0.918 (0.305)**
Observations	115	115	132
Clusters	8	8	8

Table 5 reports OLS regressions. *Slack* is defined as $100 \times \left(\frac{\#seats\ supplied}{\#seats\ demanded} - 1\right)$. It ranges from 12.6 to 29.3 across the eight sessions, and *Mean(Slack)* = 20. *Less Envy CEEI* = 100 if the subject displayed envy under the Auction but not under CEEI; *Less Envy CEEI* = 50 if the subject displayed envy under both or no envy under both mechanisms; *Less Envy CEEI* = 0 if the subject displayed envy under CEEI but not under the Auction. *Less Strong Envy CEEI* is the same as *Less Envy CEEI* but only counts subjects who report they “prefer” or “strongly prefer” another subject’s realized schedule. The envy results are restricted to the 115 subjects who saw at least one envy binary comparison under each mechanism. *Prefer CEEI* = 100 if the subject preferred their CEEI schedule to their Auction schedule both times they were asked; *Prefer CEEI* = 50 if the subject is indifferent (i.e., switched preferences between schedules) or got an identical schedule under both systems; *Prefer CEEI* = 0 if the subject preferred their Auction schedule to their CEEI schedule both times they were asked. Robust standard errors, clustered by session, are in parentheses. *, **, and *** indicate significance at 0.1, 0.05, and 0.01 respectively in two-sided tests.

3.4 Discussion of Results on Fairness and Efficiency

Subjects report their preferences “accurately enough” that CEEI outperforms the Auction on every measure of efficiency and fairness, with most comparisons statistically significant at the 5% level. The improvements generated by CEEI are especially large on the “hard” allocation problems with low slack. However, CEEI outperforms by much less based on the binary comparisons than based on the reported preferences. The percentage of students exhibiting envy in CEEI is 31% based on the binary comparisons versus just 4% based on the preference reports; the proportion of students preferring their CEEI schedule to their Auction schedule is 57% based on the binary comparisons versus 69% based on the preference reports. These large differences indicate that difficulty with preference reporting was an important factor in mechanism performance. We turn to this subject in the next section.

IV. Difficulty with Preference Reporting

In this section, we investigate why subjects failed to report their preferences perfectly.

Conceptually, there are two possible reasons why agents' preference reports might not reflect their underlying true preferences, i.e., why the "agents report their type" assumption might fail. First, agents may have difficulty using the preference reporting language we provided in the lab to express their underlying true preferences, even though in principle it is mathematically feasible to express their underlying true preferences using the language. Second, there are some kinds of preferences that mathematically cannot be expressed using the language we provided in the lab, which if present in our subject pool necessarily creates a discrepancy between subjects' reported preferences and their true preferences.[35] We present summary statistics on the overall prevalence of mistakes in Section 4.1 and then investigate each of these two sources of mistakes in turn in Sections 4.2 and 4.3. Section 4.4 discusses the results.

4.1 Summary Statistics

Every binary comparison is a test of our preference reporting language. We say a binary comparison is *consistent* if the subject's choice is consistent with their reported preferences, and otherwise is a *contradiction*. Table 6 presents summary statistics on the prevalence and magnitude of contradictions.[36] A higher percentage of contradictions suggests that the preference reporting language was less able to capture subjects' true preferences.

A few observations can be made from the pattern of data in Table 6. First, there are a significant number of contradictions. Overall, 84.4% of binary comparisons were consistent, with 15.6% contradictions (76.4% of subjects have at least one contradiction). Second, there are very few contradictions in the bottom right of the table (i.e., when preference reports assign a big utility difference between the two schedules and the binary comparison indicates that the schedule with the lower utility is "strongly preferred"), suggesting that there are few "big" contradictions. In general, as we move down rows in Table 6, the data shift to the left, meaning that the preference reports are more likely to pick the preferred schedule and contradictions are more likely to be associated with a weak preference. Of the 596 comparisons when the utility difference is 100 or more, the preference reports contradict the binary comparison responses only 7.05% of the time (and only 1.85% of cases are contradictions in which subjects report a strong preference for the disfavored schedule). In contrast, in the 123 comparisons in

[35] Returning to the analogy of footnote 8: if one's true underlying preferences are in English, and the preference reporting language is Latin, the first issue is that translating from English to Latin requires mastery of both English and Latin and skill at translation, whereas the second issue is that there are some concepts and ideas that can be expressed in English that cannot be fully expressed in Latin.

[36] Table A4 in Appendix F provides summary statistics on the use of the preference reporting language.

which the utility difference between the schedules based on the preference reports is less than 10, 29.27% of cases are contradictions.

Table 6: Prevalence and Magnitude of Preference Reporting Contradictions

Utility Difference Between Schedules	# Comparisons with this Utility Difference	Consistent	Contradictions			
			All	Weak Preference	Preference	Strong Preference
All	1,662	84.42%	15.58%	5.17%	6.92%	3.49%
1-9	123	70.73%	29.27%	9.76%	11.38%	8.13%
10-49	516	77.13%	22.87%	6.78%	11.24%	4.84%
50-99	427	85.25%	14.75%	6.32%	5.62%	2.81%
100+	596	92.95%	7.05%	2.01%	3.19%	1.85%

Table 6 shows the percentage of binary comparisons that were contradictions. For each binary comparison, the *Utility Difference Between Schedules* is the utility of the schedule with the higher utility minus the utility of the schedule with the lower utility, as determined by the subject's preference reports under CEEI. The table shows all 1,662 comparisons where this number is greater than 0 and so the preference reports suggest one schedule is preferred to the other. The *Consistent* column reports the percentage of these comparisons where the binary comparison choice confirms the preference report prediction. The *Contradictions* columns report the percentage of binary comparisons that contradicted the CEEI preference reports overall and at each level of preference.

4.2 Difficulty Using the Preference Reporting Language

To assess whether agents had difficulty using the preference reporting language we provided in the lab, we explore whether they were able to effectively use each of the components of the reporting language: cardinal values to express preferences for individual courses and pairwise adjustments to express certain kinds of substitutabilities and complementarities 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. We say that a binary comparison between schedules A and B is an *ordinal comparison* if the subject's reported preferences generates an unambiguous predicted preference between A and B based on ordinal information alone. For example, if A consists of the subject's {1^{st}, 3^{rd}, 5^{th}, 7^{th}, 9^{th}} reported favorite courses, B consists of the subject's {2^{nd}, 4^{th}, 6^{th}, 8^{th}, 10^{th}} reported favorite courses, and neither schedule triggers adjustments, then we can conclude that the subject has reported a preference for schedule A over schedule B without knowing the specific cardinal utilities the student assigned to each course. When one schedule can be determined to be preferred to the other based on ordinal information alone, we say that schedule "rank dominates" the other schedule.

We define a comparison between schedules A and B as a *cardinal comparison* if neither schedule triggers an adjustment and neither schedule rank dominates the other. For example, if schedule A consists of a subject's {1^{st}, 2^{nd}, 8^{th}, 9^{th}, 10^{th}} reported favorite courses and schedule B consists of a subject's {3^{rd}, 4^{th}, 5^{th}, 6^{th}, 7^{th}} reported favorite courses, ordinal information alone is insufficient to determine which is preferred. These are the comparisons for which the subject's ability to report cardinal preference information accurately is put to the test.

Table 7 summarizes the prevalence and magnitude of preference reporting consistencies and contradictions as a function of whether the comparison in question is an ordinal comparison or a cardinal comparison. The table shows that contradictions are much more common in the case of cardinal comparisons (31.72%) than in ordinal comparisons (10.94%), a nearly threefold increase. This difference is highly statistically significant (probability ratio test, two-sided, $p<0.001$) and is robust to controls for the difference in utility between the two schedules.[37]

Table 7: Prevalence and Magnitude of Preference Reporting Contradictions for Ordinal and Cardinal Comparisons

Type of Comparison	# Comparisons with this Utility Difference	Consistent	Contradictions			
			All	Weak Preference	Preference	Strong Preference
Ordinal	1,207	89.06%	10.94%	4.06%	4.39%	2.49%
Cardinal	373	68.36%	31.64%	9.38%	15.28%	6.97%

Table 7 shows all 1,580 binary comparisons in which neither schedule triggered an adjustment. A comparison is *Ordinal* if the CEEI preference report predicts which schedule is preferred based on ordinal information alone. Otherwise the comparison is *Cardinal*, because cardinal preference intensity information is necessary to determine which schedule is preferred.

[37] The rank dominant comparisons have larger utility differences than the cardinal comparisons (99.4 versus 82.7), but controlling for the difference in utility, we still observe that comparisons are significantly more likely to be a contradiction if they rely on cardinal information rather than just ordinal information. A linear regression that controls non-parametrically for the utility difference shows that the cardinal comparisons are 16.1 percentage points (i.e., nearly two-and-a-half times) more likely to be a contradiction ($p<0.001$).

These results suggest that subjects had significant difficulty reporting the intensity of their preferences for individual courses. The ordinal component of preferences for individual courses was substantially more accurate than the cardinal component.

Next, 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 (see details in Appendix Table A4). We ask whether, in the cases where adjustments were used, they enhanced or detracted from reporting accuracy.

Table 8 summarizes the prevalence and magnitude of preference reporting contradictions as a function of whether the comparison activated an adjustment. Due to the relatively limited use of adjustments, only 82 of the binary comparisons involved a schedule in which one or more adjustments were activated for the subject. That said, in these 82 cases, only 10.98% yielded preference reporting contradictions versus 15.82% for the comparisons that did not involve an adjustment (probability ratio test, two-sided, p=0.239). The relatively low rate of contradictions in the 82 cases when adjustments were activated suggests that adjustments did not detract from preference reporting accuracy, and may have slightly enhanced it (though the difference is not statistically significant).[38]

Table 8: Prevalence and Magnitude of Preference Reporting Contradictions for Comparisons with and without Adjustments

Type of Comparison	# Comparisons with this Utility Difference	Consistent	Contradictions			
			All	Weak Preference	Preference	Strong Preference
No Adjustment	1,580	84.18%	15.82%	5.32%	6.96%	3.54%
Adjustment	82	89.02%	10.98%	2.42%	6.10%	2.42%

Table 8 shows all 1,662 comparisons. *Adjustment* indicates that one of the schedules in the comparison activated an adjustment in the CEEI preference reports. *No Adjustment* indicates that neither schedule activated an adjustment.

4.3 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

[38] The success of those using adjustments could be driven by selection, although we find no difference in the rate of contradictions between those subjects who report adjustments and those who do not. See Table A5 in the Appendix.

that subjects would be mathematically unable to express using the language that was provided. The issue is that many kinds of non-additive preferences cannot be expressed using pairwise adjustments.[39] Additionally, there are many kinds of non-additive preferences that in principle could be expressed using the language, but for which the language does not seem especially natural.[40]

The set of potential non-expressible preferences is vast, and we do not have a disciplined way of exploring all such possibilities as a source of preference reporting contradictions.[41] Instead, we look at two specific sources of non-additive 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 the 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 he 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 supplied preference language. We therefore investigate whether there are more contradictions when the schedule we expect a subject likes less based on the preference reports has one of these elegant features (and thus should get a utility bump that is unreported).

Table 9 is broken up into two panels, one for each of the two features: whether the schedule is balanced (Panel A), and whether the schedule is contiguous (Panel B). The table summarizes the prevalence and magnitude of preference reporting contradictions as

[39] 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, we and the committee concluded that arbitrary set-wise adjustments would be too complicated for students. How best to trade-off the expressiveness of a preference reporting language and agents' ability to use the language is an interesting open area for research, as we discuss further in the conclusion.

[40] 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 $k(k-1)/2$ 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.

[41] With roughly 50,000 possible schedules in the lab, there are 50,000! possible ordinal preferences over schedules, or roughly $10^{12,499}$. As such, the up to nineteen binary comparisons we ask of subjects does not provide enough data to identify patterns in such a large set without prior guidance on where to look.

a function of which of the schedules in the binary comparison had the elegant schedule feature.

Results from Panel A show that subjects' binary comparison responses are more likely to contradict their reported preferences when the schedule that their reported preferences predict they disfavor is balanced. Subjects are more likely to make a contradiction when the schedule their reported preferences predict they disfavor is balanced and the other is not (29.41% are contradictions) than when both or neither are balanced (15.21% are contradictions; probability ratio test, two-sided, p<0.01) or when the schedule their reported preferences predict they favor is balanced and the other one is not (13.60% are contradictions; probability ratio test, two-sided, p=0.036).

Table 9: Prevalence and Magnitude of Preference Reporting Contradictions for Comparisons with and without Elegant Schedules

Type of Comparison	# Comparisons with this Utility Difference	Consistent	Contradictions			
			All	Weak Preference	Preference	Strong Preference
Panel A: Balanced Schedule						
Only higher rated has it	66	86.40%	13.60%	3.03%	6.06%	4.55%
Only lower rated has it	51	70.59%	29.41%	7.84%	15.69%	5.88%
Neither or both have it	1,545	84.79%	15.21%	5.18%	6.67%	3.37%
Panel B: Contiguous Schedule						
Only higher rated has it	199	87.44%	12.56%	4.52%	4.02%	4.02%
Only lower rated has it	192	81.77%	18.23%	7.29%	7.81%	3.12%
Neither or both have it	1,271	84.34%	15.66%	4.96%	7.24%	3.46%

Table 9 shows the 1,662 comparisons and splits them based on whether each of the schedules in the comparison was balanced (Panel A) and whether each of the schedules in the comparison was contiguous (Panel B). *Type of Comparison* indicates which schedule(s) in the comparison had the elegant feature. The "higher rated" and "lower rated" labels refer to the schedule the CEEI preference reports predicts to be favored and disfavored, respectively.

Panel B depicts a similar pattern when looking at whether a schedule is contiguous. Subjects are directionally, but not significantly, more likely to make a contradiction when the schedule their reported preferences predict they disfavor is contiguous and the other is not (18.23% are contradictions) as compared to when both or neither are contiguous (15.66% are contradictions, probability ratio test, two-sided, p=0.365) or when the schedule their reported preferences predict they favor is contiguous and the other one is not (12.56% are contradictions, probability ratio test, two-sided, p=0.120).

That subjects are more likely to make a contradiction when their reported preferences predict they favor a schedule that is not balanced or not contiguous suggests that at least some of the contradictions are due to the preference reporting language failing to provide a way for agents to report important features of their preferences. An important caveat is that each of these specific types of non-expressible preferences accounts for only a small number of contradictions each; there are likely other non-expressible preferences that we do not quantify here.

4.4 Discussion

This section has reported three sets of results about preference reporting. First, the preference reports convey significant information about subjects' preferences, with about 85% of binary comparisons consistent with the preference reports and large mistakes comparatively rare. Subjects are able to report their types relatively accurately. At the same time, however, preference reporting difficulties are prevalent, with about 15% of binary comparisons contradictory and with over three-quarters of subjects exhibiting at least one contradiction, which, as shown in Section 3, demonstrably harmed mechanism performance. Second, subjects especially had difficulty with reporting cardinal preference intensity information, being about three times more likely to have a contradiction in binary comparisons involving cardinal information relative to those just involving ordinal information. Third, when subjects did express non-additive preferences they did so with reasonable accuracy, but they did so rarely and the evidence suggests that there were some forms of non-additive preferences that were important to the subjects but that they were unable to express with the tools 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 non-additive preferences (Cantillon and Pesendorfer 2006, Reguant 2014), and the overall importance of language design (Milgrom 2009, 2011). We also hope that the overall logic of the results gives the reader additional comfort as to the validity of the experimental methodology.

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 lab, based on the overall level of accuracy of the reports taking into consideration that subjects had only ten 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, since this was such an important source of difficulty in the lab. Third, Wharton enhanced the top-ten widget in the preference reporting user interface to allow students to see substantially more than ten 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.[42] To date, Wharton has opted not to incorporate other ways to report non-additive preferences beyond the pairwise adjustment tool, fearing excessive complexity. Developing a conceptual understanding of the tradeoff between expressiveness and complexity is an interesting open area for future research.

V. Qualitative Analysis

As noted in the introduction, an additional advantage of using real market participants as experimental subjects is that we could search for "side effects" — issues not captured by the theory in Budish (2011) that could undermine the potential benefits of CEEI. For example, a mechanism might have great theoretical properties, but if participants find it frustrating or confusing they may rebel against using it. Concern about side effects was especially pronounced in our setting because the CEEI mechanism had never been used before and was complex in several ways. Additionally, fear that a new market design could lead to disgruntled participants was especially pronounced at Wharton, where student satisfaction is a top priority — in this case, the Wharton administration was concerned about satisfaction both with regard to the final allocation and to the process that lead to that allocation.

To address these concerns, our experiment looked for such "side effects" of the new mechanism by collecting a wide variety of quantitative survey data on subjects' attitudes towards both mechanisms. Subjects also had the opportunity to provide free-response comments. This survey data played a perhaps-surprisingly prominent role in the Wharton Committee's evaluation of CEEI and also suggested some improvements for implementation.

[42] In the free-response comments, several students specifically mentioned the top-ten widget as a helpful feature of the user interface.

In total, we asked 15 qualitative Likert-scale questions about the mechanisms. The seven that found a significant difference between the mechanisms are shown in Table 10 and discussed in the following sections. The complete list of questions is presented in Table A6 in the Appendix. The responses to these survey questions tell a story about the qualitative benefits of CEEI relative to the Auction and suggest an area for improvement in practical implementation. We address the relative benefits of CEEI in Section 5.1 and then address the potential improvement in Section 5.2. In Section 5.3 we investigate how preference for CEEI interacted with subject demographics.

Table 10: Qualitative Responses About CEEI and Auction

Questions about each System	CEEI Average	Auction Average	Wilcoxon p-value
Panel A: Questions Regarding Strategic Simplicity and Overall Satisfaction			
I enjoyed participating in this course allocation system.	4.72	4.37	$p = 0.095$
I like this course allocation system.	4.55	4.18	$p = 0.095$
This course allocation system is simple.	4.45	3.73	$p = 0.001$
I had to think strategically about what other students would do in this course allocation system.	2.93	6.42	$p < 0.001$
Someone with perfect knowledge of the historical supply and demand for courses could have had an advantage over me in this system.	3.67	6.04	$p < 0.001$
Panel B: Questions Regarding Transparency and Understanding			
I understand how this course allocation system works.	4.83	5.92	$p < 0.001$
I felt like I had control over my schedule in this course allocation system.	3.95	4.45	$p = 0.073$

Table 10 shows the seven qualitative questions that resulted in statistically significant differences between the two mechanisms. These seven questions are divided into two panels to facilitate discussion in the text of Sections 5.1-5.2. The other eight questions not listed yielded no significant differences with $p > 0.1$. All survey questions are listed in Appendix G. Questions were rated on a scale of: 1="Strongly Disagree," 2="Disagree," 3= "Somewhat Disagree," 4="Neither Agree or Disagree," 5="Somewhat Agree," 6="Agree," 7="Strongly Agree." *CEEI Average* and *Auction Average* take the mean of the response values across all 132 subjects in the experiment. Since each subject gave an answer for each of the mechanisms, we can use a non-parametric Wilcoxon sign-rank test that responses are equal across the two mechanisms.

5.1 Strategic Simplicity and Student Satisfaction

The area of the survey with the largest difference between CEEI and the Auction concerned strategic simplicity. The average response to the question that asked subjects' agreement with the statement "I had to think strategically about what other students would do in this course allocation system" was 2.93 for CEEI (i.e., close to "Somewhat Disagree," a 3 on the Likert scale) and was 6.42 for the Auction (i.e., close to the midpoint between "Agree" and "Strongly Agree"). Similarly, the average response to the question: "Someone with perfect knowledge of the historical supply and demand for courses could have had an advantage over me in this system" was 3.67 for CEEI and was 6.04 for the Auction. These results suggest that subjects broadly understood that the CEEI mechanism, unlike the Auction, does not require strategizing.[43] The question "This course allocation system is simple" also significantly favored CEEI over the Auction but with a smaller magnitude (4.45 vs. 3.73, Wilcoxon sign-rank test, p=0.001).[44]

Overall satisfaction was the other area of the survey in which CEEI outperformed the Auction, though with smaller magnitudes and with significance only at p<0.1. Subjects were more likely to agree with "I like this course allocation system" for CEEI than for the Auction (4.55 vs. 4.18, Wilcoxon sign-rank test, p=0.095) and more likely to agree with "I enjoyed participating in this course allocation system" for CEEI than the Auction (4.72 vs. 4.37, Wilcoxon sign-rank test, p=0.095). These results helped convince the Wharton administration that there was nothing unexpected about the CEEI mechanism that led subjects to dislike the system; that is, there was no unanticipated "side effect" that made CEEI unappealing to the Wharton student subjects.

5.2 CEEI is a Black Box

The two questions on the survey on which CEEI performed significantly worse than the Auction were "I understand how this course allocation system works" (4.83 for CEEI vs. 5.92 for Auction) and "I felt like I had control over my schedule in this course allocation

[43] One might be somewhat surprised that the difference between CEEI and the Auction on these measures is not even larger. One explanation is that at least some of our subjects were reluctant to accept or did not understand that the CEEI mechanism was not "gameable" like the Auction was (cf. Hassidim, Romm and Shorrer 2015, Li 2015). One lesson for implementation that came out of these survey responses was to do a more thorough job of explaining this fact to students, since understanding that historical information and strategizing was not necessary for CEEI was positively correlated with other measures of satisfaction with CEEI. For more discussion of the benefits of strategy-proofness in market design see, e.g., Pathak and Sönmez (2008, 2013), Roth (2008), Azevedo and Budish (2015) and Li (2015).

[44] The strategic simplicity theme also arose in many subjects' free responses. One subject wrote, "Multiple rounds and the historical price research introduced in the auction make it quite time-consuming and therefore kind of annoying." Another wrote, "Really like the idea of the new system as it removes the inherent 'gaming' aspect of the auction – I'm a believer in free markets but the auction is a disadvantage to those that don't have the time or skill required to fully research and intelligently participate in it."

system” (3.95 for CEEI vs. 4.45 for Auction). Our interpretation of these results is that subjects felt that CEEI was a bit of a “black box”, i.e., non-transparent.[45]

These findings have helped guide implementation at Wharton in two ways. First, Wharton administrators did student-wide presentations about the new mechanism to explain in detail how it works; the presentation also covered the theory behind the mechanism and the experimental evidence in support of its use at Wharton to enhance transparency. Second, in the practical implementation, Wharton made a simple change to the mechanism’s user interface. In the user interface in the lab, subjects were shown the schedule they received under CEEI but were not shown market-clearing prices. This prevented subjects from understanding why they got the specific schedule they got, and why, for example, they failed to get some particular course they valued highly. In the practical implementation, students are shown the market clearing prices.[46]

5.3 Gender

One additional set of results that arose from the survey data was on the relative preferences between CEEI and the Auction for men and women. This result turned out to be important for the Wharton administration, which was facing evidence that women at Wharton disproportionately disliked the 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 Auction than men did (7-point scale of effectiveness, 4.95 for men vs. 4.28 for women, t-test, two-sided, $p<0.001$).

We found a similar pattern in our data with regard to attitudes toward the Auction. In our Likert-scale questions, female subjects reported liking the Auction significantly less than male subjects reported liking it (4.51 for men vs. 3.81 for women; Wilcoxon rank-sum test, $p=0.032$). For CEEI, however, the gender gap disappears (4.56 for men vs. 4.53 for women; Wilcoxon rank-sum test, $p=0.854$).[47]

To our knowledge, this is the first evidence documenting a gender gap in liking of market designs. If we interpret the Auction as “competitive” because it is highly strategic and CEEI as “noncompetitive” because it is approximately strategy-proof, the finding echoes a famous finding in the gender literature (Niederle and Vesterlund 2007).

[45] Subjects’ free responses raised this issue directly. One subject wrote: “I like the idea of getting the best schedule I could afford, but didn’t like feeling like I wasn’t in control. I would feel helpless if I got a schedule that wasn’t close to what I preferred.” Another wrote: “The course matching system is just a black box where there’s one round and we rely on the computer to make judgments for us.”

[46] Gérard Cachon, the chair of Wharton’s Course Allocation Redesign Team, writes 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.’”

[47] Women reported liking CEEI significantly more than the Auction (4.53 for CEEI vs. 3.81 for the Auction; Wilcoxon sign-rank test, $p=0.027$) but there is not a significant interaction comparing the Auction to CEEI between men and women because men also like CEEI slightly more than they like the Auction.

VI. Conclusion

Wharton formally decided to adopt CEEI for use in practice after a series of administrative meetings in the few months following the experiment. This could not have been an easy decision given the complexity of the 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 CEEI was not any one experimental result but rather the full set of experimental results: the efficiency and fairness gains relative to the auction; 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 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.[48]

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.[49] However, the limited data that are available are all consistent with the claims made by the theory and 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 not getting popular courses reflects unfairness of the mechanism or their preferences, but the aggregate distribution suggests that CEEI improved equity. In the last fall of the Auction, 32% of students got zero of the top twenty most popular courses and 5% got three or more, versus 13% and 0%, respectively, under CEEI. That is, under 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 Gini index of this distribution went from 0.54 in the last fall of the Auction to 0.32 in the first fall of CEEI.[50] In addition, we used school-wide surveys to investigate the change in mechanisms. At our urging, the annual administration survey of the student

[48] The following is an excerpt from the "Course Match User Manual" provided to students: "In the Fall of 2011, Wharton faculty and staff joined with 132 MBA students and put the Course Match theories to the test…The results were clear. Students were more satisfied with their Course Match schedules than with those generated by the Auction. They were less envious of their peers' schedules and they found Course Match easier to use even though they received only minimal training on the new system."

[49] Ideally, we would have used a school-wide survey to obtain true preference from students during the last year of the Auction; this would have allowed us to compare student outcomes from actual play of the Auction to counterfactual play of CEEI, analogously to the study conducted by Budish and Cantillon (2012). 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 Auction would have been confusing — especially given that a school-wide announcement had been made concerning the adoption of the new, truthful mechanism. Due to the complexity of the equilibrium of the Auction, it is an open question whether it is possible to infer true preferences from strategic play in the absence of such a survey.

[50] For further details on these data and other details regarding the practical implementation see Budish, Cachon, Kessler and Othman (forthcoming).

body added a few questions about course allocation in the last year of the Auction's use, written in such a way that they could be used again in the first year of CEEI 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 Auction) to 64% in 2014 (the first year of CEEI). 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 lab and adopted for implementation is "accurate enough" to yield efficiency and fairness benefits from CEEI, but the results do not at all suggest that the language is optimal. One specific direction to consider based on the results from the experiment would be to allow students to report richer kinds of non-additive preferences. A more difficult conceptual question is how to think about the overall tradeoff 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.[51] Too complicated of a language would also be sub-optimal, 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, e.g., 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 (i.e. about the temporal structure of schedules as described in Section 4.3) that students had difficulty reporting suggests the potential for advancement on this front.

[51] A real-world 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 (forthcoming) 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.

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 it was a non-starter for us given the questions our experiment had to answer. We suspect that as market design continues to grow as a field — and as computers become more powerful and decision supports more sophisticated — market designs leveraging complicated preference information will become more common and many other market design researchers will find themselves in our shoes. Our greatest hope for this paper is that other market design researchers can build on the example here and help bring other useful market designs from theory to practice.

VII. References

Abdulkadiroğlu, Atila and Tayfun Sönmez. 2003. "School Choice: A Mechanism Design Approach." *The American Economic Review* 93 (3): 729-747.

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

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

Abdulkadiroğlu, Atila, Parag Pathak, Alvin E. Roth and Tayfun Sönmez. 2006. "Changing the Boston School Choice Mechanism." Working Paper.

Arnsperger, Christian. 1994. "Envy-Freeness and Distributive Justice." *Journal of Economic Surveys*, 8(2): 155-186.

Ausubel, Lawrence M., Peter Cramton, and Paul Milgrom. 2006. "The Clock-Proxy Auction: A Practical Combinatorial Auction Design." In Peter Cramton et al., eds. *Combinatorial Auctions*, MIT Press, 212-259.

Azevedo, Eduardo and Eric Budish. 2015. "Strategy-proofness in the Large." Working Paper.

Bergemann, Dirk and Stephen Morris. 2005. "Robust Mechanism Design." *Econometrica*, 73(6), 1771-1813.

Bogomolnaia, Anna and Hervé Moulin. 2001. "A New Solution to the Random Assignment Problem." *Journal of Economic Theory*, 100: 295-328.

Budish, Eric. 2011. "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes." *Journal of Political Economy*, 119(6): 1061-1103.

Budish, Eric and Estelle Cantillon. 2012. "The Multi-Unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard." *American Economic Review*, 102(5): 2237-2271.

Budish, Eric, Gérard Cachon, Judd Kessler, and Abraham Othman. Forthcoming. "Course Match: A Large-Scale Implementation of Approximate Competitive Equilibrium from Equal Incomes for Combinatorial Allocation." *Operations Research*.

Calsamiglia, Caterina, Guillaume Haeringer, and Flip Klijn. 2010. "Constrained School Choice: An Experimental Study." *American Economic Review*, 100(4): 1860-1874.

Cantillon, Estelle and Martin Pesendorfer. 2006. "Combination Bidding in Multi-Unit Auctions." Working Paper.

Chamberlin, Edward H. 1948. "An Experimental Imperfect Market." *Journal of Political Economy*, 56(2), 95-108.

Chen, Yan and Tayfun Sönmez. 2006. "School Choice: An Experimental Study." *Journal of Economic Theory*, 127(1), 202-231.

Cramton, Peter, Yoav Shoham, and Richard Steinberg (eds.). 2006. *Combinatorial Auctions*. Cambridge, MA: MIT Press.

Ehlers, Lars and Bettina Klaus. 2003. "Coalitional Strategy-Proof and Resource-Monotonic Solutions for Multiple Assignment Problems." *Social Choice and Welfare*, 21: 265–280.

Featherstone, Clayton and Eric Mayefsky. 2014. "Why Do Some Clearinghouses Yield Stable Outcomes? Experimental Evidence on Out-of-Equilibrium Truth-Telling." Working Paper.

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

Foley, Duncan. 1967. "Resource Allocation and the Public Sector." *Yale Economic Essays*, 7: 45-98.

Fragiadakis, Daniel E. and Peter Troyan. 2014. "Improving Welfare in Assignment Problems: An Experimental Investigation." Working Paper.

Fudenberg, Drew and Jean Tirole. 1991. *Game Theory*. Cambridge, Mass.: MIT Press.

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

Goeree, Jacob K. and Charles A. Holt. 2010. "Hierarchical Package Bidding: A Paper & Pencil Combinatorial Auction." *Games and Economic Behavior*, 70, 146-169.

Hassidim, Avinatan, Assaf Romm, and Ran I. Shorrer. 2016. "'Strategic' Behavior in a Strategy-Proof Environment." Working Paper.

Hatfield, John William. 2009. "Strategy-Proof, Efficient, and Nonbossy Quota Allocations." *Social Choice and Welfare*, 33 (3): 505–515.

Hatfield, John William and Scott Duke Kominers. Forthcoming. "Contract Design and Stability in Many-to-Many Matching." *Games and Economic Behavior*.

Kagel, John H., Yuanchuan Lien, and Paul Milgrom. 2010. "Ascending Prices and Package Bidding: A Theoretical and Experimental Analysis." *American Economic Journal: Microeconomics*, 2(3): 160-185.

Kagel, John H. and Alvin E. Roth. 2000. "The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment." *Quarterly Journal of Economics*, 115(1): 201-235.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy*, 98(6): 1325-1348.

Kessler, Judd B. and Alvin E. Roth. 2012. "Organ Allocation Policy and the Decision to Donate." *American Economic Review*, 102 (5): 2018–2047.

Kessler, Judd B. and Alvin E. Roth. 2014. "Loopholes Undermine Donation: An Experiment Motivated by an Organ Donation Priority Loophole in Israel." *Journal of Public Economics*, 114 (2014) 19–28.

Klemperer, Paul, 2004. *Auctions: Theory and Practice*. Princeton and Oxford: Princeton University Press.

Kojima, Fuhito. 2009. "Random Assignment of Multiple Indivisible Objects." *Mathematical Social Sciences*, 57(1): 134-142.

Krishna, Aradhna and Utku Ünver. 2008. "Improving the Efficiency of Course Bidding at Business Schools: Field and Laboratory Studies." *Management Science*, 27 (2): 262–282.

Levin, Jonathan and Andy Skrzypacz. Forthcoming. "Properties of the Combinatorial Clock Auction." *American Economic Review*.

Li, Shengwu, 2015. "Obviously Strategy-Proof Mechanisms." Working Paper.

Milgrom, Paul. 2004. *Putting Auction Theory to Work*. Cambridge, UK: Cambridge University Press.

Milgrom, Paul. 2009. "Assignment Messages and Exchanges." *American Economic Journal: Microeconomics*, 1(2): 95-113.

Milgrom, Paul. 2011. "Critical Issues in the Practice of Market Design." *Economic Inquiry*, 49: 311–320.

Moulin, Hervé. 1995. *Cooperative Microeconomics*. London: Prentice Hall.

Myerson, Roger. 1991. *Game Theory: Analysis of Conflict.* Cambridge, Mass.: Harvard University Press.

Narita, Yusuke. 2016. “Match or Mismatch: Learning and Inertia in School Choice.” Working paper.

Niederle, Muriel and Lise Vesterlund. 2007. “Do Women Shy away from Competition? Do Men Compete too Much?” *Quarterly Journal of Economics*, 122(3):1067-1101.

Nisan, Noam. 2006. "Bidding Languages for Combinatorial Auctions." In Peter Cramton et al., eds. *Combinatorial Auctions*, MIT Press, 215-232.

Othman, Abraham, Eric Budish, and Tuomas Sandholm. 2010. “Finding Approximate Competitive Equilibria: Efficient and Fair Course Allocation.” Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems.

Pápai, Szilvia. 2001. “Strategyproof and Nonbossy Multiple Assignments.” *Journal of Public Economic Theory*, 3(3): 257–271.

Pathak, Parag A. and Tayfun Sönmez. 2008. “Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism.” *American Economic Review*, 98(4): 1636-1652.

Pathak, Parag A. and Tayfun Sönmez. 2013. “School Admissions Reform in Chicago and England: Comparing Mechanisms by their Vulnerability to Manipulation.” *American Economic Review*, 103(1): 80-106.

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

Prendergast, Canice. 2015. “The Allocation of Food to Food Banks.” Working Paper.

Reguant, Mar. 2014. “Complementary Bidding Mechanisms and Startup Costs in Electricity Markets.” *Review of Economic Studies*, 81(4): 1708-1742.

Roth, Alvin E. 2002. “The Economist as Engineer.” *Econometrica*, 70(4), 1341-1378.

Roth, Alvin E. 2008. “What Have We Learned from Market Design?” *The Economic Journal*, 118: 285–310.

Roth, Alvin E. 2015. *Who Gets What—and Why: The New Economics of Matchmaking and Market Design*. Boston: Houghton Mifflin Harcourt.

Roth, Alvin E. Forthcoming. “Experiments in Market Design.” Prepared for the *Handbook of Experimental Economics*, Volume 2.

Roth, Alvin E. and Elliott Peranson. 1999. “The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design.” *American Economic Review*, 89(4), 748-782.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review*, 81(5): 1068-1095.

Roth, Alvin E., Tayfun Sönmez, and Utku Ünver. 2004. "Kidney Exchange." *The Quarterly Journal of Economics*, 119(2): 457-488.

Roth, Alvin E., Tayfun Sönmez, and Utku Ünver. 2005. "Pairwise Kidney Exchange." *Journal of Economic Theory*, 125(2): 151-188.

Roth, Alvin E., Tayfun Sönmez, and Utku Ünver. 2007. "Efficient Kidney Exchange: Coincidence of Wants in Markets with Compatibility-Based Preferences." *The American Economic Review*, 97(3): 828-851.

Smith, Vernon. 1962. "An Experimental Study of Competitive Market Behavior." *Journal of Political Economy*, 70(2): 111-137.

Sönmez, Tayfun and Tobias Switzer. 2013. "Matching with (Branch-of-Choice) Contracts at the United States Military Academy." *Econometrica*, 81(2): 451-488.

Sönmez, Tayfun and Utku Ünver. 2003. "Course Bidding at Business Schools." Working Paper.

Sönmez, Tayfun and Utku Ünver. 2010. "Course Bidding at Business Schools." *International Economic Review,* 51(1): 99–123.

Wharton. 2014. "Course Match User Manual." Accessed at https://spike.wharton.upenn.edu/mbaprogram/course_match/Course%20Match%20Manual%20-%20v1_4.pdf

ONLINE APPENDIX

Appendix A: Study Instructions

Study Instructions

Thank you for participating in this study.

If you have a question about the study at any time, please raise your hand.

In this study you will be constructing hypothetical class schedules for the spring semester of your second year at Wharton.

You will construct a schedule twice, once under each of two different course allocation systems.

One course allocation system is a simplified version of Wharton's current MBA "Course Auction". The other is an alternative course allocation system for Wharton MBA courses called the "Course Matching System".

Half the sessions in the study will use the "Course Auction" first and half will use the "Course Matching System" first.

After you construct a schedule under each system, you will answer a series of questions about the schedule you have constructed and about the system that you used.

After you have constructed schedules under both systems, you will be asked to compare around 15 to 20 pairs of schedules. For each pair of schedules you will be asked which of the two you prefer.

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.

We are using a subset of 25 spring semester course sections. These course sections were selected to be representative in terms of scheduling, department, and popularity level.

There may be some courses that you would be interested in taking that are not included on this list. There is a limited set of courses because there are only approximately 18 students in the study today and so we cannot replicate the entire course offerings of a normal spring semester. (Note that the actual roster for this spring may differ in terms of which courses are offered, the professors teaching them, and their meeting times.)

We ask you to imagine that these are the only courses available in the spring semester of your second year at Wharton, and to construct your most preferred schedule given these courses. Since this is your last semester, any budget points that you do not use are worthless.

Please imagine that you do not need to take any particular courses for your major or any other graduation requirements, but that you do need to take 5 credit units. If you have already taken one of the courses in the sample, then you should assume that you cannot take the course again in the spring semester. On the other hand, you should assume that you can take any course in the sample that you have not already taken, that is, ignore any prerequisite requirements. Notice that all of the courses are semester length and worth one credit unit.

Imagine that this is the schedule you would construct the week before classes begin. Once classes start you would be able to drop a course, but you would have to replace it with a course that had an open seat.

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.

Note: Neither the schedules you construct nor the decisions you make in this experiment will have any impact on your actual spring semester courses or your point budget in the actual Wharton MBA Course Auction.

The course sections that are available are listed in the packet that has been given to you. Please take five minutes to look through the packet of courses that are available. Think about how interested you are in each of the courses and what would be your ideal schedule or schedules. We will begin with the first system in five minutes.

Instructions for the Course Auction

This procedure is a simplified version of Wharton's current MBA Course Auction. It is similar to the Course Auction that you have already used during your time at Wharton, but with a few differences:

- Every student starts with the same number of budget points (5,000)
- There are 4 rounds of auction activity
- All students are considered second-year students bidding on courses for their last semester
- All students need 5 credit units (CUs)

You are given a budget of 5,000 points. There are then 4 rounds of the auction, all of which we will play today. In the first round you can bid on as many courses as you would like so long as the sum of your bids is less than or equal to your budget. In the next three rounds, you can buy and sell courses with other students.

Instructions for Round 1

Submitting Bids

In the first round, you can submit bids for as many different course sections as you like. The sum of your bids cannot exceed your budget of 5,000 points.

How are prices calculated?

Prices are calculated the same way as in the current Wharton Course Auction. The price of a section is set at the highest losing bid or 100 points, whichever is higher. For example, if a section has 5 seats, the price for the section is set equal to the sixth highest bid for it, if that bid is at least 100 points, otherwise the price is 100. For example, if the sixth highest bid is 120, then the five highest bidders would each get a seat and be charged 120 points. If fewer students bid for a section than it has seats, then the price of the section is set to 100.

What sections do I get?

You get any section for which your bid is greater than or equal to the price. In the event of a tie, where two or more students submit exactly the same bid and there is not enough space for all of them, the computer randomly assigns the available seats to students who bid that amount.

What happens to my budget?

For each section that you receive, your budget will be decreased by the price of the section. For example, if you bid 1000 for the only section of Course A and its price is 400, then you will receive a seat in Course A, and your budget will be decreased by 400 points. If you do not get a seat in the course then you will not give up those 400 points.

Instructions for Rounds 2, 3, and 4

Submitting Bids and Asks

In Rounds 2 through 4, you can submit bids for as many different sections as you like, just as in Round 1. You can also submit asks, which are offers to sell, for any section that you currently have. The sum of your bids cannot exceed your current budget. You can ask whatever amount you like.

How are prices calculated?

For any section where there are both bids and asks, a trading price is set if there is at least one bid higher than the lowest ask. When this is the case, the computer sets a price to make as many trades as possible. This involves finding a price such that the number of bids higher than that price is the same as the number of asks lower than that price.

Suppose the following bids and asks are submitted for a section during a round.
Bids: 101, 323, 143, 103, 187, 280, 156, and 152.
Asks: 225, 64, 298, 171, and 0.
To see which bids and asks are successful and what the clearing price is, first arrange all the bids in descending order and the asks in ascending order as shown in the table below:

Only three trades can take place at a uniform price.

Bids	Asks
323	0
280	64
187	171
156	225
152	298
143	
103	
101	

Since only the top three bids are higher than the three lowest asks (and the fourth highest bid is lower than the fourth lowest ask), only three trades can go through. The clearing price is determined as the larger of the first losing bid and the highest winning ask; in this case, the first losing bid is 156, and highest winning ask is 171 — hence the clearing price is 171. The clearing price amount is transferred from each of the successful bidders to each successful seller (the accounts of unsuccessful bidders and sellers remain unaffected).

If there are extra seats in a section, for example if a section does not reach capacity in Round 1, then those seats are treated as if they are being offered for an ask of 100 points.

You can always be guaranteed to drop a section by submitting an ask of “0”.

What should my schedule look like at the end of Round 4?

At the end of Round 4 you should have: (1) no more than 5 credit units in your schedule; (2) no sections that have a time conflict with each other; and (3) no more than one section in each course.

Is my schedule after Round 4 my final schedule?

Not necessarily. Recall, you should imagine that this is the schedule you would construct the week before classes begin. Once classes start you would be able to drop a course, but you would have to replace it with a course that had an open seat.

If you have any questions, please raise your hand.

Instructions for Between Systems

You have just constructed a schedule under the first system and answered some questions about the schedule and the system. You will now construct a schedule under the other system.

You are constructing a schedule in this system starting “from scratch” such that the decisions you and the other students in this session made while using the first system do not affect anything about activity in this system.

You should again construct the best schedule you can for your spring term of your second year at Wharton. The same course sections are available for this system as were available for the last one.

Instructions for the Course Matching System

The Course Matching System is different from the Wharton Course Auction with which you may be familiar.

The Course Matching System works differently from an auction in that you do not directly bid for course sections. Instead, the computer acts as your agent to buy the best schedule of courses you can afford.

Your job is to tell the computer how much you value individual course sections and whether you assign extra value (or negative value) to having certain course sections together. This process will be explained in detail below.

Since you can tell the computer how much you like every course or pair of courses that might be in your schedule, the Course Matching System only needs one round. In that round, the computer will use your preferences to buy you the best schedule you can afford.

Since the computer is going to optimally buy courses for you, your job is to provide the computer with all the information it needs about how much you value the courses. This is obviously very important, since the computer is going to buy the optimal schedule for you given only what it knows about how you value courses.

The way to communicate your values to the computer is as follows:

1) You tell the computer how much you value each course section that you have any interest in taking.

- First, you pick a favorite course section and assign it a value of 100.
- Second, you assign all other course sections that you have any interest in taking a value between 1 and 100.

The reason that you assign your favorite course section a value of 100 and all other sections a number between 1 and 100 is that all values are relative.

For example, if you value every course at 100 then you are telling the computer that you value all courses equally. If you value one course at 100 and another course at 50, you are telling the computer you value the course at 100 twice as much as the course at 50.

Unlike using other course allocation systems, when using the Course Matching System, you do not need to think about what other people are doing. All you need to do is communicate how you value course sections to the computer so it knows how to make tradeoffs for you.

How does assigning value to courses work?

Suppose that among the many course sections you assign a positive value, you tell the computer the following values for the single section courses A through E:

Course A = 100
Course B = 80
Course C = 60
Course D = 15
Course E = 10

This tells the computer that you are particularly interested in Courses A, B and C, and somewhat interested in Courses D and E. In particular, it tells the computer that you prefer getting Courses A, B, and C (100 + 80 + 60 = 240) than getting Courses A, D, and E (100 + 15 + 10 = 125).

It also tells the computer that you prefer getting Courses B and C (80 + 60 = 140) than Courses A, D, and E, which only sum to 125. <u>For any two schedules, the computer thinks you prefer whichever schedule has a larger sum.</u>

For simplicity, this example valued only 5 course sections. You should list a positive value for as many courses that you have any interest in taking. We recommend that you assign a positive value to at least 12 course sections. This way the computer can distinguish between a section that has low positive value to you and a section that has zero value to you.

Can I assign values for multiple sections of the same course?

Yes, and you will probably want to do this. To explain, suppose three sections of a course are offered, all on Mondays and Wednesdays. Professor Smith teaches the 10:30-12:00 and 12:00-1:30 sections while Professor Jones teaches the 3:00-4:30 section. You may assign values of 90, 80 and 15 to these three sections, respectively, to signify that you greatly prefer Professor Smith to Professor Jones, and slightly prefer 10:30 to 12:00. Because you can only take one section of a course, you will be assigned at most one of these three course sections, even though you entered values for all three.

Again, there is no limit to the number of course sections that you may assign a positive value.

2) You tell the computer if you assign extra (or negative) value to certain pairs of classes.

To do this, you check the boxes next to any two sections and indicate an extra positive or negative value to having both sections together. These “adjustments” are shown at the top of the page of your valuations.

Why might I assign extra value to two courses together?

Some students might get extra value from having two courses that are back-to-back in their schedule (e.g. they do not like breaks between classes).

Some students might get extra value from having two courses that are related in their schedule (e.g. they might get extra value from taking two courses from the same department if each one becomes more useful with the other).

You can think of these courses as complements, i.e. the combination of the two courses together is greater in value than the sum of their values.

How does assigning extra value work?

Suppose you specify the following values for single section courses A through C:

Course A = 40
Course B = 30
Course C = 85

And suppose you assign an extra value of 20 for getting Course A and Course B together.

Then you are telling the computer that getting Course A and Course B together in your schedule has a value of 90 (90 = 40 for Course A + 30 for Course B + 20 for getting both together).

This means that the computer would try to get you Course A and Course B together before trying to get you Course C. If you had not assigned the extra value to Courses A and B together, the computer would have tried to get you Course C before trying to get you Courses A and B.

Why might I assign negative value to two courses together?

Some students might get negative value from having two courses that are back-to-back in their schedule (e.g. they prefer to take breaks between classes).

Some students might get negative value from having two courses that are related in their schedule (e.g. they might decide that they only want to take one class from a certain department).

You can think of these courses as substitutes, i.e. the second course is worth less when you already have the first.

How does assigning negative value work?

Suppose you specify the following values for single section courses A through C:

Course A = 40
Course B = 30
Course C = 55

And suppose you assign a negative value of -20 for getting Course A and Course B together.

Then you are telling the computer that getting Course A and Course B together in your schedule has a value of 50 (50 = 40 for Course A + 30 for Course B - 20 for getting both together).

This means that the computer would try to get you Course C before getting you Course A and B together. If you had not assigned the negative value to Courses A and B together, the computer would have tried to get you Courses A and B before trying to get you Course C.

You can also use an adjustment to tell the computer "I want to take at most one of these two courses". Using the example above, suppose you want to take either Course A or Course B, but you absolutely do not want to take both. Then you should assign a negative value of -70 for Course A and B together. That negative adjustment tells the computer that the combination has value 0 to you (0 = 40 for Course A + 30 for Course B – 70 for getting both together). Therefore, you may get Course A or Course B, but the computer will never get both for you.

When do I not need to enter in an adjustment?

You do not need to enter an adjustment when two sections are from the same course or two sections are offered at the same time. The computer already knows that you cannot take these sections together. For example, if Professor Baker teaches two sections of the same course, one from 9:00-10:30 and the other from 10:30-12:00, then you can assign a positive value for each of them, but you don't need to assign a positive or negative adjustment for the combination.

Once the computer knows how much you value each course section, it will buy the best schedule you can afford.

How do I know that I am reporting my values right?

To help make sure you are reporting your values right, you can click a button on the navigation bar to see your top 10 schedules. Given the values you reported, the computer thinks that these are your 10 favorite schedules, ranked in order. This means that the computer will try to buy you these schedules in this order. If the order of these schedules does not look right to you, go back and adjust your values until they appear in the right order.

What is my budget that the computer will use to buy courses for me?

Each student is given a budget of 5,000 points.

How are prices determined?

The Course Matching System sets prices based on demand for the courses so that demand equals supply. Courses that are more highly demanded get higher prices and courses that are less popular get lower prices or prices of zero.

One way to think about how prices are set is that each student's computer asks for the best possible schedule for its student. When everyone has their best possible schedule, some courses will have too many students. The price of those courses will rise. Then, given the new set of prices, each student's computer asks again for the best possible schedule for its student at the new set of prices. Some courses will be undersubscribed or oversubscribed and prices will adjust again. This process repeats until there is a set of prices where all popular courses are full and every student gets their best possible schedule given those prices.

Given the set of prices, it may be necessary to break a tie between two or more students who want a course section. These potential ties are broken by assigning a randomly selected small budget increase to each student.

Shouldn't the values I report to the computer depend on the prices of courses or other student's values?

No! The Course Matching System is designed so 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.

To see this, notice that if your favorite course, to which you assign a value of 100, is a course whose demand is less than the number of available seats, then it will have a price of zero and you will get that course without using any of your budget. The computer can then use the remainder of your budget to try to get the other course sections that you value highly.

Another way to think about reporting your values to the computer is to imagine you are sending the computer to the supermarket with your food budget and a list of your preferences for ingredients for dinner. You want to report your true values so that the computer can make the right tradeoffs for you when it gets to the supermarket and observes the actual prices for each ingredient.

Are my values equivalent to "bids"?

No! As mentioned above your values are only compared to each other and never compared with other students' values.

Is the schedule I receive after I report my values my final schedule?

Not necessarily. Recall, you should imagine that this is the schedule you would construct the week before classes begin. Once classes start you would be able to drop a course, but you would have to replace it with a course that had an open seat.

If you have any questions, please raise your hand.

Please use this page to write any additional comments about your experience during this session. These are anonymous comments, so please do not include your name.

Appendix B: List of Course Sections Available in Experiment and Excerpt of Course Descriptions

At the beginning of each session, along with the instructions reproduced as Appendix A, we distributed to students the list of course sections available in the experiment as well as course descriptions. This list and the first two course descriptions are reproduced below and on the following pages. The number of available seats was selected by the Wharton Committee to create scarcity in the laboratory environment anticipating 20 subjects per session. Our actual turnout varied between 14-19 subjects per session. In order to maintain scarcity with fewer subjects we adjusted course capacities as follows. If 18-19 subjects attended, we used the capacities below (107 seats total). If 16-17 subjects attended, we turned five-seat courses into four-seat courses (97 seats total). If 14-15 subjects attended we turned five-seat courses into four-seat courses and turned four-seat courses into three-seat courses (86 seats total).

Course	**Title**	**Instructor**	**Day Code**	**Start Time**	**Stop Time**	**Available Seats**
ACCT742	PROBLEMS IN FIN REPORTIN	LAMBERT R	MW	0130PM	0300PM	5
ACCT897	TAXES AND BUS STRATEGY	BLOUIN J	MW	1200PM	0130PM	4
FNCE726	ADVANCED CORP FINANCE	VAN WESEP,E	TR	1200PM	0130PM	5
FNCE728	CORPORATE VALUATION	CICHELLO M	MW	0300PM	0430PM	4
FNCE750	VENT CAP & FNCE INNOVAT	WESSELS D	MW	0130PM	0300PM	4
FNCE750	VENT CAP & FNCE INNOVAT	WESSELS D	MW	0300PM	0430PM	4
FNCE891	CORPORATE RESTRUCTURING	JENKINS M	TR	0130PM	0300PM	4
LGST806	NEGOTIATIONS	DIAMOND S	R	0300PM	0600PM	3
LGST806	NEGOTIATIONS	BRANDT A	W	0300PM	0600PM	3
LGST809	SPORTS BUSINESS MGMT	ROSNER S	TR	0300PM	0430PM	5
LGST813	LEG ASP ENTREPRENRSHP	BORGHESE R	M	0300PM	0600PM	5
MGMT691	NEGOTIATIONS	MUELLER J	TR	1030AM	1200PM	3
MGMT721	CORP DEV: MERG & ACQUIS	CHAUDHURI S	TR	0900AM	1030AM	4
MGMT721	CORP DEV: MERG & ACQUIS	CHAUDHURI S	TR	1030AM	1200PM	4
MGMT782	STRATEGIC IMPLEMENTATION	MURMANN J	TR	1200PM	0130PM	5
MGMT833	STRAT & PRAC OF FAMILY	ALEXANDER W	TR	0130PM	0300PM	4
MKTG756	MARKETING RESEARCH	IYENGAR R	MW	1030AM	1200PM	5
MKTG773	CUSTOMER BEHAVIOR	REED A	TR	1030AM	1200PM	5
MKTG776	APPL PROB MODELS MKTG	FADER P	W	0300PM	0600PM	5
MKTG778	STRATEGIC BRAND MGMT	MOGILNER C	TR	0130PM	0300PM	5
OPIM690	MANAG DECSN MAKING	MILKMAN K	MW	0130PM	0300PM	5
OPIM692	ADV TOPICS NEGOTIATION	SCHWEITZER M	TR	0130PM	0300PM	4
REAL721	REAL ESTATE INVESTMENTS	FERREIRA F	MW	0130PM	0300PM	4
REAL721	REAL ESTATE INVESTMENTS	WONG M	TR	0130PM	0300PM	4
REAL821	REAL ESTATE DEVELOPMENT	NAKAHARA A	W	0300PM	0600PM	4

ACCT742: PROBLEMS IN FIN REPORTIN - LAMBERT R

Financial statements are a primary means for firms to communicate information about their performance and strategy to investors and other groups. In the wake of numerous accounting scandals and the recent financial meltdown (which accounting both helped and hindered), it is more important than ever for managers and investors to understand (i) the financial reporting process, (ii) what financial statements do and do not contain, and (iii) the types of discretion managers have in presenting transactions they have undertaken. This course is designed to help you become a more informed user of accounting numbers by increasing your ability to extract, interpret, and analyze information in financial statements.

While this is not a course in equity valuation per se, equity valuation is one of the most common uses of financial statement data. Accordingly, we will examine the relation between Accounting 742 -stock prices and financial statement information. We will also study the use of financial ratios and forecasted financial statement data in models of distress prediction.

ACCT897: TAXES AND BUS STRATEGY - BLOUIN J

Traditional finance and strategy courses do not consider the role of taxes. Similarly, traditional tax courses often ignore the richness of the decision context in which tax factors operate. The objective of this course is to develop a framework for understanding how taxes affect business decisions.

Part of being financially literate is a having a basic understanding of how taxation affects business decisions that companies typically face: forming the business and raising capital, operating the firm, distributing cash to shareholders through dividends and share repurchases, expanding through acquisition, divesting lines of business, and expanding internationally. Taxes have a direct impact on cash flow and often divert 40% to 50% of the firm’s pretax cash flow to the government. Having an understanding of taxation and how firms plan accordingly is important whether you will be running the firm (e.g., executive in large company, entrepreneur, or running a family owned business) or assessing it from the outside (e.g., financial analyst, venture capitalist, or investment banker). Taxes are everywhere and it pays to have some understanding of them.

…

[Subjects received course descriptions like the above for all 21 distinct courses in the experiment.]

Appendix C: Recruitment Materials

From: Kaufold, Howard
Sent: Thursday, November 17, 2011 3:09 PM
To: whg12; whg13
Subject: Do Wharton Research Study, Get Free Food, and Earn Your Chance at Cash Prize!

Dear Students,

We would like to ask for your help in a research study that is recruiting current Wharton MBA students. The research, conducted by a Wharton faculty member along with one of our curricular committees of faculty, department chairs and students, is attempting to understand the decisions of Wharton MBA students as they relate to pending changes in the MBA program. Through this study we will learn valuable information that we will use to improve the experience of Wharton students for years to come.

We want to emphasize that your participation is strictly voluntary. However, as a token of our appreciation, at the end of each session we will randomly choose two students and each one will receive $250. (Each session will have approximately 20 students.) In addition, we will provide you with lunch (noon sessions) or dinner (6pm sessions). Your help will also be greatly appreciated as we want to ensure that we understand as best as possible the preferences of our MBA students with respect to these important design changes in the MBA program.

The study will last 90 minutes and take place in either Room F80 or F375 of Jon M. Huntsman Hall. Sessions will begin at 12 **noon** and 6**pm** on
Monday 11/21 – F375 JMHH
Monday 11/28 – F80 JMHH
Tuesday 11/29 – F80 JMHH
Wednesday 11/30 – F80 JMHH
Thursday 12/1 – F80 JMHH

Please click http://mktgweb.wharton.upenn.edu/mba-bhlab/ to sign up for any available time slot on one of the days listed above. (You need only participate in one session.)

We understand that this a busy time of the year for all students, but we do very much hope you will be able to help us with this valuable research study for our MBA program. Thanks in advance.

Yours,
[SIGNATURE] [SIGNATURE]
Thomas S. Robertson, *Dean* Howard Kaufold, *Vice Dean*

Appendix D: Subject Representativeness

Subjects were representative of all Wharton MBA students on demographics as well as attitudes towards, and behavior in, the Wharton Auction. Using data provided by the Wharton Dean's Office, Table A1 shows the demographics of our 132 subjects as well as the universe of Wharton MBA students in the 2011-2012 academic year. The final column reports the p-value of either a *test of proportions* or a *t-test* comparing our subjects to the universe of students. We see that based on demographics, our subjects are representative of the Wharton student body with $p>0.1$ for each variable except race.

Important for our purposes, our subjects look identical to the student body with regard to Auction behavior: namely, the number of points they had at the start of the Spring Auction (which began before the study took place) and the number of points they had when our study took place (points in the fourth round of the Spring Auction). For the second-year students in our study, we also examine data on their attitude towards the Wharton Auction as measured on the preceding spring's stakeholder survey. Our second-year subjects were almost identical to the universe of second-year subjects in reports in the effectiveness of the Wharton Auction.

Table A1: Representativeness of Experimental Subjects

	Subjects	Wharton MBAs	p-value (two-sided)
N	132	1660	
Panel A: Demographics			
First Year Student	51.7%	50.8%	0.83
Female	42.0%	47.0%	0.27
From United States	37.1%	34.3%	0.52
Finance Major	23.5%	25.7%	0.57
Total Registered Credits	17.1	17.0	0.96
Wharton Credits	11.5	11.3	0.56
White	48.5%	37.2%	0.01***
Asian	20.5%	27.0%	0.10*
Black, Non-Hispanic	5.3%	4.0%	0.46
Hispanic	3.0%	3.4%	0.83
Multi-Race	8.3%	7.2%	0.62
No race reported	14.4%	21.1%	0.07*
GPA	Subjects directionally higher		0.14
Panel B: Auction Behavior			
Points at Start of Spring Auction	6899.6	6966.4	0.79
Points in 4th Round of Spring Auction	4992.3	4960.7	0.92
	Subjects	**Wharton MBAs**	p-value (two-sided)
N	62	731	
Panel C: Auction Beliefs (Second years only)			
Reported Auction Effectiveness (0 to 7)	4.69	4.68	0.96

Table A1 reports data provided by Wharton. Due to Wharton's policy of grade non-disclosure, GPA levels cannot be reported. The auction beliefs data in Panel C came from the stakeholder survey completed by rising second year students the preceding spring, so we only have it for the second-year students. Tests are two-sided t-tests (for continuous variables) or two-sided tests of proportions (for binary variables).

Appendix E: Order Effects of Main Results

In four of our eight sessions, subjects used the Auction first; in the other four sessions subjects used CEEI first. If using CEEI forces subjects to think about their preferences more deeply than using the Auction — and this deeper thought contributes to better outcomes — then we might expect CEEI to do particularly well relative to the Auction when subjects use the Auction before CEEI (i.e. before they have engaged in the deep thought) as compared to when they use the Auction after CEEI. Here we show that our main fairness and efficiency results are nearly identical regardless of which mechanism was used first.

Table A2: Order Effects on Envy — Results from Binary Comparisons

	All data			Auction first			CEEI first		
Panel A: By Subject									
	Auction n=119	CEEI n=117	p-value	Auction n=56	CEEI n=56	p-value	Auction n=63	CEEI n=61	p-value
% of subjects any envy	42%	31%	0.036	41%	34%	0.218	43%	28%	0.041
% of subjects large envy	34%	21%	0.008	32%	23%	0.145	37%	18%	0.011
Panel B: By Comparison									
	Auction n=499	CEEI n=475	p-value	Auction n=240	CEEI n=221	p-value	Auction n=259	CEEI n=254	p-value
% of comp any envy	19%	12%	0.002	18%	14%	0.154	20%	11%	0.001
% of comp large envy	14%	8%	0.002	13%	10%	0.124	14%	6%	0.002

Table A2 reproduces the results from Table 1 in the first three columns and then splits the data by whether subjects used the Auction or CEEI first. All p-values are from one-sided tests of proportions.

Table A3: Order Effects on Allocative Efficiency — Results from Binary Comparisons

	All data (n=132 subjects, 8 sessions)				Auction first (n=66 subjects, 4 sessions)				CEEI first (n=66 subjects, 4 sessions)			
	Auct	CEEI	None	p	Auct	CEEI	None	p	Auct	CEEI	None	p
Individual preference	42	56	34	0.094	22	29	13	0.201	20	27	19	0.191
Session preference	0	6	2	0.016	0	2	2	0.250	0	4	0	0.063

Table A3 reproduces results from Table 3 in the first four columns and then splits the data by whether subjects used the Auction or CEEI first. All p-values are from one-sided binomial probability tests.

Appendix F: Preference Reporting Summary Statistics

Table A4 presents summary statistics describing how subjects used the preference reporting language. The data suggest that subjects generally followed the instructions we provided. We advised subjects to report positive cardinal values for at least twelve courses. The median number of courses assigned positive values was 12 and the vast majority of subjects (76.5%) reported positive values for 11 or more courses. In addition, we advised subjects to assign their favorite course a value of 100 and to assign all other courses a relative value. Again, the vast majority of subjects (75.0%) reported a value for 100 for one and only one course. Generally speaking, subjects spread their values of courses evenly from 0 to 100. The last three rows suggest that most subjects chose not to use any adjustments (the median subject used 0 adjustments) and the average number of adjustments across all subjects was slightly more than 1. Of those adjustments that were made, they were evenly split between positive and negative adjustments.

Table A4 Use of the Preference Reporting Language

	Mean	Min	25th Pct.	Median	75th Pct.	Max
# courses valued $v>0$	12.45	7	11	12	14	24
# courses valued $v=100$	1.40	0	1	1	1	8
# courses valued $50 \leq v \leq 99$	4.87	0	3	5	7	10
# courses valued $0<v<50$	6.17	0	4	6	8	17
# adjustments	1.08	0	0	0	2	10
# adjustments > 0 (complements)	0.55	0	0	0	1	10
# adjustments < 0 (substitutes)	0.53	0	0	0	1	6

Table A4 reports use of the preference reporting language for the 132 subjects in the experiment. v is the cardinal value assigned to a particular course section.

Table A5 performs an analysis of preference reporting contradictions as a function of whether a subject used the adjustments feature of the preference reporting language. The results suggest that subjects who used the adjustment feature were directionally less likely to make preference reporting contradictions, although the differences are not statistically significant. This result is consistent with the results in the main text, which found that binary comparisons were directionally less likely to contradict preference reports when at least one schedule triggered an adjustment.

Table A5: Prevalence and Magnitude of Preference Reporting Contradictions for People who Use (and do not Use) Adjustments

Type of Comparison	# Comparisons with this Utility Difference	Accurate	Contradictions			
			All	Weak Preference	Preference	Weak Preference
All	1,662	84.42%	15.58%	5.17%	6.92%	3.49%
Did not use Adjustments	878	85.31%	14.69%	5.01%	6.38%	3.30%
Used Adjustments	784	83.42%	16.58%	5.36%	7.53%	3.70%

Table A5 shows all 1,662 comparisons. *Did Not Use Adjustments* indicates that the subject did not make an adjustment in the CEEI preference reports. *Used Adjustments* indicates that the subject made at least one adjustment in the CEEI preference repots. *Accurate* reports the percentage of these comparisons where the binary comparison choice confirms the CEEI preference report prediction. The *Contradictions* columns report the percentage of binary comparisons that contradicted the CEEI preference reports overall and at each level of preference.

Appendix G: Qualitative Questions

After subjects used each course allocation mechanism they answered qualitative questions about the schedule they received and the mechanism they had just used. After subjects had used both mechanisms, they answered additional questions. All of these qualitative questions were asked to explore for "side effects" as discussed in the body of the paper. The full list of questions asked of subjects is listed in Table A6. In addition, subjects were given a page to write free responses at the end of the experiment.

Table A6: Qualitative Questions Subjects Answered

Question Wording	Timing	Scale
The way courses are allocated through this course allocation system is fair.	After using the first mechanism and again after using the second mechanism	“Strongly Disagree” “Disagree” “Somewhat Disagree” “Neither Agree or Disagree” “Somewhat Agree” “Agree” “Strongly Agree”
This course allocation system is easy for me to use.		
I understand how this course allocation system works.		
This course allocation system led to the best outcome I could hope for.		
I am satisfied with my course outcome.		
I enjoyed participating in this course allocation system.		
I like this course allocation system.		
My fellow students will like this course allocation system.		
I felt like I had control over my schedule in this course allocation system.		
This course allocation system is simple.		
I had to think strategically about what other students would do in this course allocation system.		
Someone with perfect knowledge of the historical supply and demand for courses could have had an advantage over me in this system.		
Which course allocation system did you prefer?	After using both mechanisms and completing binary comparisons	“Strongly Prefer 1st” “Prefer 1st” “Slightly Prefer 1st” “Unsure Which I Prefer” “Slightly Prefer 2nd” “Prefer 2nd” “Strongly Prefer 2nd”
Which course allocation system do you think your fellow students would prefer?		
In which course allocation system did you get a better schedule?		

Table A6 shows all the qualitative questions subjects were asked, when the questions were asked, and the responses available to the subjects.