Describing Deferred Acceptance and Strategyproofness to Participants: Experimental Analysis*

Yannai Gonczarowski† Ori Heffetz‡ Guy Ishai§ Clayton Thomas¶

July 9, 2024

Abstract

We conduct an incentivized lab experiment to test participants' ability to understand the DA matching mechanism and the strategyproofness property, conveyed in different ways. We find that while many participants can (using a novel GUI) learn DA's mechanics and calculate its outcomes, such understanding does not imply understanding of strategyproofness (as measured by specially designed tests). However, a novel *menu* description of strategyproofness conveys this property significantly better than other treatments. While behavioral effects are small on average, participants with levels of strategyproofness understanding above a certain threshold play the classical dominant strategy at very high rates.

*Keren-Or Barashi Gortler, Itamar Bellaiche, Yehonatan Caspi, Gabriela Cohen-Hadid, Ayala Goldfarb, Michael Khalfin, Ido Leshkowitz, Josef Mccrum, Shenhav Or, Yonatan Rahimi and Ohad Weschler provided excellent research assistance. The authors thank Eric Budish, Ben Enke, Nicole Immorlica, David Laibson, Assaf Romm, Shigehiro Serizawa, Ran Shorrer, Alex Teytelboym, and Leeat Yariv for helpful discussions; participants at the Stanford Institute for Theoretical Economics (SITE) 2023 Experimental Economics, SITE 2023 Market Design, WZB Berlin Matching Workshop, Crown Family Israel Center for Innovation (ICI) 2024 Academic Conference, Virtual Market Design Seminar, 2024 Marketplace Innovations Workshop, 8th Solomon Lew Conference on Behavioral Economics (Tel Aviv), 1st Annual Chicago Booth Market Design Conference, and seminar participants at Bar Ilan, Cornell, and the Hebrew University for comments that significantly improved the paper; and Adam Chafee and his team at the Cornell Business Simulation Lab for their help with running the experiment. The authors gratefully acknowledge research support by the following sources. Gonczarowski: National Science Foundation (NSF-BSF grant No. 2343922), Harvard FAS Inequality in America Initiative, and Harvard FAS Dean's Competitive Fund for Promising Scholarship. Heffetz: Israel Science Foundation (grant No. 2968/21), US-Israel Binational Science Foundation (NSF-BSF grant No. 2023676), Cornell's S.C. Johnson School, and Cornell's Center for Social Sciences. Ishai: Barbara and Morton Mandel Doctoral Program, Bogen Family, and Federmann Center for Rationality. Thomas: NSF CCF-1955205, Wallace Memorial Fellowship in Engineering, and Siebel Scholar award; part of his work was carried out while in Princeton's Department of Computer Science.

The full online experimental materials are available at the authors' websites.

†Department of Economics and Department of Computer Science, Harvard University. E-mail: yannai@gonch.name.

‡Johnson Graduate School of Management, Cornell University, Bogen Department of Economics and Federmann Center for Rationality, The Hebrew University of Jerusalem, and NBER. E-mail: oh33@cornell.edu.

§Bogen Department of Economics and Federmann Center for Rationality, The Hebrew University of Jerusalem. E-mail: guy.ishai@mail.huji.ac.il.

¶Microsoft Research. E-mail: clathomas@microsoft.com.

1 Introduction

To what extent do participants in the widely used Deferred Acceptance matching mechanism (Gale and Shapley, 1962; henceforth, DA) understand how it works? To what extent do they understand one of the mechanism's most celebrated properties, namely, its strategyproofness? Are there principled ways to change the structure of how DA and strategyproofness are described to participants that could improve their understanding? In this paper we experimentally study these questions. We compare a traditional DA description, and a description of the definition of strategyproofness itself, with new *menu* versions of both, theoretically developed in Gonczarowski et al. (2023). We study effects of changing descriptions on participants' understanding of the mechanism, on their understanding of strategyproofness, and on the strategies they play in the mechanism.

DA is a mechanism used in many real-world matching systems, from residency matching to school choice. A crucial property of DA is that it is strategyproof: under classical assumptions on participants' preferences, straightforward (henceforth, SF) reporting—i.e., ranking options from highest to lowest value—is a dominant strategy.[1] In theory, strategyproofness obviates any need to strategize, promoting ease of participation, reducing inequality across those with different levels of strategic sophistication, and increasing robustness of theoretical predictions on play—assuming, of course, that participants are *aware* that the mechanism is strategyproof.

Unfortunately, growing evidence from both the lab and the field shows that DA participants frequently do not play straightforwardly. This behavior has been the focus of much recent work in market design (for surveys, see Hakimov and Kübler, 2021; Rees-Jones and Shorrer, 2023). Various theoretical explanations of non-straightforward (henceforth, NSF) strategies have been suggested—from cognitive failures to nonclassical preferences—and based on these explanations, several papers have suggested that (in some contexts and with important caveats) NSF play might be lowered by using dynamic, interactive mechanisms.[2] Other pragmatic approaches to mitigating NSF behavior have also been suggested, such as providing participants with trusted, concrete advice on how to determine their rank-

[1] Throughout this paper, we refer to the term (non-)straightforward, abbreviated (N)SF, to refer to the strategy of (not) ranking options from highest to lowest value. The SF strategy is sometimes described in past research as the "truthtelling strategy," but we adopt different terminology to separate participants' behavior from any notions of dishonesty.

[2] For examples in various mechanisms, Li (2017) and Pycia and Troyan (2023) suggest that NSF behavior can be explained by failures of contingent reasoning, and Dreyfuss et al. (2022b) and Meisner and von Wangenheim (2023) suggest expectations-based loss aversion. Based on their respective explanations, these papers propose changing the dynamic implementation of the mechanism. Other papers investigating dynamic implementations include Kagel and Levin (1993), Breitmoser and Schweighofer-Kodritsch (2022), and Bó and Hakimov (2023). See Section 4.

ing.[3]

Despite ample progress towards a better understanding of NSF play, our knowledge remains limited about one potential channel inducing it: To what extend do participants play NSF due to general *misunderstandings*, of either the mechanism itself or the property of strategyproofness? This question, and the question of how to mitigate such misunderstanding, are the focus of the current paper.

We conduct an incentivized lab experiment with five between-subjects treatments. We first expose participants to one of five descriptions of DA or its strategyproofness property, accompanied with specifically tailored training modules. We then elicit ranking behavior in ten incentivized DA rounds. Finally, we test understanding of strategyproofness. We use a static DA setting in which four participants—one human and three computerized—are matched to four prizes based on participants' submitted rankings of the prizes and their exogenous priorities for getting the prizes.

Our design features two novel aspects. First, inspired by prior real-world approaches to explaining DA (see, e.g., NMS, 2020), we construct a new GUI (graphical user interface) in which participants can perform the sequence of DA proposals and tentative acceptances for themselves to calculate the outcome. This GUI gives us effective tools towards not only *teaching* the DA algorithm to participants, but also *assessing* their understanding of its (taught) mechanics—providing us with new, DA-understanding outcome variables. Second, we directly test participants' understanding of *strategyproofness* (separately from assessing their understanding of DA mechanics and from their actual play), via eighteen quiz-style questions that ask participants either (i) which outcomes are logically possible in different abstract scenarios, or (ii) how participants should practically play to maximize their earnings—all solvable using the definition of strategyproofness alone. These tests provide a rich, second set of new outcome variables that allow us to assess measures of strategyproofness understanding (henceforth, SP understanding) separate from both (our new) DA-understanding outcomes and (routinely studied) SF-play outcomes. Together with our experimental flow, our framework provides a general and extensive method of testing participants' understanding of different facets of mechanisms based on descriptions, which we view as a methodological contribution of our paper (in addition to the substantive contribution of our main findings below).

In our first treatment—Traditional DA Mechanics (Trad-DA)—we examine participants' response to the traditional, complete and unambiguous description of (the mechanics of, i.e.,

[3]See, e.g., Guillen and Hing (2014); Masuda et al. (2022); Guillen and Veszteg (2021); Rees-Jones and Skowronek (2018); for details, see Section 4.

how the outcome is calculated in) the participant-proposing DA algorithm.

Our second treatment is motivated by recent theoretical work by Gonczarowski et al. (2023) (henceforth, GHT). GHT construct novel descriptions of (static, direct-revelation) mechanisms where strategyproofness holds via a simple mathematical proof. In particular, GHT constructs *menu descriptions*, which, inspired by Hammond (1979), present the mechanism to participant i via two steps:

- **Step (1)** uses only the reports of other participants to describe, in complete and explicit detail, the set of outcomes participant i might receive, called i's *menu*.
- **Step (2)** describes how to award participant i her favorite outcome (according to her report) from her menu.

The main idea behind menu descriptions is that strategyproofness for participant i follows from a menu description via a one-sentence proof: i's menu in Step (1) cannot be affected by her report, and SF reporting guarantees i her favorite outcome from the menu in Step (2). We investigate participants' reactions to a menu description in order to compare it to the traditional description of DA, where the proof is more challenging.[4]

This second treatment—Menu DA Mechanics (Menu-DA)—relays to participants a menu description of DA that follows the main positive result of GHT, i.e., it relays the two-step outline above while providing full details on (the mechanics of) how the menu is calculated in Step (1). This menu description constructed by GHT is more involved than the traditional description, and we convey it to participants using a modified and extended version of the GUI from Trad-DA. Both of our DA Mechanics treatments provide in-depth coaching to participants, who learn to calculate their DA outcomes for themselves within the GUI. These two treatments are designed to be as directly comparable to each other as reasonably possible, including visually and in the mechanics-understanding measures they elicit.

Our third treatment—Menu SP Property (Menu-SP)—departs from our two DA Mechanics treatments by focusing on relaying *only* that the matching mechanism satisfies the strategyproofness property, without specifying how one could calculate the mechanism's outcome. In particular, the Menu-SP treatment relays the two-step outline above, while providing no details on *how* the menu is calculated in Step (1). This treatment provides a potential real-world approach to conveying strategyproofness, allowing comparison between Menu-DA and its stripped-down version that contains only the information directly relevant for knowing that the mechanism is strategyproof. (And, since every strategyproof mech-

[4] Katuščák and Kittsteiner (2020) also experiment with a menu description (of the Top Trading Cycle mechanism) in a matching environment, though without our new design features and outcome variables. For details and other related work, see Section 4.

anism has a menu description, this treatment conveys no more information than that the mechanism is strategyproof.)

Our fourth treatment—Textbook SP Property (Textbook-SP)—also relays (only) the fact that the mechanism is strategyproof. It does so using a definition inspired by classical textbook ones (but written in simpler, everyday language), providing a mathematically equivalent version for comparison with Menu-SP.

Importantly, none of our treatments gives explicit strategic advice on how participants should construct their reported rankings. In particular, our two SP Property treatments are designed with an important conceptual distinction in mind between *advice*—such as recommending participants play the SF strategy—which the treatments avoid; and *the strategyproofness property*—a particular property of the mechanism, i.e., of the mapping from reported rankings to allocations—that the treatments aim to teach. Both of our SP-Property treatments also provide coaching to participants, using questions similar in nature to those in the SP-understanding tests, accompanied with detailed feedback. As with our two DA-Mechanics treatments, the two SP-Property treatments are designed to be as directly comparable to each other as reasonably possible.

Finally, our fifth treatment is a control, close-to-zero-information treatment, termed Null. It provides a generic description explaining that the mechanism generates a matching while attempting to accommodate participants' submitted rankings, without any details explaining either the mechanics of how the matching mechanism works or that it is strategyproof.

We emphasize that the incentivized DA-rounds module of our experiment is identical in all five treatments. Thus, for example, while the DA Mechanics treatments (Trad-DA and Menu-DA) convey to participants (only) that the mechanism is DA, this mechanism is (also) in fact strategyproof; similarly, while the SP-Property treatments (Menu-SP and Textbook-SP) convey to participants (only) that the mechanism is strategyproof, this mechanism is (also) implemented using DA. More generally, our experimental design tests how participants respond when different facets of a *fixed environment* are described in different ways.

We conduct our experiment (combined $N = 542$; pre-registered) on two different subsamples: US participants recruited on Prolific and participating remotely via live video sessions ($N = 255$), and Cornell students participating physically via in-person lab sessions ($N = 287$). These different settings, as well as populations, are meant to induce variation in focus and attention, as well as cognitive resources and education—all key moderators of our hypothesized mechanisms—to investigate the extent to which they affect our main results. Our main findings replicate across the two sub-samples, hence we pool them in our main analysis.

We highlight four main results. First, participant training scores in the Traditional

DA Mechanics treatment suggest that our training is indeed effective in teaching many participants the mechanics of DA. For example, the last task in our DA Mechanics training modules asks participants to calculate their DA outcome using our GUI completely on their own for a specific DA scenario. Three quarters of participants (76% ; SE = 4%) succeed with no mistakes on first attempt. Of the remaining quarter (24%), who are only informed "Incorrect allocation. Please try again," another third (8 percent of the total) succeed with no mistakes on the next attempt. Crucial to this result is our new GUI and extensive instruction and quizzing of participants, which enables a much more systematic reporting and analysis of participants' understanding levels compared with prior works.

Second, we find that understanding the full, explicit mechanics of DA *does not* imply understanding its strategyproofness. Our SP-understanding tests contain (i) thirteen *abstract* questions asking participants to apply the logical definition of strategyproofness (e.g., "If you ranked the four prizes {A, B, C, D} in order B-D-C-A and received Prize C, is there another ranking that could have gotten you Prize B? Yes/No"); and (ii) five *practical* questions asking participants about the implications of strategyproofness (e.g., "If you want to maximize your earnings, will you sometimes have to rank the prize that earns you the most in second place or lower? Yes/No"). In Trad-DA, mean overall SP-understanding score is 56% (SE = 2%), and in Menu-DA it is 58% (SE = 2%). Despite strategyproofness holding via a one-sentence proof in Menu-DA, the two are statistically indistinguishable, and are very close to the baseline set by the Null treatment, of 54% (SE = 1%). Thus, to the extent that our tests are effective in measuring understanding of strategyproofness, instructing participants on the full mechanics of how their outcome is calculated has (perhaps startlingly) little effect on their SP understanding.

Third, we find that our SP Property treatments, particularly Menu-SP, *can* to some extent teach participants strategyproofness. Participants' mean scores on the SP-understanding tests in Menu-SP is 71% (SE = 2%)—the highest of any treatment. Participants in Textbook-SP (which is based on the classical definition of strategyproofness) have a mean score of 62% (SE = 2%)—the second highest, but significantly lower than in Menu-SP. Investigating the two subsets of SP-understanding questions separately, we find that participants in Menu-SP fare better both in drawing inferences from the abstract logical properties of strategyproofness in hypothetical problems, and in realizing the practical implications of the strategyproofness property for maximizing earnings.[5] We also find that scores on the SP-understanding test, particularly the sub-test with five practical questions regarding maximizing earnings,

[5]For example, we can consider the fraction of participants with perfect or one-less-than-perfect scores, separately for the two sub-tests. In Menu-SP, the respective fractions are 53% (SE = 5%) and 35% (SE = 5%). The next-highest fractions in any other treatment are 28% (in Textbook-SP; SE = 4%) and 19% (in Trad-DA; SE = 4%), respectively.

exhibit a form of all-or-nothing bimodality, suggesting a degree to which participants either consistently "get strategyproofness" or consistently "get it wrong."

Interestingly, despite significant shifts in participants' SP-understanding scores, we do *not* see a similarly significant shift in the distributions of SF play. In particular, the rate of SF play seems surprisingly close to uniformly-distributed across participants in all treatments, and mean rates of SF play across treatments range from 48% in Null to 59% in Menu-SP (SE = 3% in all treatments).[6] Hence, although treatment differences in mean SF rate are largely directionally consistent with mean SP-understanding rates, the overall distributions do not generally suggest large statistical differences between rates of SF play.

Fourth, we find that, despite the relatively similar distributions of rates of SF play across treatments, there is a strong positive relationship, across participants, between SP-understanding and rates of SF play. In particular, there is a step-function-like pattern in the relation between these two variables: at low and mid-rate SP-understanding levels, higher understanding score is not associated with significantly higher SF behavior rates (which are around 50% on average); however, for SP-understanding scores above roughly 80%, SF behavior rates are dramatically higher (roughly 80–100%). Moreover, underlying the findings above, we find that Menu-SP shifts the largest group of participants from the region of low-to-medium SP-understanding and relatively low SF rates to the region of high rates in both measures. When coupled with the form of bimodality we see in the SP-understanding test, this suggests that participants not only *perceive* strategyproofness in a binary way, but that they additionally *act on* their perceptions to play SF at dramatically different rates.

Our results section and supplementary materials include further details and analysis, as well as robustness tests of our findings. Shedding more light on our second main finding above, we show that, beyond average treatment effects, even the individuals with highest DA-understanding levels in our Mechanics treatments do not typically understand strategyproofness well. Related to our third and fourth findings, we investigate SP-understanding patterns in detail and show that among all eighteen SP-understanding questions, the five on practical implications are better predictors of both overall SP understanding and behavior; additionally, we explore whether ranking behavior (beyond the binary SF/NSF distinction) jointly depends on treatment and round-parameters variation, but we do not find meaningful patterns. We also show that our findings are robust to controlling for a rich set of demographic characteristics and session fixed-effects. Specifically, the findings are robust to focusing on only one of the Cornell and Prolific sub-samples, despite their different participant characteristics—while, notably, baseline understanding levels, as well as treatment

[6]Mean rates of SF play are: 56% in Trad-DA, 50% in Menu-DA, 59% in Menu-SP, 53% in Textbook-SP, and 48% in Null.

effects, are lower on Prolific than at Cornell.[7]

Our findings highlight that understanding the mechanics of calculating the outcome of a mechanism—in our case, DA—is very different from understanding that it is strategyproof. While traditional DA descriptions can be effective in teaching the former, i.e., how outcomes are calculated from inputs, we find that they do not make an effective difference in teaching the latter (relative to our close-to-zero-information Null treatment). In contrast, a stripped-down menu description does make an effective difference in teaching strategyproofness.

Our paper proceeds in the usual order. Section 2 outlines our experimental design. Section 3 presents our results. In Section 4, we overview much of the long line of literature in behavioral mechanism design related to and inspiring our work. We conclude in Section 5 by discussing some implications of our findings, for both future research and real-world applications.

2 Experimental Design

We use an information-provision experiment to study the effects of traditional vs. menu descriptions of DA and SP on three outcomes of interest: (i) participants' understanding of the description; (ii) their understanding of strategyproofness; and (iii) their behavior.

Our experiment (a) introduces the matching environment; (b) describes the matching mechanism using one of our five descriptions; (c) attempts to improve, and tests, participants' understanding of the mechanism using novel training questions and/or GUI; (d) elicits behavior in standard, incentivized DA rounds; (e) elicits understanding of strategyproofness using novel tests; and finally (f) elicits reflections, perceptions, cognitive-abilities measures, and a set of demographic characteristics. Only steps (b) and (c) differ across treatments.

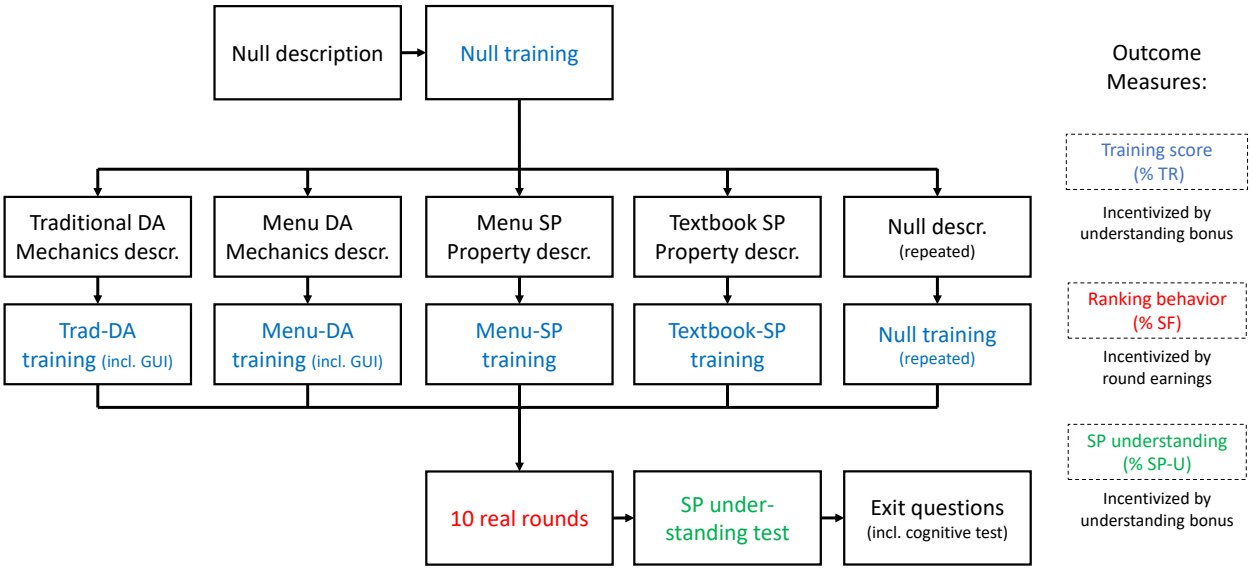
Participants earn money mainly from incentivized rounds of DA in (d) (up to £/$9.90 on Prolific/at Cornell, respectively), but they are also incentivized with a monetary comprehension bonus for all questions they answer correctly on the first attempt in training questions in (a)–(c) and in the SP-understanding tests in (e) (up to £/$4.50). We add these (somewhat nonstandard) comprehension bonuses in order to encourage participants to base their behavior on their understanding of the description (as opposed to, e.g., learning by playing) by increasing participants' attention to and understanding of the experiment and the descriptions, which are quite complicated in some treatments. Participants are shown their cumulative DA-round earnings after each round; their understanding bonus is only

[7]We find similar results when splitting the sample into top vs. bottom scorers in cognitive-ability exit questions and into more vs. less attentive participants as measured by attention tests embedded in our experiment.

calculated at the end of the experiment, based on their fraction of correct answers.

Figure 1 provides a summary of the overall experimental flow across the five treatments: two DA Mechanics treatments, two SP Property treatments, and a Null treatment. DA Mechanics treatments describe the details of the allocation process using either a traditional algorithm (Trad-DA) or menu algorithm (Menu-DA). SP Property treatments convey only the fact that the mechanism is strategyproof—i.e., without specifying the details of the assignment procedure—using either a definition inspired by menu descriptions (Menu-SP) or a benchmark definition inspired by conventional textbook definitions of strategyproofness (Textbook-SP). Our experiment's three main outcome variables are participants' performance on the training questions (% TR), their performance on the strategyproofness understanding test (% SP-U), and their fraction of straightforward ranking behavior (% SF).

Figure 1: Overview of experimental flow.



Our experiment is programmed in oTree (Chen et al., 2016). The rest of this section describes the experiment parts in chronological order. For the full experimental materials with screenshots of all screens in all treatments, see the online Appendix A accompanying this paper.

2.1 Setting (All Treatments)

The experiment starts with consent and introductory screens. Then, participants see a description of the matching environment, i.e., the inputs and outputs of the matching mechanism and how the outcomes will earn them money, without any description of how the

actual matching will be determined. We call this the "Null description."

The Null description presents the matching setting as a static setting with four participants: one human participant (to whom we refer below simply as "the participant" when no confusion can arise) and three computerized participants. This setting is presented as a game in which the human and computerized participants each submit a ranking of the four prizes, and each participant wins one prize. The human participant is informed that the prizes are worth different amounts of money to the different participants, and in each round the human participant is shown how much each prize earns them.

The participant is told that the allocation process depends on their ranking, on the computerized participants' rankings, and on the "prize priorities." The prize priorities (which fill the role of the preferences of the prizes in the matching mechanism) are also shown to the participant before they submit their ranking. The computerized participants' rankings are not shown at any point. See Section 2.4 for information on the distribution of the round parameters (the prize values for the human participant, prize priorities, and computerized participants' rankings). To avoid potentially misleading language, we explain to the participant that computerized participants' rankings are "determined beforehand"—rather than using language that refers to randomization, which may implicitly (and incorrectly) suggest uniform distributions. We highlight to the participant that their ranking cannot affect the priorities or computerized participants' rankings, and provide an intuitive description of what the rankings and priorities mean.[8] (See Figure 5 in Section 2.4 for examples of the information presented to participants in each real round, and the ranking interface.)

After learning about the environment (Null description), the participant plays two practice rounds (Null training) where they gain basic experience with the environment and answer a few basic training questions about the facts they have thus far learned (which count towards their understanding bonus).

2.2 Descriptions (by Treatment)

After completing the basic Null training, the participant is randomly assigned into one of five treatments. Treatments differ in the way that they describe DA Mechanics or the SP Property; these descriptions are summarized in Table 1. For the full content of the main description text in all treatments, see Appendix C within this document.

[8]These intuitive descriptions read as follows. For rankings: "The allocation process tries to give each participant a prize that they ranked higher rather than a prize that they ranked lower, while taking into account the rankings of all participants." For priorities: "The prize priorities can affect the allocation of prizes. The higher your priority is for getting some prize, the more likely you are to get that prize at the end of the process."

Table 1: Expository versions of main description texts, by treatment.

Traditional DA Mechanics (Trad-DA): You will receive a prize according to the following process: [The participant-proposing DA algorithm is then explained in detail.]
Menu DA Mechanics (Menu-DA): A temporary allocation will be calculated using the reported lists of all the participants *except for you*, according to the following process: [The prize-proposing DA algorithm excluding the participant is then explained in detail.] Your Obtainable Prizes are all those where you have higher priority than the temporary match of that prize. You will receive your highest-ranked Obtainable Prize.
Menu SP Property (Menu-SP): Some set of Obtainable Prizes will be calculated using the reported lists of all the participants *except for you*. You will receive your highest-ranked Obtainable Prize.
Textbook SP Property (Textbook-SP): The prize you receive upon submitting a list L is always *at least as high*, according to list L, compared to the prize you would receive submitting another list.
Null: [A repeat of the Null description initially shown to all participants in order to convey the basic setting and environment (see Section 2.1).]

In the Traditional DA Mechanics (Trad-DA) description, the mechanism is described via the participant-proposing DA algorithm, similarly to how it is described in real-world contexts and in previous experiments (see, e.g., NMS, 2020; Chen and Sönmez, 2006). In the Menu DA Mechanics (Menu-DA) description, the mechanism is described via the menu description of GHT (2023). This algorithic description proceeds in two steps: (1) a modified DA algorithm is run (with the proposing and receiving sides flipped, and with the human participant excluded), without using the human participant's ranking, in order to determine a menu of prizes; and (2) the human participant gets their highest-ranked prize out of the menu.

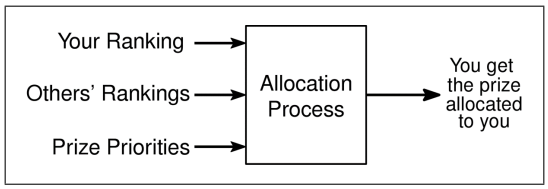
In the Menu SP Property (Menu-SP) description, the participant is informed only that the mechanism is strategyproof, by explaining that *some* menu of prizes will be determined without using the participant's ranking (with no details on how this is done), and the participant will get their highest-ranked prize out of the menu. In the Textbook SP Property (Textbook-SP) description, participants are informed of the mechanism's strategyproofness via a description based on a standard game-theory definition. As in all other description treatments, we use plain English with no mathematical terms; here we explain that ranking prizes according to some list gets you a prize at least as high on that list compared to any different ranking.

Each non-Null description is also accompanied by a chart that conveys either: the fact that the participants' allocation is determined based on their rankings, the computerized

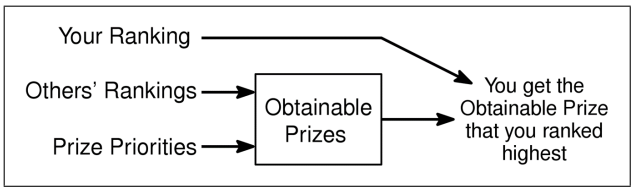
participants' rankings, and the prize priorities (in Trad-DA and Textbook-SP); or the fact that the menu is determined based on the computerized participants' rankings and the prize priorities, and the human participant receives the prize they rank highest among these (in Menu-DA and Menu-SP). See Figure 2 for screenshots.

Figure 2: Charts illustrating to participants the overall matching process.

(a) Trad-DA and Textbook-SP.



(b) Menu-DA and Menu-SP.



Our Null treatment gives no additional information on the allocation mechanism, and in place of such a description, it repeats the Null description discussed in Section 2.1.

2.2.1 Description Rationale

Our treatments are each designed to measure the participant's response to different information. Our DA Mechanics treatments relay explicit, detailed algorithms that determine the participant's allocated prize, with Trad-DA serving as a status-quo approach, and Menu-DA serving as an alternative. Our SP Property treatments relay the definition of strategyproofness, with Menu-SP serving as an abbreviated form of Menu-DA, and Textbook-SP serving as an equivalent, baseline definition of strategyproofness. Our Null treatment relays a basic description with only minimalist information about the allocation process, serving as a broader baseline to compare all other treatments against.

In more detail, our SP Property treatments investigate the participant's responses to learning (only) that the mechanism is strategyproof, but without providing any recommendation on how the participant should rank the prizes. In other words, in both SP Property treatments, we focus on the *definition* of strategyproofness, and avoid providing participants with any *advice* on how participants should rank the prizes (e.g., "from highest to lowest value"). We do this for three reasons. First, our purpose is to investigate, and try to improve, participants' *understanding* (rather than, e.g., their reaction to, or trust in, advice; or their tendency to yield to authority). In particular, we compare a "partial menu description," i.e., one where the details of how the menu is calculated are not conveyed (Menu-SP), to another SP Property treatment (Textbook-SP) that conveys identical mathematical information—that the mechanism is strategyproof. Second, our approach relies on weaker assumptions regarding subjects' preferences; in particular, it avoids the conceptual problem of assuming

that participants wish to play classically predicted, expected-utility-maximizing strategies. Instead, our approach relays a concrete property of the allocation rule; we argue that this approach is worth investigating both experimentally and for potential use in real-world markets as a complement to more conventional, advice-based approaches. Third, importantly, our approach may help mitigate an experimenter-demand effect.

2.2.2 Comparisons of Treatment Descriptions

As each of our descriptions conveys different information, it is challenging to make them look and read similarly. However, we make an effort to increase the similarity between the two DA Mechanics treatments and between the two SP Property treatments as much as we reasonably can. In the DA Mechanics treatments, when describing a DA algorithm containing proposals and rejection, we maintain very similar wording in both DA Mechanics descriptions (other than interchanging the role of participants vs. prizes, and excluding the participant). Still, Menu-DA conveys elements that Trad-DA does not; namely, Menu-DA conveys in addition that the overall process has a menu structure, and also describes exactly how the menu is calculated from the reversed-DA allocation, and how the participant's prize is chosen from the menu. Therefore, Menu-DA is longer.

The two SP Property descriptions are completely different in their wording since they convey different definitions of strategyproofness, but we design them to be as similar as we can in paragraph structure, sentence lengths, and overall description length. Still, as discussed above, Menu-SP describes a concrete two-step outline, while Textbook-SP relays the classical mathematical definition of strategyproofness (using as ordinary a language as we reasonably can).

2.3 Training Rounds and Training Score (by Treatment)

After reading the description, the participant completes a rich set of comprehension exercises that test understanding while also providing further opportunities to learn. The participant's total scores on these exercises are our first main outcome variable, which we term their Training Score, denoted % TR.

Specifically, the participant encounters three training rounds after the description component. The training rounds differ from real rounds in that prizes are not worth money, the computerized participants' rankings are fixed, and the human participant is instructed which particular ranking of the four prizes to submit. We then ask the participant specific questions about this allocation scenario, and the participant receives points towards their

understanding bonus for each question that they answer correctly on their first attempt.[9] During these training screens (and the real-round screens that follow), the participant is always able to open a pop-up window to remind them of the text they saw in the description component.

DA Mechanics training rounds. In the DA Mechanics training rounds, the round is stopped just before the allocation is determined, and the participant is asked to calculate the outcome of DA for themselves. (In Menu-DA, they are also asked to calculate their menu.) The first training round includes a detailed step-by-step walk-through, using a specially developed GUI. It guides the participant in calculating the allocation, based on the DA Mechanics description component they previously saw, providing detailed feedback in each step. The second training round offers the participant an optional video illustrating how to solve the training questions in this round, but the participant must implement the solution without additional hand-holding. The third training round asks the participant to use the GUI to calculate the final allocation, without any guidance. Thus, the successive training rounds increase in difficulty and in the independence required from the participant.

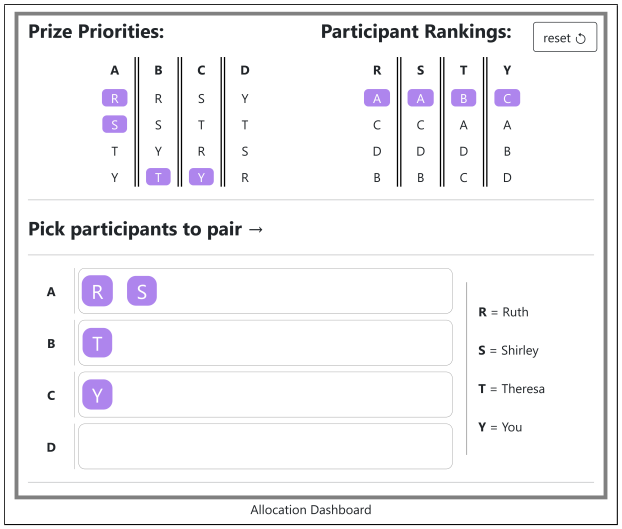
The same DA scenarios (i.e., the human and computerized participants' rankings and the prize priorities) are presented in both Trad-DA and Menu-DA, but the training questions are specifically tailored to the treatment, since the treatments involve different algorithms. That said, these scenarios are chosen so that running each of these algorithms (whether participant- or prize-proposing) would be of comparable complexity. Figure 3 shows screenshots of the DA Mechanics training GUI (along with links to the optional video available to participants in the second training round). In Trad-DA (respectively, Menu-DA), the participant iteratively clicks on the purple labels signifying participants (prizes) to tentatively pair them with prizes (participants), until the final match is calculated.

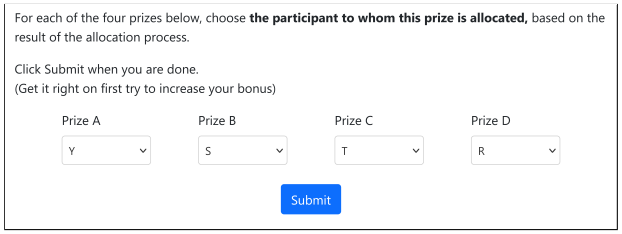
SP Property training rounds. In contrast to the DA Mechanics treatments, in the SP Property training rounds, the participant experiences none of the above allocation-calculation GUI. Instead, after submitting the particular ranking they were instructed to submit, the allocation is calculated behind the scenes by the computer, they are informed of the determined allocation, and the round is stopped right *after* this allocation is determined. Then, the participant is asked a series of multiple-choice questions about the definition of strategyproofness and its implications, based on the SP Property description they received. They are asked a few questions aimed to reveal some common misconceptions about strat-

[9] In the most complicated, multi-step questions of the DA Mechanics treatments, the participant earns 5 points (i.e., the equivalent of five correct answers) toward their understanding bonus for answering correctly on their first attempt, and 2 points for answering on their second attempt.

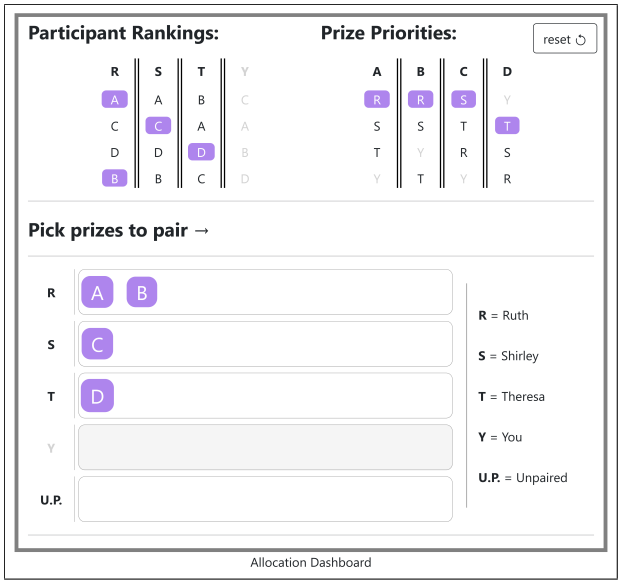
Figure 3: Samples of DA Mechanics training.

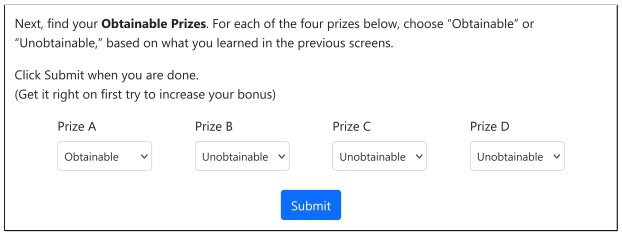
(a) Trad-DA training GUI and allocation entry screen.





(b) Menu-DA training GUI and menu entry screen.

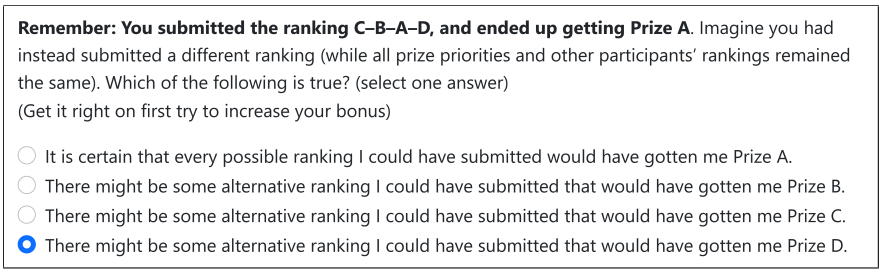




Note: The optional videos used in the second training round (created using this GUI) are available at https://youtu.be/qK9JL32oxJg (Trad-DA) and https://youtu.be/dYePFVdqm5I (Menu-DA).

egyproof mechanisms,[10] and many questions about counterfactual outcomes of the round they played, had they submitted other rankings. See Figure 4 for an example SP Property training question.

Figure 4: A sample SP Property training question.



Following each correct response, the participant gets detailed feedback that explains why this answer is correct. Following each incorrect response to a multiple-choice question, they are asked to try again until they answer correctly. With minimal exceptions, the training questions are the same in Menu-SP and Textbook-SP.[11] However, the feedback following correct answers differs across treatments (in both wordings and illustrative examples) in order to explain the answer in terms of the specific Menu-SP vs. Textbook-SP framing. The answer to each of the identical questions is completely determined by either SP Property description because these two descriptions are mathematically equivalent (Hammond, 1979).

Null training rounds. In the Null treatment, recall that the description of the mechanism is replaced with a repeat of the Null description and two training rounds discussed in Section 2.1.

Training rounds rationale. The main goal of our training rounds is to maximize the participant's understanding of, and ability to apply, the (often complicated) information we present them with. Additionally, we use the participant's training score (% TR) as a measure of understanding. However, due to the way the training questions differ across treatments, % TR is most informative within a treatment, somewhat informative across the two DA Mechanics treatments or across the two SP Property treatments, and uninformative for other cross-treatment comparisons.

[10] These include viewing rankings as able to "insure" against the worst outcome, and viewing a strategy of flipping the first and second prizes in a ranking as increasing the chances of getting the second in case the first is impossible to get.

[11] The exceptions are: (1) the first training question is different in Menu-SP vs. Textbook-SP, because each refers directly to the text of the description; and (2) one later question contains a hint that references the menu in Menu-SP, and references features of the Textbook-SP description in Textbook-SP.

2.4 Real Rounds and Ranking Behavior (All Treatments)

After reading the descriptions and completing all training questions, the participant plays ten rounds of the DA mechanism. Figure 5 provides a screenshot of a completed round. Our second main outcome measure is the fraction of rounds in which the participant ranks straightforwardly, i.e., in highest-earning to lowest-earning order, denoted % SF.

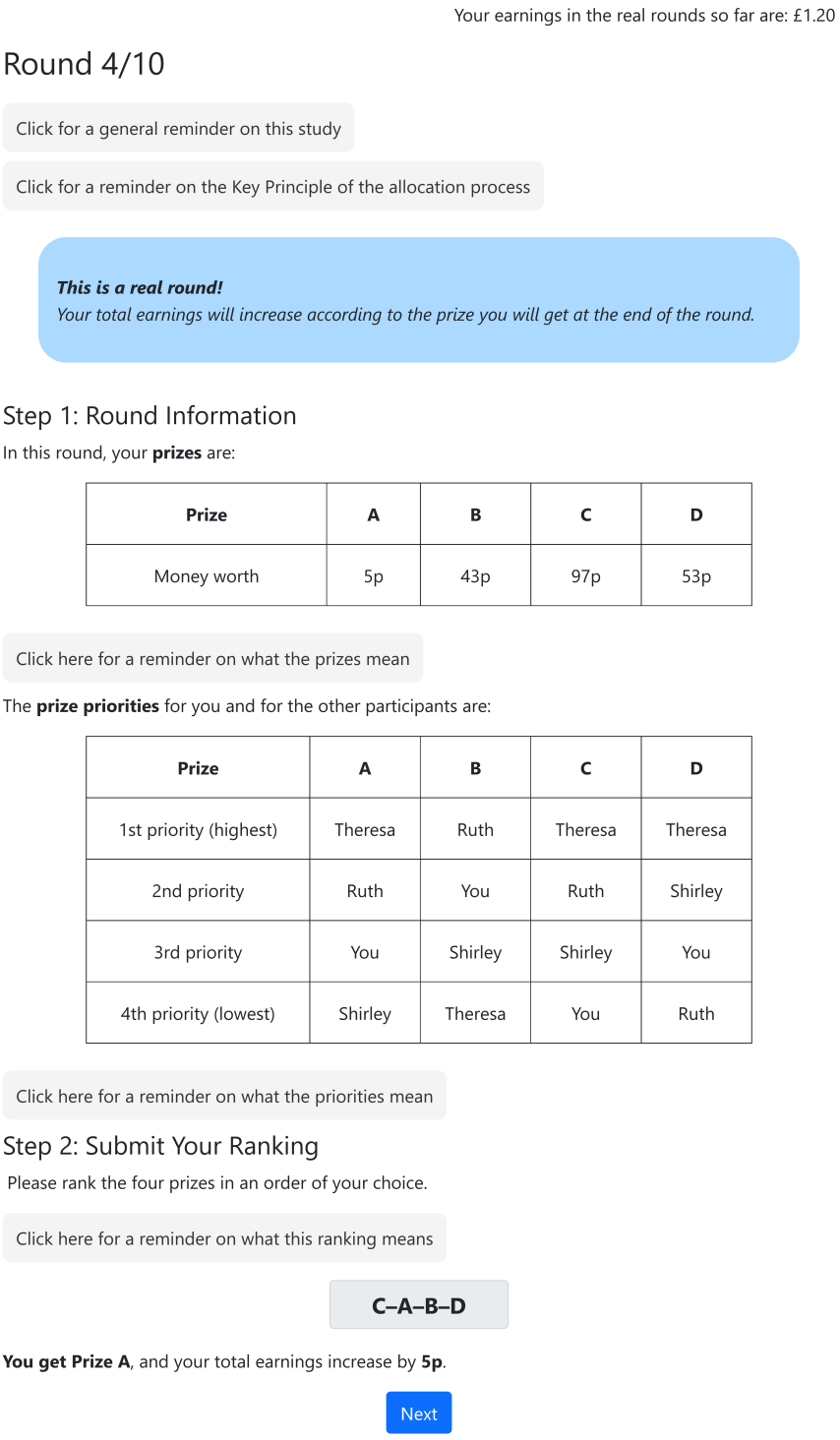
Prize values for the human participant, prize priorities, and computerized participants' rankings are drawn from a specially-tailored joint distribution (for full details, see Section C.3). This distribution is designed to make the setting feel competitive, in order to encourage the participant to carefully consider their strategy and play NSF at rates bounded away from zero and from one—a necessary condition to identify our treatments' effect. For example, the distribution gives the human participant a lower-than-uniform probability of having a high priority at their highest-earning prize, and makes it more likely that the computerized participants rank the human participant's highest-earning prize first. Additionally, the distribution we use often induces large differences between prize values, so that allocation outcome meaningful affect a participant's earnings.

As discussed in Section 2.1, at the beginning of each round, the participant observes each prize's value for them and all prize priorities. We present the prize priorities both to imitate a real-world context in which the participant has an idea about their likelihood to get different outcomes (e.g., in a school-choice context, how well-positioned they are relative to other students), and to leave the participant with the possibility to strategize based on this information if they so choose. At the end of each round, the participant receives limited feedback: they only learn their own outcome (i.e., their assigned prize), with no information on the computerized participants' rankings or outcomes. This is done to minimize learning from experience (through feedback), and hence maximize the effect our different descriptions may have on behavior.

2.5 Strategyproofness-Understanding Test (All Treatments)

After the ten real DA rounds, the participant completes a newly designed strategyproofness-understanding test. Their test score is our final main outcome variable, denoted % SP-U. The test consists of four parts, each on its own screen and containing its own set of questions. For each question the participant answers correctly on the first attempt, they receive 2 points (i.e., the equivalent of two correct answers) towards their understanding bonus. The first three sets are on abstract logical properties of strategyproofness, and we refer to their thirteen questions jointly as Abstract; Figure 6's top part provides examples. They are similar in nature to the comprehension questions asked during the SP Property training rounds, but

Figure 5: Real round (number 4). The participant entered the ranking C-A-B-D and received Prize A.



broader in scope, and common across all five treatments. For SP Property treatments, these can be thought of partially as further measures of participants' understanding of the description (similar to those asked during training rounds), and partially as tests of whether the participant can apply their knowledge in different, novel scenarios. For DA Mechanics treatments, these can be thought of as measures of whether the participant correctly infers (or guesses) the features relevant to the strategyproofness properties from their mechanical description. The final set of questions is on practical implications of strategyproofness, and specifically, how one could maximize their earnings. We refer to this five-questions sub-test as Practical; Figure 6's bottom part provides a complete screenshot.

The goal of our strategyproofness-understanding test is to measure and disentangle the participant's grasp of both the abstract logical properties of strategyproofness, and its practical implications for how an earnings-maximizing participant should rank the prizes.

2.6 Exit Questions and Cognitive Score (All Treatments)

Finally, additional questions elicit reflections on, and perceptions of, the mechanism; measure cognitive abilities and numeracy; and collect demographic and other data.

The reflection questionnaire asks the participant several questions about the strategies they played and about how they perceived the mechanism. Some of these question are adapted from from previous studies (e.g., Rees-Jones and Skowronek, 2018) and some are new to this study. The next questionnaire measures cognitive abilities using the three questions of the Cognitive Reflection Task (CRT; Frederick, 2005), and measures numeracy using one question from the Berlin Numeracy Test. We use this questionnaire to give participants a cognitive score between 0 and 4. Next, a questionnaire collects a long list of demographics, as well as social, economic, and political leanings. The final questionnaire elicits general feedback on the experiment.

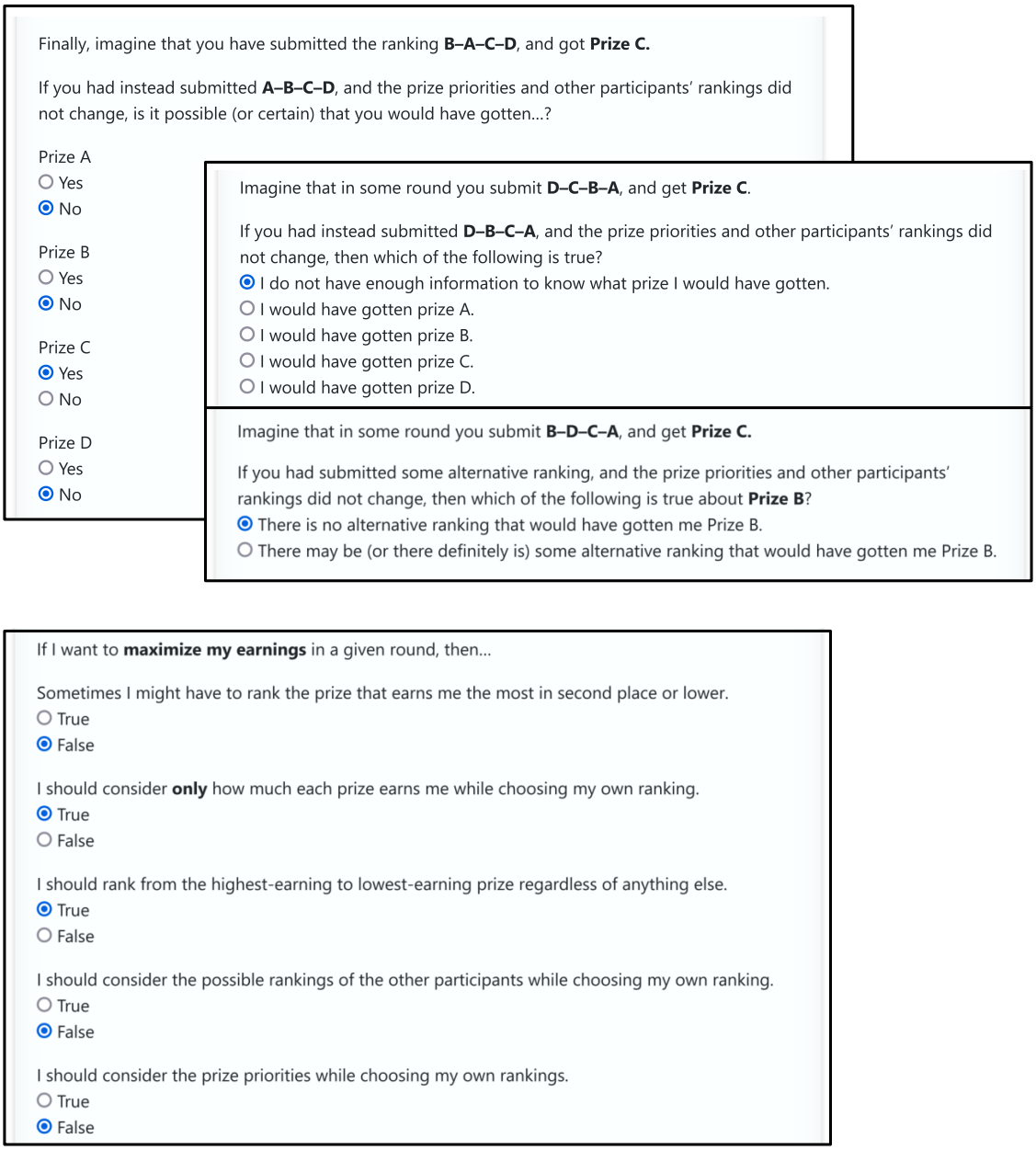
3 Results

3.1 Sample

We collected our data (a total of $N = 542$ participants) from two sources: Prolific ($N = 255$)—a crowd-sourcing platform designed for scientific use—and the Cornell Johnson Business Simulation Lab (BSL; $N = 287$).[12] We use these two different sources to increase

[12]Our pre-registration can be found at https://aspredicted.org/G1V_SBC. According to the pre-registration, we stopped collection at each platform after obtaining at least 50 responses per treatment, resulting in at least 100 responses per treatment overall. For a summary of our pre-registration and com-

Figure 6: Samples of questions from the strategyproofness understanding test.



Notes: The first three boxes show questions directly concerning the abstract logical properties of strategyproofness ("Abstract"), while the fourth shows questions on practical implications, i.e., how participants can maximize earnings ("Practical"). The content of each bordered-box appeared, in top-to-bottom order, on a separate screen. The set of questions in the first box were accompanied by another similar set and by an attention-check question. The question in the second (respectively, third) box was accompanied by one (two) other similar question(s). The last box shows the entire set of questions of that screen (the Practical sub-test).

the variation of factors such as engagement, education level and cognitive skill within our sample, as they may all play a key role in how our descriptions affect understanding and behavior. Prolific participants were recruited from August 3 to August 8, 2023, and Cornell participants were recruited from September 26 to November 17, 2023. For the distribution of demographic characteristics in our sample, see Section B.1.

At Prolific, 291 participants started the experiment and 257 completed it. Of the 34 who dropped out, 27 dropped before reaching any treatment-specific parts—at or before an optional exit point, right after the common Null training, where they could quit for partial payment; we included this exit point to reduce self selection conditional on treatment. We had to drop additional 2 observations that included corrupt DA round data due to technical issues, leaving a final sample of $N = 255$. At Cornell, 296 participants started the experiment and 294 completed it; of the two incompletes, one voluntarily dropped out, and one stopped due to a computer crash. We had to drop additional 7 observations due to computer crashes and other technical problems unrelated to the experiment content, leaving a final sample of $N = 287$.[13]

Recruitment and experiment text were as identical as possible across platforms; changes were limited to differences in currency—Prolific using British pounds (£) and Cornell using US dollars ($)—differences in the fixed participation fee, and that the optional early exit point for partial payment was enabled for Prolific participants only. Prolific prescreening included only participants from the US, with a past Prolific approval rating of at least 99% and at least 50 approved past tasks. In addition, Prolific participants were required to participate in a video-conference meeting with the experiment conductor during the whole experiment. Cornell prescreening included only students, who participated in physical lab sessions. For recruiting materials, a detailed description of the experiment protocol, and more details on the sessions, see Section C.2.

Prolific participants earn an average of £15.0 (a fixed £7 participation fee, an average of £3.0 for TR and SP-U questions, and an average of £5.0 for incentivized rounds). Cornell participants earn an average of $25.5 (a fixed $17 participation fee, an average of $3.4 for TR and SP-U questions, and an average of $5.1 for incentivized rounds). Median completion times differ across treatments, where Trad-DA and Menu-DA take 48 minutes and 53 minutes, respectively, Menu-SP and Textbook-SP take 40 minutes and 39 minutes, respectively, and Null takes 35 minutes. The overall median time in Prolific is 4 minutes higher than in Cornell. For more details on the duration of specific parts of the experiment, see Section B.2.

parison with our analysis and findings, see Section C.1.

[13] In the Prolific sample, of the 9 participants who dropped out after allocation into treatment or whose data was corrupt, 3 were allocated to Trad-DA, 3 to Menu-DA, 2 to Menu-SP and 1 to Textbook-SP. In the Cornell sample, the one participant who voluntarily dropped out was allocated to Trad-DA.

In our main analysis, we pool the two (Prolific and Cornell) sub-samples.[14]

3.2 Training Rounds and Training Score (% TR)

The left panel of Figure 7 shows the average levels across treatments of our first main outcome variable, training score (% TR).[15] The figure, along with others like it throughout this section, includes a "Random" benchmark which indicates what results would have looked like had participants chosen actions uniformly randomly. The training modules differ dramatically between DA Mechanics treatments and SP Property treatments, and hence % TR is only meaningfully comparable within treatment groups.[16] For detailed performance metrics of our participants at the individual question level, see Section B.5.

We make three simple observations. First, we see that mean training scores are far above the Random benchmark in all treatments. Second, Menu-DA seems noticeably more challenging than Trad-DA, likely reflecting the more complicated algorithm used in Menu-DA. Third, participants fare better on the comprehension questions in Menu-SP than in Textbook-SP.

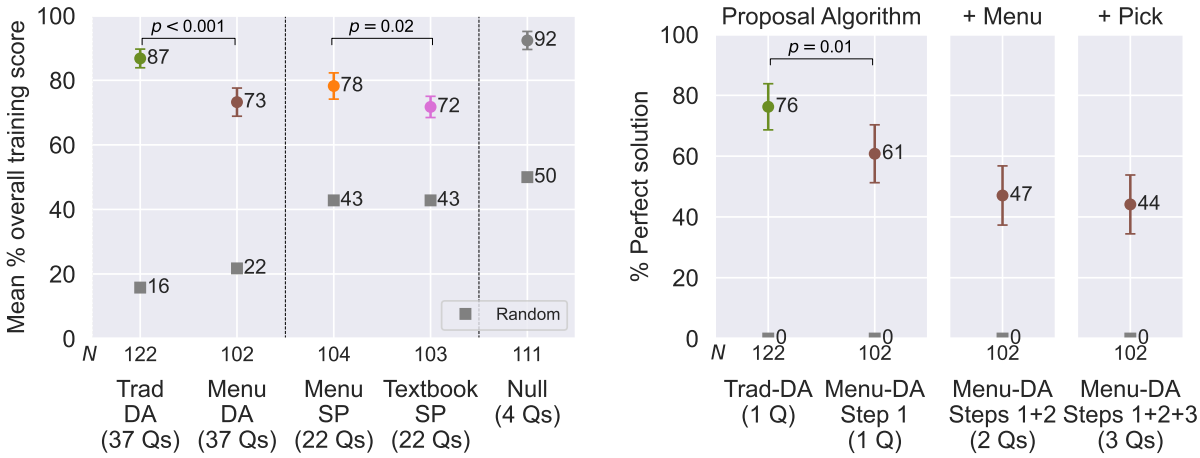
Next, we look deeper into the data regarding our richest training modules, namely, our two DA Mechanics treatments. The right panel of Figure 7 show the means of different sub-measures of % TR for these treatments. These sub-measures focus on our final training round, which requires participants to use our interactive GUI to manually calculate DA outcomes completely by themselves. We track the fraction of participants with perfect scores on this training round. For Trad-DA, this calculation is all done within our training GUI, and we term this outcome "Proposal Algorithm". For Menu-DA, participants first use the same GUI, second calculate their menu from the outcome of the GUI (cumulatively termed "+ Menu"), and third select their final matching by picking from the menu (cumulatively

[14] We investigate between-sample differences in Section B.3. In terms of differences between treatments and relations between important variables, we get similar main results for both sub-samples. The two main differences are: (1) The baseline % TR and % SP-U levels are lower in Prolific and higher in Cornell (2) Menu-SP's effect in moving participants to both high % SP-U and high % SF is less pronounced in Prolific (in accordance with the fewer participants achieving high % SP-U in this sub-sample). In this appendix we also also investigate differences by top vs. bottom cognitive score of our participants (elicited at the end of the experiment) and by top vs. bottom attention scores (elicited in two attention checks planted in the experiment) and find similar results to the Prolific vs. Cornell comparison.

[15] Section B.4 shows that the mean cross-treatment differences of % TR shown in Figure 7, as well as those of % SP-U and % SF shown below in Figure 8 and Figure 9, change very little when controlling for demographic characteristics, cognitive and attention scores, and session-date fixed effects.

[16] As discussed in Section 2.3, while the DA Mechanics training modules in Trad-DA and Menu-DA are designed to be as similar as possible, they still must differ significantly due to the different algorithms involved, with the Menu-DA description being longer and having more steps. The SP Property training is much closer between treatments.

Figure 7: Training Score (% TR) by treatment.



Notes: *Left Panel:* Mean overall % TR by treatment. *Right Panel:* Mean % TR sub-measures, based on manually calculating a DA outcome using our training GUI perfectly in the final training round (relevant to DA Mechanics treatments only). "Proposal Algorithm" indicates the percentage of participants perfectly calculating the DA algorithm. In Trad-DA, this is the final DA outcome. In Menu-DA, this is only the first step; "+ Menu" indicates the percentage of participants who also perfectly calculate the menu based on it; "+ Pick" adds also the last step of perfectly picking the participant's highest ranked prize from the menu. *All Panels:* Error bars: 95% confidence intervals. p-values: two-sided equality-of-means. "N": number of observations. "(# Qs)": the relevant training consisted of # questions, over which the score is averaged. "Random": the expected score of answering every question uniformly at random. Averages are not comparable across vertical dotted lines.

termed "+ Pick").[17] Training scores are lower on the harder Menu-DA questions; however, 61% of participants are still able to calculate the proposal-algorithm outcome by themselves (where we note that getting this correct by random chance is essentially impossible) on their first attempt (with additional 5% succeeding in the second attempt, with no additional feedback), and 44% of participants succeed in addition to calculate their menu and the prize picked from the menu, resulting in a correct final outcome.

In contrast to Menu-DA, in Trad-DA 76% of participants succeed in answering the training question on their first attempt (and an additional 8% on second attempt). We view this as evidence that with sufficient care and coaching, participants can be taught (the mechanical step-by-step method of calculating) the traditional description of DA, something often taken for granted in prior work.

Holistically, these results on % TR for our DA Mechanics treatments convey the first main result from Section 1, namely, that with sufficient care and coaching, we can teach participants the mechanics of DA in its Traditional format (as well as in its Menu format, albeit with larger effort and somewhat less success).

3.3 Strategyproofness Understanding Test (% SP-U)

Next, we look at perhaps our most important outcome variable, % SP-U, i.e., participants' score on the strategyproofness understanding, standardized across all treatments, test conducted after rounds of playing DA. We investigate the full % SP-U measure and its two main components separately.

3.3.1 Full Measure % SP-U

We start by looking at the full % SP-U measure. The top panel of Figure 8 shows the average levels across treatments. This figure shows that Menu-SP, and to a lesser extent Textbook-SP, lead to a noticeable increase in % SP-U. Indeed, from the baseline of 54% (SE = 1%) set by the Null treatment, participants in Textbook-SP move to 62% (SE = 2%), and participants in Menu-SP move to 71% (SE = 2%). In contrast, both DA Mechanics treatments are indistinguishable from each other in % SP-U, while improving understanding levels relative to Random by little, and are essentially indistinguishable from the Null treatment. This is

[17] We recall from Section 2.3 and Gonczarowski et al. (2023) that for Trad-DA, the proposal-algorithm questions have participants execute the participant-proposing DA algorithm, and for Menu-DA, these questions have participants execute the prize-proposing DA algorithm in which proposals to the human participant are omitted. Despite our attempts to keep these DA-algorithm calculations as consistent as possible across DA Mechanics treatments, the lower score of Menu-DA in Proposal Algorithm suggests that participants may find the DA algorithm used by Trad-DA easier to conceptualize and/or execute than that of Menu-DA (even disregarding the + Menu or + Pick components, which are only present in Menu-DA).

true even for Menu-DA, where strategyproofness follows from the description via a simple, one-sentence proof.[18]

To get a more in-depth picture of % SP-U, the middle panel of Figure 8 shows a histogram for the distribution of % SP-U across treatments. We see a noticeably larger fraction of participants get perfect or near-perfect scores in the SP-U test in the Menu-SP treatment, but much smaller differences among other treatments.

These results for % SP-U give the high-level picture of two of our main results. First, based on the overall high levels of of training scores, but low levels of strategyproofness understanding, we conclude that understanding the mechanics of DA does *not* imply understanding strategyproofness (the second main result from Section 1). Second, at the same time, the abstract strategyproofness property can be taught in a way that moves understanding levels upward, as demonstrated by our specially tailored Menu-SP description (the third main result from Section 1). We delve further into these results below.

3.3.2 Sub-Measures of % SP-U

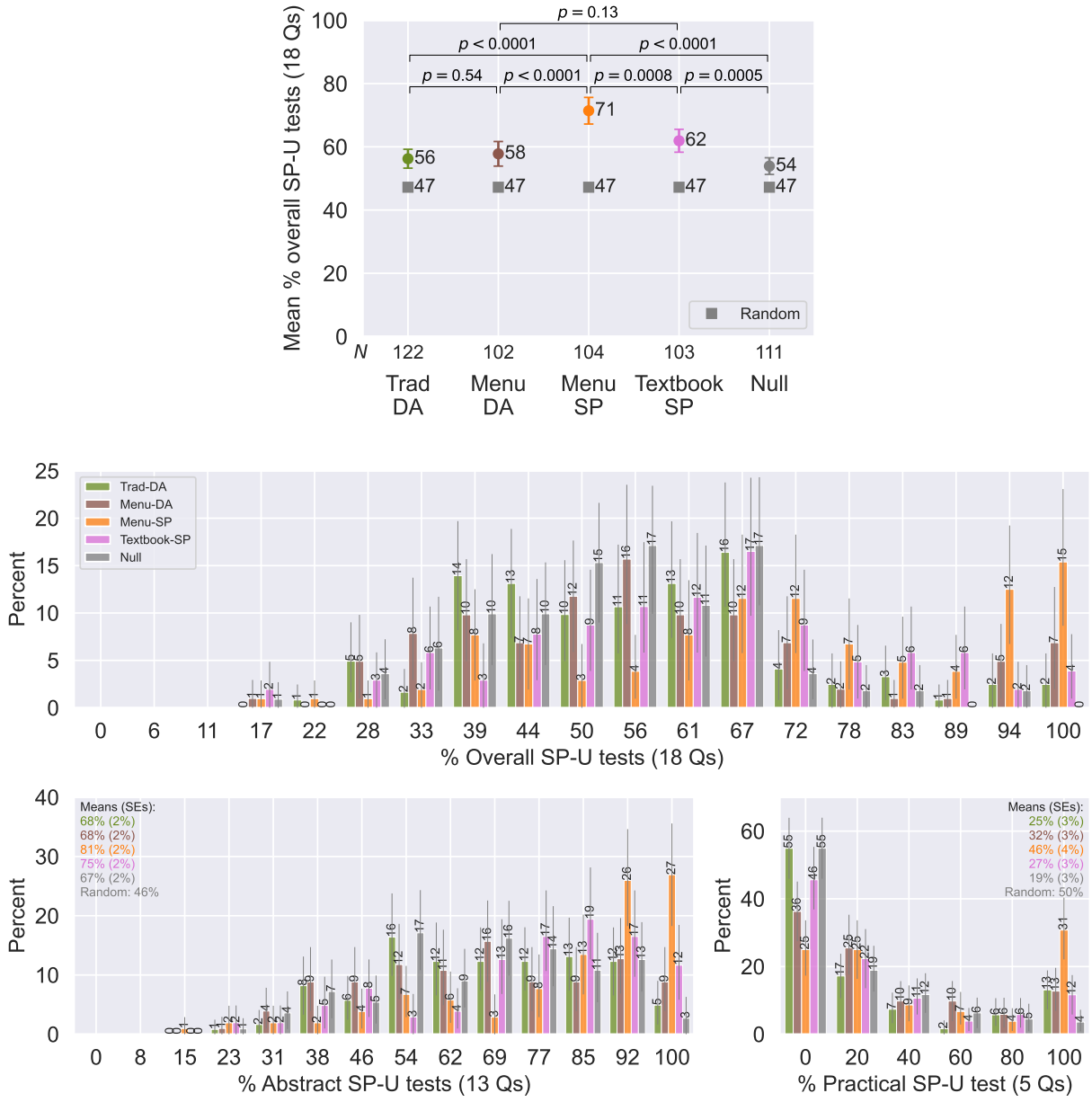
Next, we investigate the distribution of two sub-measures of % SP-U. First, we consider questions directly concerning the abstract logical properties of strategyproofness (which we recall that we refer to as Abstract), and second, we consider practical-implication questions concerning how participants can maximize earnings (which we recall that we refer to as Practical). For sample questions by sub-measure, see Figure 6.

The bottom two panels of Figure 8 report participants' performance on Abstract and Practical separately. Participants' performance in the Abstract tests seems approximately bell-shaped, with a peak that smoothly moves across treatments. In contrast, performance on Practical is strikingly bimodal, with many participants scoring close-to-zero and many other participants scoring close-to-perfect.

The bimodality in Practical may suggest that participants typically have one of two mental models of how one can maximize earnings: one in which they perfectly perceive strategyproofness, and one in which they perceive exactly the opposite of strategyproofness (e.g., that one should consider the prize priorities and use them to strategically report an NSF ranking). Moreover, we observe that in the Trad-DA treatment—which is based upon the traditional description used to explain DA in real-world markets—more than *half* of our participants may have this "opposite-SP" perception of DA—a rate similar to that among Null participants, who are told almost nothing about the mechanism! In contrast, in the SP

[18]Since there are ten possible comparisons of % SP-U across pairs of treatments, p-values may need to be adjusted for multiple testing. Applying an arguably over-conservative Bonferroni correction would multiply all p-values in Figure 8 by 10, maintaining the differences between Menu-SP and other treatments with a statistical significance level of at least $p < 0.01$.

Figure 8: Understanding of SP (% SP-U) by treatment.



Notes: *Top Panel:* Mean of % SP-U by treatment. p-values: two-sided equality-of-means. "N": number of observations per treatment. "Random": the expected score of answering every question uniformly at random. *Middle Panel:* Histogram of % SP-U by treatment. *Bottom Left Panel:* Sub-measure of % SP-U restricting attention to the first three screens of the test, which concerned the abstract logical properties of strategyproofness. *Bottom Right Panel:* Sub-measure of % SP-U restricting attention to the last screen of the test, which concerned practical implications of strategyproofness, i.e., how participants could maximize earnings. *All Panels:* Error bars: 95% confidence intervals. "(# Qs)": the relevant part(s) of the strategyproofness understanding test consisted of # questions, which the score is averaged over.

Property treatments, and Menu-SP in particular, participants move higher in both Abstract (recalling and applying the strategyproofness property we taught them in novel scenarios) and Practical (drawing implications from the strategyproofness property regarding how they can maximize earnings).

In Section B.6, we also investigate the joint distribution between Abstract and Practical, and how these sum to % SP-U. Our main finding on this topic is that the variation in % SP-U is largely driven by different sub-measures for different regions of the % SP-U distribution. Namely, among observations with % SP-U $< 75\%$, Practical is typically low—approximately at its "opposite-SP" perception mode (with an average of 17%)—and variation in % SP-U is mostly determined by Abstract. In contrast, when % SP-U $\geq 75\%$, Abstract is close to its maximal value and changes only by little, while the bimodality in Practical drives the weaker form of bimodality seen in the full distribution of % SP-U, namely, the "dip" between 78% and 89%. At higher levels of % SP-U, Practical has already shifted to its "correct SP" perception mode; for instance, Practical's average in the % SP-U $\geq 75\%$ range is 83%.[19,20]

Holistically, our results on the sub-measure of % SP-U enhance main result (3) from Section 1, by showing that Menu-SP moves participants understanding higher according to both measures.

3.4 Real Rounds and Ranking Behavior (% SF)

Figure 9 shows the mean levels and distribution of our final main outcome measure, rate of straightforward play (% SF). Perhaps surprisingly, % SF seems roughly uniformly distributed in all treatments, and is never particularly high on average. The baseline % SF rate is 48% in the Null treatment. In the Menu-SP treatment, in spite of the significant increase in % SP-U we observe in Section 3.3, the mean of % SF is still only 59%, and only slightly above the Trad-DA treatment which has mean 56%. Menu-DA and Textbook-SP remain even lower at 50% and 53% respectively—statistically indistinguishable from Null. (SE = 3% for % SF in all treatments.) This gives our result, discussed in Section 1, that global treatment effects on the distribution of behavior are small.

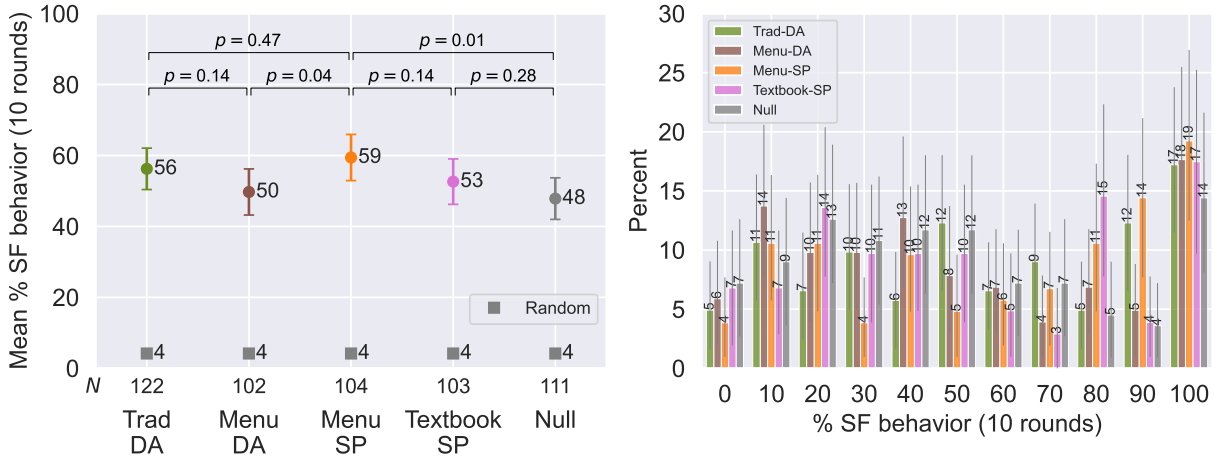
As in most other DA experiments, our main measure of behavior is % SF. In Section B.7, we look for patterns in participants' ranking behavior beyond playing SF vs. NSF, and for patterns of % SF over the 10 played rounds. By far the most common NSF ranking flips

[19]We conjecture that this pattern corresponds to a more general feature of participants understanding of strategyproofness, namely, that instructing participants on the strategyproofness property to an intermediate level of understanding does not "partially" teach them how to maximize earnings; rather, (some) participants realize how to maximize their earnings only when they understand the strategyproofness property quite well.

[20]We remark that the specific value of (approximately) 75% seems to be largely due to the relative scales of the sub-measures—13 questions for Abstract, and 5 for Practical.

the first and second prizes, but we do not find additional strong trends in ranking behavior beyond those conveyed by % SF. In particular, while most participants play SF in some rounds and NSF in others, the specific choice of when to play NSF does not seem to depend strongly on parameters of the specific round of DA.[21] In addition, we do not find strong trends of % SF over the 10 rounds in any treatment, suggesting weak effects of learning from playing.

Figure 9: Straightforward ranking behavior (% SF) by treatment.



Notes: Left Panel: Mean of % SF by treatment. Error bars: 95% confidence intervals. p-values: two-sided equality-of-means. "N": number of observations per treatment. "Random": the expected % SF from ranking the prizes uniformly at random. Right Panel: Histogram of % SF by treatment.

3.5 The Relations Between % TR, % SP-U and % SF

We now investigate the joint distributions of our main outcome variables. Throughout our analysis, we explore the conceptual causal flow—in spite of our empirical *correlational*, rather than causal, evidence—from understanding the description (% TR) to understanding strategyproofness (% SP-U) to some participants' straightforward play (% SF).[22]

[21] For example, we find that across all treatments, participants, and rounds, the fraction of times the participant ranks their highest-earning prize first is 40%. In rounds where the participant has highest priority of getting their highest-earning prize, this fraction is not much lower, at 34%, and we see little variation in these fractions across treatments. We also find little variation in ranking patterns conditional on the difference in money value between the highest-earning prize and the second-highest-earning prize.

[22] Analysis in this section is mostly exploratory and not pre-registered. However, we consider the result on the global relation between % SP-U and % SF as a fourth main result of the paper due to the strong pattern it indicates, and its robustness to different treatments and sub-samples.

3.5.1 Correlations

Table 2 shows three OLS regressions using our main outcome variables: % SP-U and % SF regressed on % TR, and % SF regressed on % SP-U, as well as the correlation coefficients between these pairs. Focusing on the non-Null treatments, we make three observations motivating further investigation.[23] First, higher training scores are associated with higher SP understanding. This is unsurprising for the SP Property treatments—since these treatments deliberately teach SP and train on questions similar in nature to those in the SP-U tests—but perhaps more surprisingly suggests that despite the overall low % SP-U in our DA Mechanics treatments, they do have some ability to convey SP. We investigate this in Section 3.5.2 below and find that the detailed relation looks weaker than implied by the linear regression.

Table 2: Regression and correlation coefficients between our main outcome variables, (non-causally) investigating the hypothetical chain of influence % TR $\mapsto$ % SP-U $\mapsto$ % SF.

Indep.	Dep.		Trad DA $N = 122$	Menu DA $N = 102$	Menu SP $N = 104$	Textbook SP $N = 103$	Null $N = 111$
% TR	% SP-U	β	0.34 (0.08)	0.36 (0.07)	0.70 (0.10)	0.77 (0.08)	0.09 (0.07)
		r	0.35 (0.09)	0.38 (0.08)	0.73 (0.10)	0.80 (0.08)	0.10 (0.07)
% SP-U	% SF	β	0.61 (0.15)	0.73 (0.14)	0.73 (0.14)	0.64 (0.16)	0.25 (0.22)
		r	0.24 (0.06)	0.28 (0.05)	0.28 (0.05)	0.25 (0.06)	0.10 (0.08)
% TR	% SF	β	0.38 (0.19)	0.38 (0.15)	0.58 (0.17)	0.50 (0.18)	0.07 (0.18)
		r	0.15 (0.08)	0.15 (0.06)	0.23 (0.07)	0.20 (0.07)	0.03 (0.07)

Note: β: Estimated coefficients using OLS regressions of Dep. var on Indep. var, within each of the five treatments separately. r: Estimated Pearson regression coefficients. Robust standard errors in parentheses.

Second, the relation between % SP-U and % SF seems quite consistent across all (non-Null) treatments, suggesting that understanding SP better corresponds to more SF behavior, regardless of the description. We investigate this in Section 3.5.3 below and find that a strong pattern between high SP understanding and high SF-play rates underlies the average linear relation.

Third, the relation between % TR and % SF seems overall weaker than the other relations in the table. This is consistent with the idea that this relationship is a noisy composition

[23] Null coefficients are all either statistically indistinguishable from zero, or nearly so. For rows involving % TR, this likely reflects the fact that the Null training questions are very easy and hence have low variation. For the Null % SP-U vs. % SF row, this is consistent with the hypothesis that Null % SP-U variation originates mostly from noise, since Null provides no useful information for understanding SP.

of those from first two rows of the table, and is hence consistent with our conceptual chain in which training affects behavior only through SP understanding (and given the noise in the % SP-U vs. % SF relation). A more detailed view on this relation also supports this interpretation; we defer these results to Section B.10. This suggests some degree of explanation for our second main finding from Section 1, namely, that teaching participants DA Mechanics does not imply teaching them strategyproofness.

3.5.2 The Detailed Relationship Between % TR vs. % SP-U in DA Mechanics Treatments

As motivated by Table 2, we investigate the relation of % TR and % SP-U within the DA Mechanics treatments. Figure 10 shows the full joint distribution scatterplot of training score and SP understanding within these treatments. The figure displays the regression line from Table 2, and also contour lines estimating the smoothed distribution. It also splits each treatment's distribution into four quadrants, depending on whether participants are above or below a certain threshold in % SP-U and in % TR. Motivated by our findings in Section 3.3.2, we choose a threshold of 75% in % SP-U, and we arbitrarily use the same threshold in % TR as well.

The main takeaway from Figure 10 is that in both DA Mechanics treatments, the fraction of participants with high % TR and high % SP-U is fairly small. This within-treatment result adds additional support for our between-treatments result (2) from Section 1. Even a high level of understanding DA Mechanics descriptions, conditional on getting them, does not imply understanding of strategyproofness.

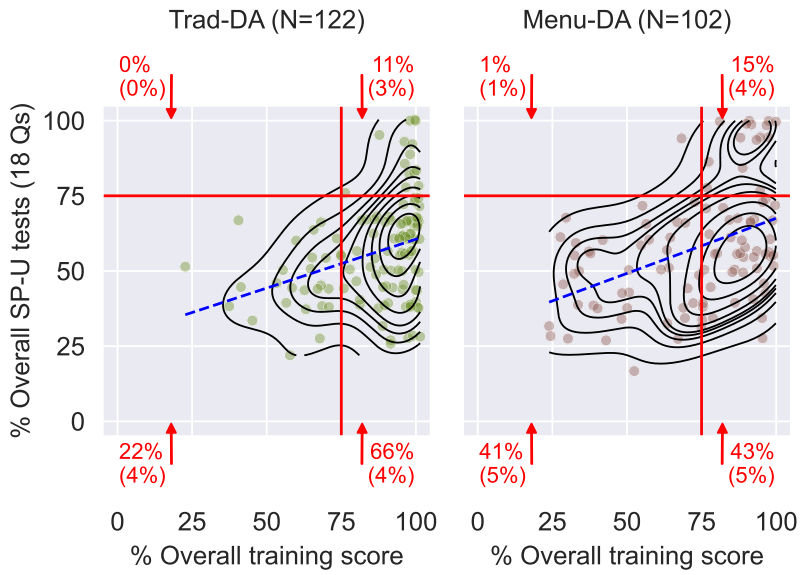
3.5.3 The Detailed Relationship Between % SP-U and % SF

Finally, we investigate in more detail the joint distribution of % SF and % SP-U, beginning with Figure 11. The bottom panel reports the histogram of % SP-U by treatment (replicating the middle panel of Figure 8). The top panel of Figure 11 shows mean % SF, by % SP-U-score value (from the bottom panel).

The top panel of Figure 11 shows that mean % SF "jumps" to noticeably high levels once participants score sufficiently high in % SP-U, with a threshold around roughly 75–80% in the SP-U tests.[24] This result suggests the fourth main result from Section 1: that those who grasp strategyproofness and its implications very well play SF much more than those who do not. We next make a few observations regarding this relation.

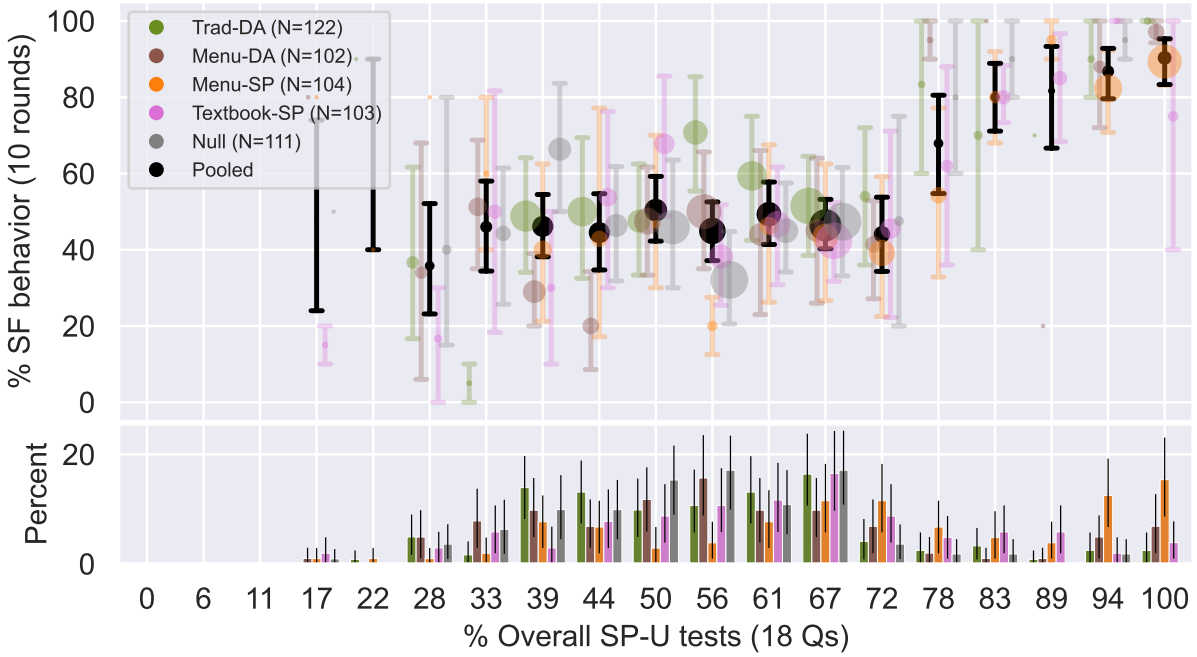
[24] Section B.4 shows that the step-function shape of the % SF vs. % SP-U relation, as well as the magnitude of the "jump" in % SF in that relation, change by little when controlling for demographic characteristics, cognitive and attention scores, session-date fixed effects and treatment indicators.

Figure 10: Joint distribution of % TR and % SP-U for our DA Mechanics treatments.



Notes: Each panel contains a jittered scatter plot of the two measures and estimated contours smoothing the two-dimensional distribution. *Dashed blue lines*: Predicted % SP-U values according to OLS regression results from Table 2. *Solid Red lines*: % TR = 75% and % SP-U = 75% lines. *Red numbers*: fractions of participants in the four regions separated by the solid red lines (with SE in parentheses). For a version of the figure including the SP Property and Null treatments, see Figure B.27.

Figure 11: Relationship between % SP-U and % SF.



Notes: *Bottom Panel:* Histogram of % SP-U by treatment. *Top Panel:* Mean % SF by treatment and by % SP-U-score value from the bottom panel. The circular marker in the top panel is proportional in diameter to the corresponding fraction from the bottom panel. *All Panels:* Error bars: bootstrapped 95% confidence intervals.

First and importantly, the relation seems global and not treatment-specific, supporting the conceptual-chain view that better understanding of SP may increase SF behavior of some participants, regardless of the description or training that improved SP understanding.[25]

Second, the combination of this pattern with the % SP-U distribution in the lower panel (also indicated by marker sizes in the top panel) suggests that of our five descriptions, Menu-SP is the most effective at pushing participants to the high % SP-U, high % SF region. The full, un-binned joint distribution of % SP-U and % SF, shown in Figure 12, further illustrates this by displaying the fractions of participants falling into four quadrants in the full joint distribution of % SP-U and % SF (defined using the 75% SP-U cutoff from Figure 11 and a similar cutoff of 75% to indicate high vs. low SF rates). 32% (SE = 5%) of participants in the Menu-SP treatment are in the upper right quadrant of high SP understanding and SF rates, while in all other treatments this fraction is 17% (SE = 4%) or less.[26]

Third, these findings suggest a relation between ranking behavior and the bimodality of SP understanding. Figure 11 shows that the specific % SP-U level at which % SF starts to dramatically increase is about 75%—the same level at which the bimodality of SP understanding kicks in (see Section 3.3.2). Figure 12 shows in fact a form of two-dimensional bimodality, which seems consistent across the treatments (albeit being strongest in Menu-SP). In other words, these results suggest that not only that participants' *understanding* of SP is bimodal—"SP-perceiving" or "opposite-SP-perceiving"—they additionally *act* on these two disparate modes of understanding in two very different ways—either using a variety of ranking strategies with a mean % SF around 50%, or playing SF almost exclusively.

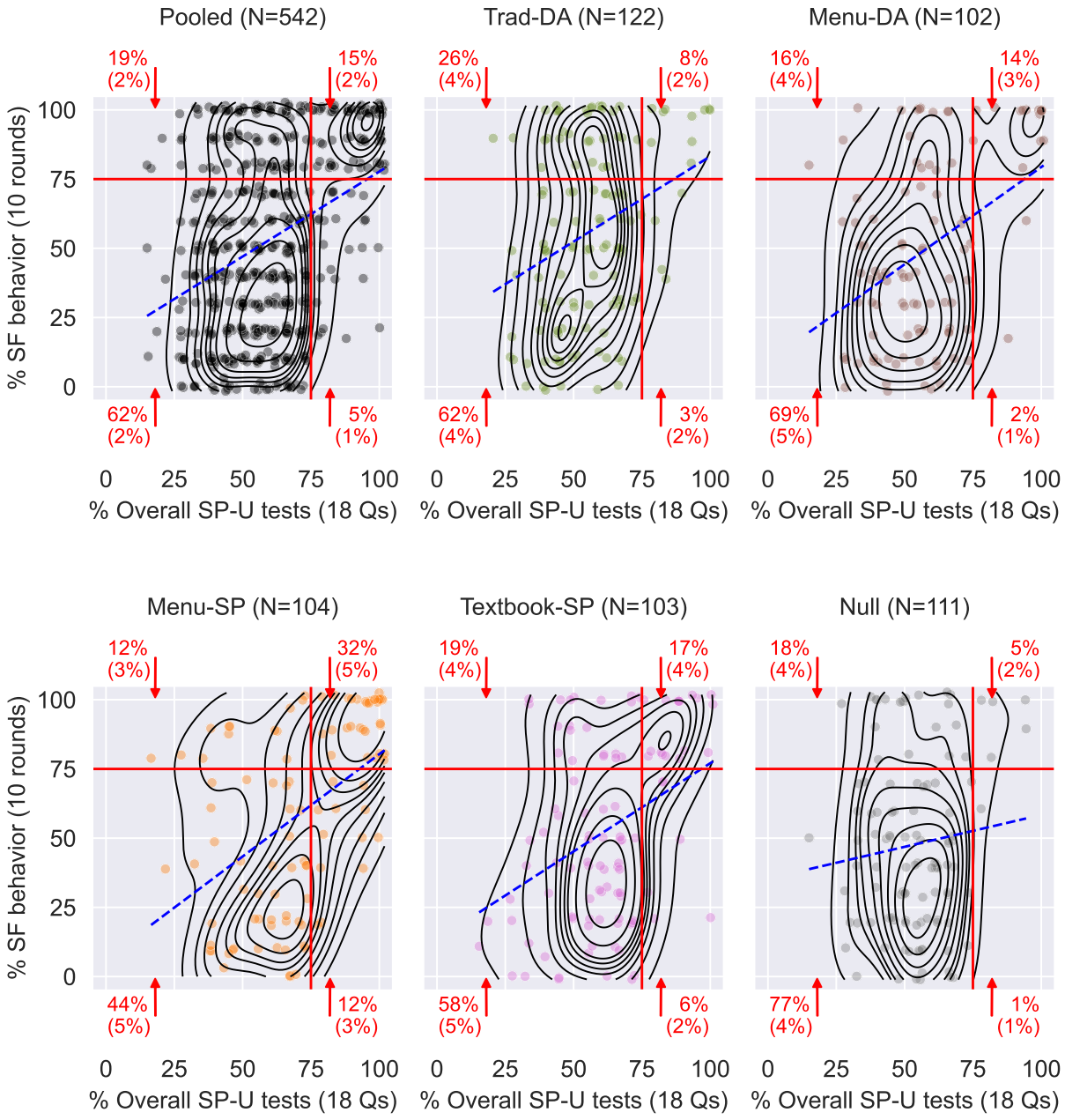
% SP-U Sub-Measures and % SF In Section B.9, we investigate how the Abstract and Practical sub-measures of % SP-U differentialy contribute to the overall % SP-U vs. % SF relation shown in Figure 11 and Figure 12. We find that the Practical sub-measure can by itself explain most of the overall relation in Figure 11, and in particular, it effectively separates participants into high-SF-playing and low-SF-playing based on a high vs. a low test score. However, we find that including Abstract may increase the sharpness of this

[25] An alternative interpretation, contrary to our conceptual chain, is that participants who play SF (for unclear reasons) simply reflect this in their SP-U test responses, e.g. by scoring highly in Practical. However, we see that among the participants with high % SF, the fraction of them with high/low % SP-U differs quite a bit by treatment, which runs contrary to this interpretation.

Of course, given our correlational data, we cannot rule out all alternative non-cause interpretations. For example, it's possible that SP-U-improving descriptions are mostly effective for those participants who tend to play SF in the first place. However, this seems somewhat implausible. For example, we see in Section B.3 that many different populations (e.g., low cognitive scores, and the Prolific sample) *can* experience upwards movement in % SP-U in our Menu-SP treatment, so it is not clear why people who play low SF would *not* experience this.

[26] As mentioned in footnote 14, this effect of Menu-SP is stronger among participants with top (as opposed to bottom) cognitive or attention scores, and among Cornell (as opposed to Prolific) participants.

Figure 12: Full joint distribution of % SP-U and % SF, along with fractions of participants in four quadrants of the distribution, and linear-regression lines.



Notes: Joint distribution of % SP-U and % SF by treatment. Each panel contains a jittered scatter plot of the two measures and estimated contours smoothing the two-dimensional distribution. *Dashed blue lines*: Predicted % SF values according to OLS regression results from Table 2. *Solid Red lines*: % SP-U = 75% and % SF = 75% lines. *Red numbers*: fractions of participants in the four regions separated by the solid red lines (with SEs in parentheses).

separation, especially among participants who score in the middle range of Practical.[27]

4 Related Literature

Our paper sits most directly within the literature on laboratory experiments in behavioral elements of mechanism and market design. Excellent reviews can be found in Hakimov and Kübler (2021); Rees-Jones and Shorrer (2023). Early works within this literature compare behavior across different social choice rules, e.g., Kagel and Levin (1993); Chen and Sönmez (2006); Pais and Pintér (2008); Featherstone and Niederle (2016), among others. More recent papers often consider a fixed social choice rule, and vary some feature of how the rule is implemented. Our paper focuses on varying the framing and information provided to participants, while keeping the rounds of play identical across treatments. We now discuss and compare to the most directly-related works and themes of the literature.

The most-related paper to ours is the behavioral mechanism design paper Katuščák and Kittsteiner (2020). Like our study, that paper conducts an experiment with a menu description of a matching mechanism, (namely, Top Trading Cycles, henceforth TTC).[28] Beyond differing in the mechanism used, that paper's experiment has only behavior as an outcome variable, while our focus (and hence that of our four main results) is on the interplay between training score, SP Understanding, and behavior; that paper does not conduct training, and does not test SP Understanding separate from behavior. That paper finds higher rates of SF play in their menu treatment of TTC, especially for participants with higher cognitive-ability scores.

Training. Many papers on behavioral market design conduct pen-and-paper training quizzes that ask participants to solve simple instances of DA, but either do not record participants' performance on such quizzes, or do not report any analysis based on such quizzes. All exceptions of which we are aware give a single pen-and-paper task without providing feedback or coaching on the descriptions.[29] Our training GUI extends these papers' descriptions

[27] Among participants whose Practical scores are between 40% and 60%, the average SF rate moves from 49% (SE = 4%) to 80% (SE = 7%) when Abstract changes from below 90% to above 90%.

[28] In more detail, and phrased in the language of our paper, one treatment of Katuščák and Kittsteiner (2020) is a Traditional (TTC) Mechanics treatment, and one is a Menu SP Property treatment with an optional supplemental appendix given to participants, which specifies the details of one form of Menu (TTC) Mechanics.

[29] These exceptions include Guillen and Hing (2014), Bó and Hakimov (2020), and Guillen and Veszteg (2021), which each report some analysis based on participants' performance in a single-question training quiz. Guillen and Hing (2014) study TTC, and report that roughly half of participants complete the quiz correctly, but that performance on the quiz does not correlate with behavior. Bó and Hakimov (2020) study DA, and report a positive correlation between answering the quiz correctly and SF play. Guillen and Veszteg

and their training tasks, and formalizes them into a computerized interface with hand-held coaching and feedback. The recent paper Serizawa et al. (2024) conducts an experiment in a strategyproof auction setting with two types of quizzes depending on the treatment; in only one treatment, their quiz asks participants how they can maximize their earnings, constituting a "hint" towards strategyproofness. They find this quiz "with hint" increases SF play substantially above the baseline.

Advice. Many papers explore advice that explicitly recommends participants play straightforwardly. Strategic advice is conceptually related to our SP Property treatments, however, our work departs from prior ones in that we choose to describe the strategyproofness property, and avoid providing any explicit advice. Prior works that do give advice in matching mechanisms include Guillen and Hing (2014); Guillen and Hakimov (2018). Masuda et al. (2022) conduct similar experiments in a second-price auction, while aiming to reduce the experimenter-demand effect (by prefacing the advice with a disclaimer indicating that the advice may or may not be true, and the participant should freely choose whether or not to follow the advice). Most works studying advice find that it is helpful, but does not eliminate NSF behavior. Our approach, by focusing on the strategyproofness property, facilitates more direct comparisons across our treatments, and provides an approach to reducing the experimenter-demand effect (complementary to Masuda et al. (2022)).

Framing and information provision changes. Various papers investigate framing effects in mechanism design settings in ways different from menu descriptions. Breitmoser and Schweighofer-Kodritsch (2022) study ascending-price-based framings of both static and dynamic auctions, and find that some framings of static auctions in terms of dynamic ones can influence behavior nearly as much as actually implementing the mechanism via dynamic auctions. Guillen and Veszteg (2021) study the theoretically-trivial change to DA of reversing the order in which participants submit their rankings, and find that a lower fraction of participants submit (reversed) straightforward rankings. Danz et al. (2022) study different information treatments in belief elicitation mechanisms, and find that providing more information (on the elicitation rule's mechanics or even strategyproofness property) can *decrease* rates of straightforward behavior. A similar insight was also achieved for the Top Trading Cycles (henceforth TTC) mechanism in Guillen and Hakimov (2018), who find that telling participants only that the mechanism is strategyproof (via an advice-style description) yields higher SF play rates than telling them that it is strategyproof *and* specifying

(2021) study TTC and DA, and report that performance on the quiz does not seem to strongly change under the framing change of reversing the order in which each participant's preference list is considered by the algorithm.

the details of the mechanism (via a mechanics-style description). In non-mechanism-design contexts, Esponda and Vespa (2023) show that some classical anomalies of behavioral economics can be mitigated by framing changes that emphasize that decisions of participants only affect their payoffs in certain contingencies; these alternative framings are distantly reminiscent of menu descriptions.

Behavioral mechanism design more broadly. Other papers propose behavioral models of behavior or confusion in strategyproof mechanisms, and investigate these models empirically or theoretically. Dreyfuss et al. (2022b,a); Meisner and von Wangenheim (2023) study loss aversion in matching mechanisms and DA in particular. The designs of these papers typically give participants more information on their chances of getting different prizes, since their behavioral theories concern participants' beliefs about the expected earnings. Since we focus here on participants' misunderstandings of strategyproofness, we provide less explicit information on probabilities of getting different prizes compared to other works, and instead focus on explicit information about the DA algorithm. Bó and Hakimov (2020, 2023) study interactive mechanisms that ask participants to repeatedly indicate their favorite object from some set of still-available objects, and find that such mechanism can increase SF play.[30] Kloosterman and Troyan (2023) study how participants' values for their outcomes in a matching mechanism depend on their submitted ranking, using a novel real-goods experimental design.

Our work is indirectly inspired by a vast literature of empirical papers that study real-world implementations of matching mechanisms; see Pathak (2017) for a review. Most related to our work are those that study how real world participants perceive strategyproofness—e.g., Hassidim et al. (2017); Shorrer and Sóvágó (2017); Rees-Jones (2018); Hassidim et al. (2021)—or other features of the mechanism—e.g., Robertson et al. (2021); Arteaga et al. (2022); Grenet et al. (2022), among others.

Finally, a substantial theoretical literature has arisen in recent years concerning interpretability of strategyproofness. A major strand in this literature concerns a theoretical notion of simple-to-play interactive mechanism termed (strong) obvious strategyproofness Li (2017); Pycia and Troyan (2023).[31] As discussed, our Menu DA Mechanics treatment is directly based on the mechanism design theory paper Gonczarowski et al. (2023), and our

[30] Bó and Hakimov (2023) consider general interactive mechanisms where participants select objects from among some set. They call these sets "menus"; we note that these are different from the notion of menu we consider (since, for example, these sets change throughout the run of the mechanism).

[31] The literature studying obviously strategyproof matching mechanisms includes Ashlagi and Gonczarowski (2018); Bade and Gonczarowski (2017); Troyan (2019); Mandal and Roy (2021); Thomas (2021). While classifying these mechanisms is often an intricate theoretical task, the main finding of this literature is that obviously strategyproof matching mechanisms do not exist in many settings.

Menu SP Property treatment is conceptually inspired by their menu descriptions framework more broadly.

5 Conclusion

Describing designed markets such as DA to participants is a challenging task. For one, a challenge arises from the fact that calculating matches in DA requires a detailed combinatorial algorithm. We view the relatively high training scores in all our treatments as a successful indication that both the Mechanics of DA, as well as the strategyproofness property, can be taught to participants with careful instruction. Building on the tools we have developed, such as our experimental flow, novel training GUI, and strategyproofness understanding tests, future work may further refine methods enabling participants to gain familiarity with DA and other strategyproof mechanisms.

We view the relation we find between participants' understanding of strategyproofness and straightforward play as an important contribution of our work. Other recent works have suggested that NSF play may be partly intentional, for example, as a strategy played by loss-averse participants in order to avoid disappointment (Dreyfuss et al., 2022b,a; Meisner and von Wangenheim, 2023). We believe it is *also* important to investigate the extent to which NSF play may be the result of simple misunderstandings of strategyproofness. The step-function-like behavior illustrated in Figure 11 may shed some light in this direction: At a certain high-enough level of strategyproofness understanding, participants indeed begin to play SF at quite high rates.

Our paper leads to a simple, natural policy suggestion: describe strategyproofness to participants in a manner similar to our Menu SP Property treatment, likely in combination with the more-conventional "advice" based approach.[32] The Menu DA Mechanics description of Gonczarowski et al. (2023) was a crucial step in our formulation of this suggestion, and facilitated our rich experimental investigation into how participants respond to detailed mechanical descriptions of DA; however, our results suggest that mechanical descriptions may remain too complicated to convey strategyproofness effectively. In contrast, Menu SP Property simply informs participants of a concrete property satisfied by the matching algorithm: Each participants' allocation will always be their highest-ranked choice among

[32] To be concrete, one conceivable description in a school-choice setting which combines Menu SP Property and advice is: "Your school will be assigned using a special algorithm. First, some set of schools where you earn admission will be computed without using your ranking. Then, you will be matched to your highest-ranked school where you are admitted. You may wish to rank the schools from highest-quality to lowest-quality according your own individual needs and wants, since this will not affect your admission to any school, and will always be the ranking that matches you to the maximum-quality school where you were admitted."

some set of possibilities which their ranking cannot influence. Future work may develop such an SP Property description further, study real-world deployments, or study broadly different new approaches to conveying other important properties of mechanisms, such as fairness.

References

Felipe Arteaga, Adam J Kapor, Christopher A Neilson, and Seth D Zimmerman. 2022. Smart matching platforms and heterogeneous beliefs in centralized school choice. *Quarterly Journal of Economics* 137, 3 (2022), 1791–1848.

Itai Ashlagi and Yannai A. Gonczarowski. 2018. Stable matching mechanisms are not obviously strategy-proof. *Journal of Economic Theory* 177 (2018), 405–425.

Sophie Bade and Yannai A. Gonczarowski. 2017. Gibbard-Satterthwaite Success Stories and Obvious Strategyproofness. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC)*. 565.

Inácio Bó and Rustamdjan Hakimov. 2020. Iterative Versus Standard Deferred Acceptance: Experimental Evidence. *The Economic Journal* 130, 626 (07 2020), 356–392. https://doi.org/10.1093/ej/uez036

Inácio Bó and Rustamdjan Hakimov. 2023. Pick-an-object mechanisms. *Management Science* (2023).

Yves Breitmoser and Sebastian Schweighofer-Kodritsch. 2022. Obviousness Around the Clock. *Experimental Economics* 25 (2022), 483–513.

Daniel L. Chen, Martin Schonger, and Chris Wickens. 2016. oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9, C (2016), 88–97.

Yan Chen and Tayfun Sönmez. 2006. School choice: An experimental study. *Journal of Economic Theory* 127, 1 (2006), 202–231.

David Danz, Lise Vesterlund, and Alistair J Wilson. 2022. Belief elicitation and behavioral incentive compatibility. *American Economic Review* 112, 9 (2022), 2851–2883.

Bnaya Dreyfuss, Ofer Glicksohn, Ori Heffetz, and Assaf Romm. 2022a. Deferred Acceptance with News Utility. (2022). In preparation.

Bnaya Dreyfuss, Ori Heffetz, and Matthew Rabin. 2022b. Expectations-based loss aversion may help explain seemingly dominated choices in strategy-proof mechanisms. *Forthcoming in American Economic Journal: Microeconomics* (2022).

Ignacio Esponda and Emanuel Vespa. 2023. Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory. *The Review of Economic Studies* (10 2023).

Clayton R Featherstone and Muriel Niederle. 2016. Boston versus deferred acceptance in an interim setting: An experimental investigation. *Games and Economic Behavior* 100 (2016), 353–375.

Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic perspectives* 19, 4 (2005), 25–42.

D. Gale and Lloyd S. Shapley. 1962. College Admissions and the Stability of Marriage. *Amer. Math. Monthly* 69 (1962), 9–14.

Yannai A Gonczarowski, Ori Heffetz, and Clayton Thomas. 2023. Strategyproofness-Exposing Mechanism Descriptions. *arXiv preprint arXiv:2209.13148* (2023). Extended abstract appeared in EC 2023..

Julien Grenet, YingHua He, and Dorothea Kübler. 2022. Preference discovery in university admissions: The case for dynamic multioffer mechanisms. *Journal of Political Economy* 130, 6 (2022), 1427–1476.

Pablo Guillen and Rustamdjan Hakimov. 2018. The effectiveness of top-down advice in strategy-proof mechanisms: A field experiment. *European Economic Review* 101 (2018), 505–511.

Pablo Guillen and Alexander Hing. 2014. Lying through their teeth: Third party advice and truth telling in a strategy proof mechanism. *European Economic Review* 70 (2014), 178–185.

Pablo Guillen and Róbert F Veszteg. 2021. Strategy-proofness in experimental matching markets. *Experimental Economics* 24 (2021), 650–668.

Rustamdjan Hakimov and Dorothea Kübler. 2021. Experiments on centralized school choice and college admissions: A survey. *Experimental Economics* 24 (2021), 434–488.

Peter J Hammond. 1979. Straightforward individual incentive compatibility in large economies. *Review of Economic Studies* 46, 2 (1979), 263–282.

Avinatan Hassidim, Déborah Marciano, Assaf Romm, and Ran I Shorrer. 2017. The mechanism is truthful, why aren't you? *American Economic Review, Papers and Proceedings* 107, 5 (2017), 220–224.

Avinatan Hassidim, Assaf Romm, and Ran I Shorrer. 2021. The Limits of Incentives in Economic Matching Procedures. *Management Science* 67, 2 (2021), 951–963.

John H Kagel and Dan Levin. 1993. Independent private value auctions: Bidder behaviour in first-, second-and third-price auctions with varying numbers of bidders. *Economic Journal* 103, 419 (1993), 868–879.

Peter Katuščák and Thomas Kittsteiner. 2020. Strategy-Proofness Made Simpler. (2020). Mimeo.

Andrew Kloosterman and Peter Troyan. 2023. Rankings-Dependent Preferences: A Real Goods Matching Experiment. In *Proceedings of the 24th ACM Conference on Economics and Computation*. 956–956.

Shengwu Li. 2017. Obviously strategy-proof mechanisms. *American Economic Review* 107, 11 (2017), 3257–87.

Pinaki Mandal and Souvik Roy. 2021. Obviously Strategy-proof Implementation of Assignment Rules: A New Characterization. *International Economic Review* 63, 1 (2021), 261–290.

Takehito Masuda, Ryo Mikami, Toyotaka Sakai, Shigehiro Serizawa, and Takuma Wakayama. 2022. The net effect of advice on strategy-proof mechanisms: An experiment for the Vickrey auction. *Experimental Economics* 25 (2022), 902–951.

Vincent Meisner and Jonas von Wangenheim. 2023. Loss aversion in strategy-proof school-choice mechanisms. *Journal of Economic Theory* 207 (2023), 105588.

NMS. 2020. The Matching Algorithm - Explained. https://www.youtube.com/watch?v=kVTwXNawpbk Video produced by National Matching Services.

Joana Pais and Ágnes Pintér. 2008. School choice and information: An experimental study on matching mechanisms. *Games and Economic Behavior* 64, 1 (2008), 303–328.

Parag A Pathak. 2017. What really matters in designing school choice mechanisms. *Advances in Economics and Econometrics* 1 (2017), 176–214.

Marek Pycia and Peter Troyan. 2023. A theory of simplicity in games and mechanism design. *Econometrica* (2023). Abstract ("Obvious Dominance and Random Priority") at Proceedings of the 20th ACM Conference on Economics and Computation (EC 2019).

Alex Rees-Jones. 2018. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. *Games and Economic Behavior* 108, C (2018), 317–330.

Alex Rees-Jones and Ran Shorrer. 2023. Behavioral Economics in Education Market Design: A Forward-Looking Review. *Journal of Political Economy Microeconomics* 1, 3 (2023), 557–613.

Alex Rees-Jones and Samuel Skowronek. 2018. An experimental investigation of preference misrepresentation in the residency match. *Proceedings of the National Academy of Sciences* 115, 45 (2018), 11471–11476.

Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Article 589, 14 pages.

Shigehiro Serizawa, Natsumi Shimada, and Tiffany Tsz Kwan Tse. 2024. *Toward an Understanding of Insincere Bidding in a Vickrey Auction Experiment*. ISER Discussion Paper. Institute of Social and Economic Research, Osaka University.

Ran I Shorrer and Sándor Sóvágó. 2017. *Obvious mistakes in a strategically simple college admissions environment*. Discussion Paper 2017-107/V. Tinbergen Institute.

Clayton Thomas. 2021. Classification of Priorities Such That Deferred Acceptance is OSP Implementable. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 860.

Peter Troyan. 2019. Obviously Strategy-Proof Implementation Of Top Trading Cycles. *International Economic Review* 60, 3 (2019), 1249–1261.

A Full Experimental Materials

For Appendix A, which contains screenshots of every screen of every treatment, see the supplementary material on the authors' websites.

B Additional Analysis

In this appendix, we present some supplemental analyses of our data.

B.1 Demographic Characteristics

The last screen of the experiment elicits demographic characteristics of our participants. The set of demographic characteristics and their possible values are as follows:

- State of residence (a choice of US states)
- Age ($[18, 30)$, $[30, 40)$, $[40, 50)$, $[50, 60)$, $[60, 70)$, 70 or above; elicited as birth year and converted into age groups in the analysis).
- Number of people in household (1, 2, 3, 4, 5, 6 or above; elicited as any number and converted into one of these groups in the analysis).
- Number of people in household aged at least 18 (same categories).
- Gender (male, female, non-binary, other, prefer not to answer).
- Race (White and/or European-American, Black and/or African-American, Native American and/or First Nations, Hispanic and/or Latino, Asian and/or Pacific Islander, Middle-Eastern and/or North African, Multiracial and/or Mixed, other, prefer not to answer).
- Education (middle school or less, some high school, high school diploma, GED (HS equivalent), some college without finishing, two-years college degree / Associate degree / A.A. / A.S., four-year college degree / B.A. / B.S., some graduate school, Master's degree (MA / MS / MBA / MFA / MDiv), advanced degree (PhD / MD / JD)).
- Primary education focus (humanities, social sciences, natural sciences or math, applied science or engineering, none).
- Knowledge of the DA mechanism (never heard, heard but don't remember, have vague knowledge, know some details, familiar with it, know it well).

- Participation in real-life DA application in the past or in the future (no participation, past participation, planned future participation).
- Marital status (married, widowed, separated, divorced, single, living with a significant other).
- Employment status (working (besides Prolific, if applicable), unemployed, retired, stay-at-home parent, student, other).
- Social views (very liberal, liberal, slightly liberal, moderate, slightly conservative, conservative, very conservative, other).
- Economic views (same categories).
- Identification with a political party (Republican, Democrat, Independent, other, none of the above).
- Vote in the 2020 presidential elections (Joe Biden, Donald Trump, other, did not vote).
- combined household income ([0, \$20k), [\$20k, \$40k), [\$40k, \$60k), [\$60k, \$80k), [\$80k, \$100k), [\$100k, \$150k), [\$150k, \$200k), \$200k or above).

Figure B.1, Figure B.2, Figure B.3 and Figure B.4 show the distribution of these characteristics in our sample across treatments.

B.2 Experiment Duration

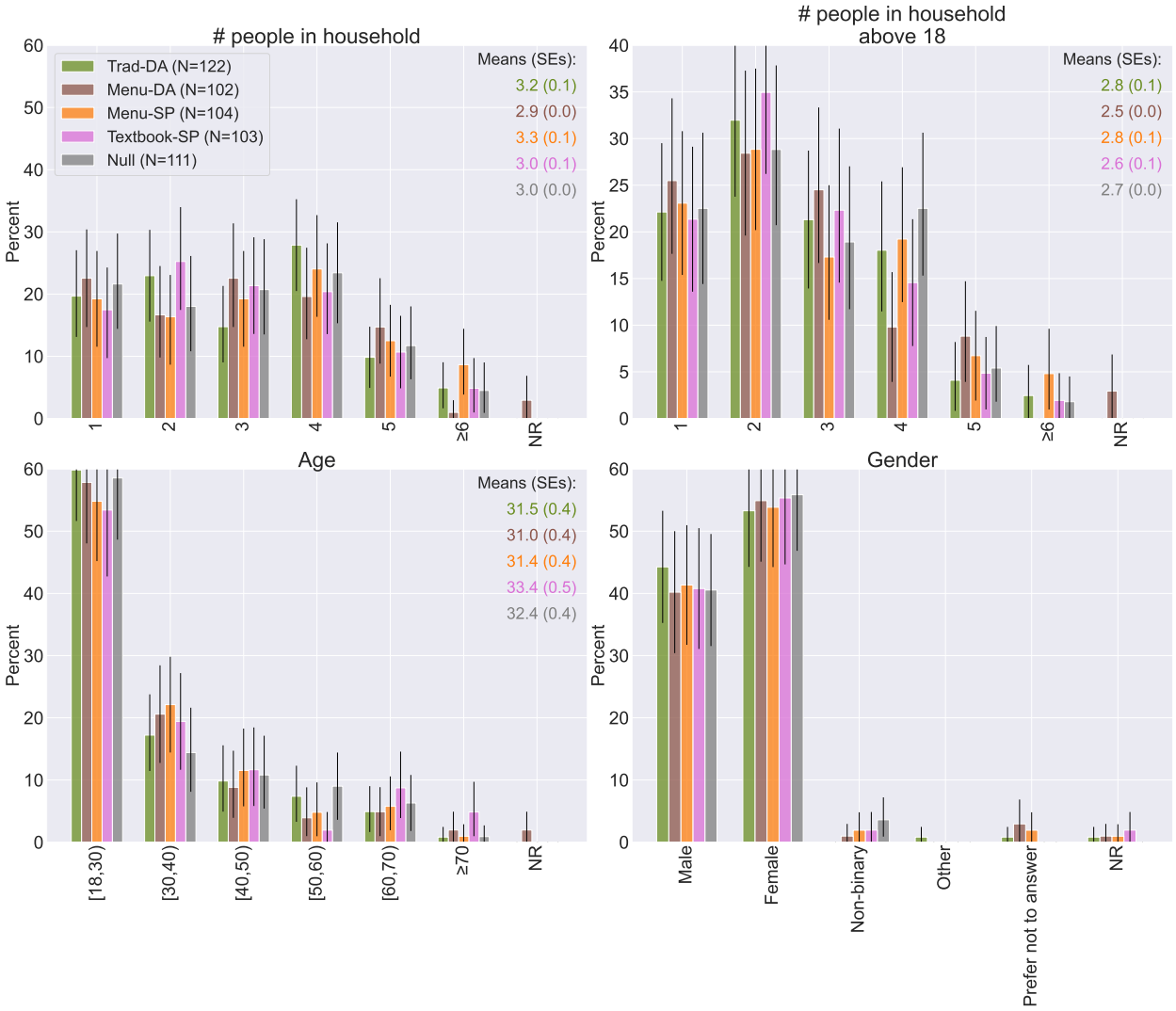
Figure B.5 shows the distribution of experiment duration across treatments and the different main parts of the experiment.

B.3 Prolific vs. Cornell Sub-Samples, Cognitive Score, Attention Score, and their Mediating Effects on Main Results

Among possible variables mediating our treatment effects, focus and attention, as well as cognitive resources and human capital seem highly important to effectively learning from any description, since they are related to the ability to learn, process, and comprehend the complicated information conveyed in the experiment.[33]

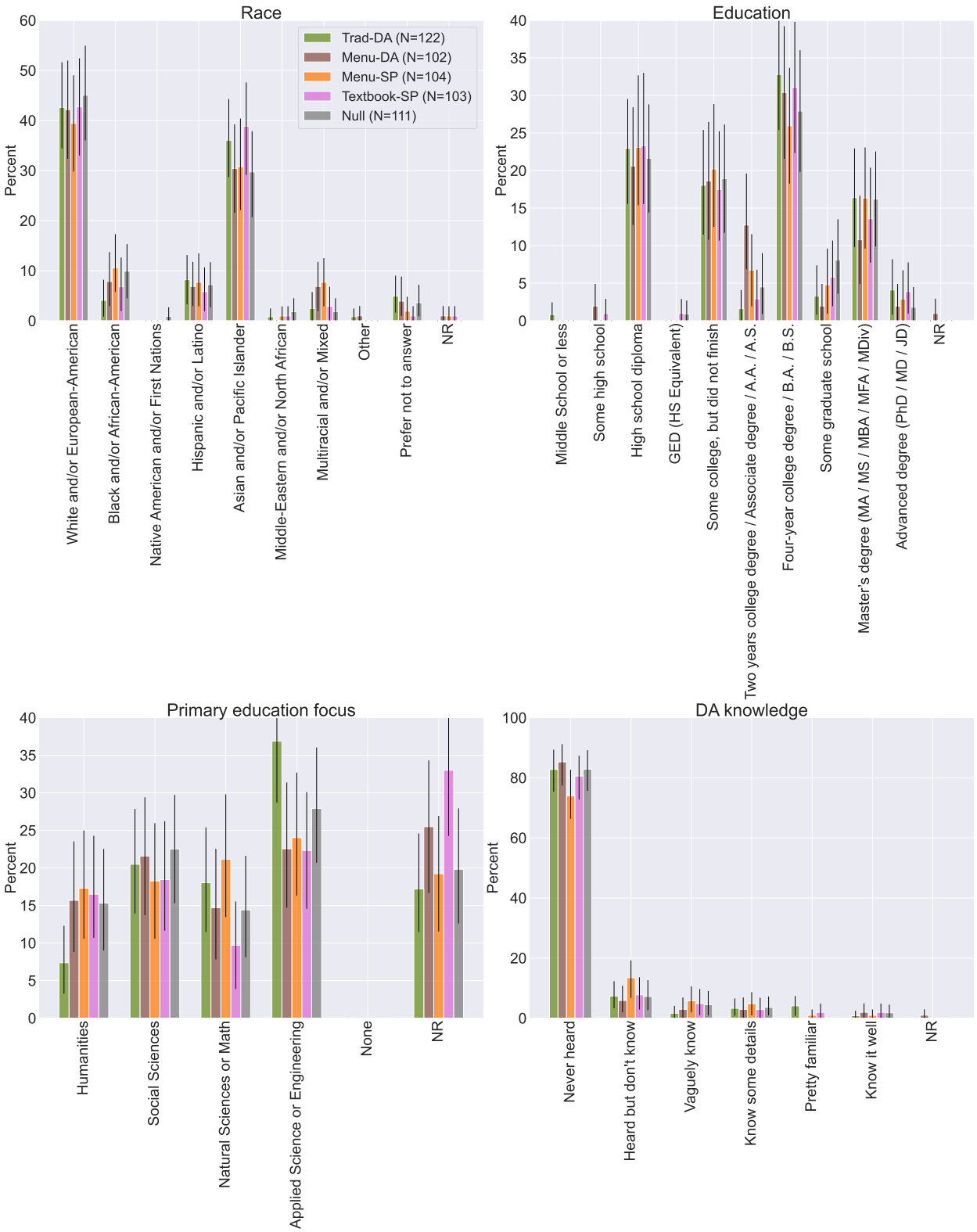
[33]Moreover, we would expect the difference between our descriptions to matter the most for participants whose human capital, cognitive level and attention are neither too low—which would render explanation attempts useless—nor too high—which would render all descriptions equally effective at promoting understanding. Exploring multiple levels of these mediators is thus useful to investigate the full potential of conveying one description rather than another one.

Figure B.1: Distribution of demographic characteristics in our sample (1/3).



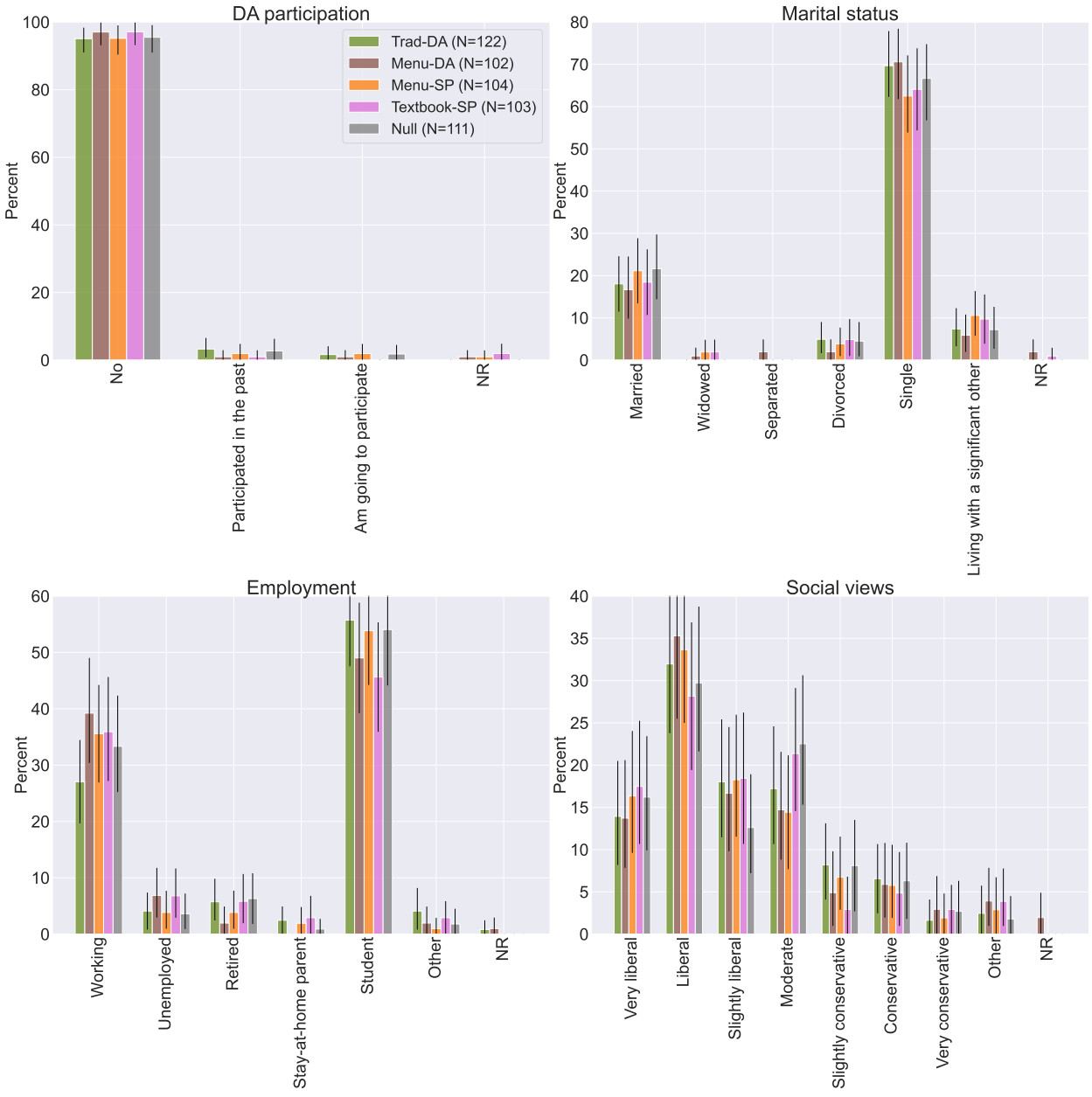
Notes: The panels describe the distribution of responses to all demographic questions except for state of residence.

Figure B.2: Distribution of demographic characteristics in our sample (2/3).



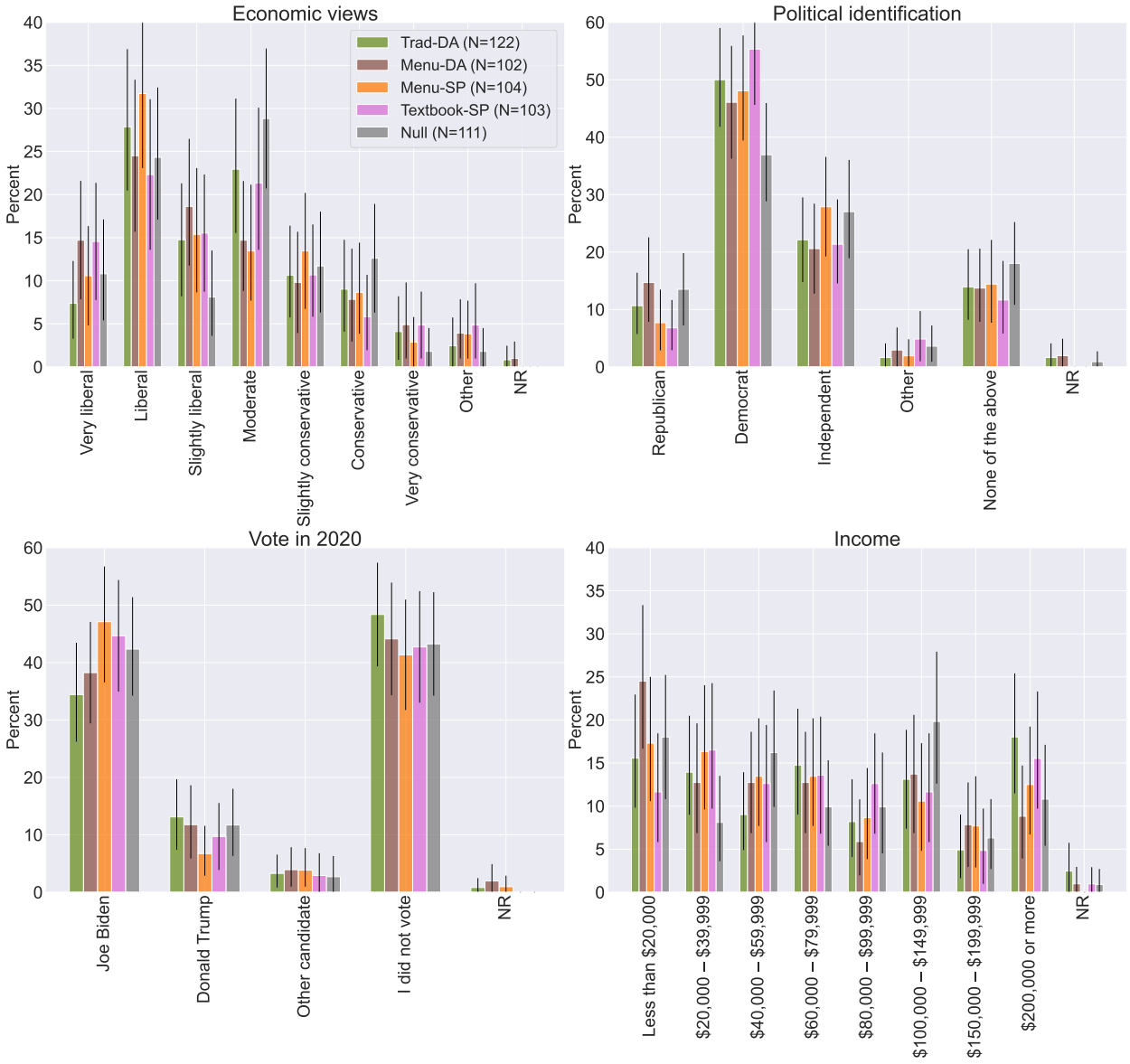
Notes: See Figure B.1.

Figure B.3: Distribution of demographic characteristics in our sample (3/3).



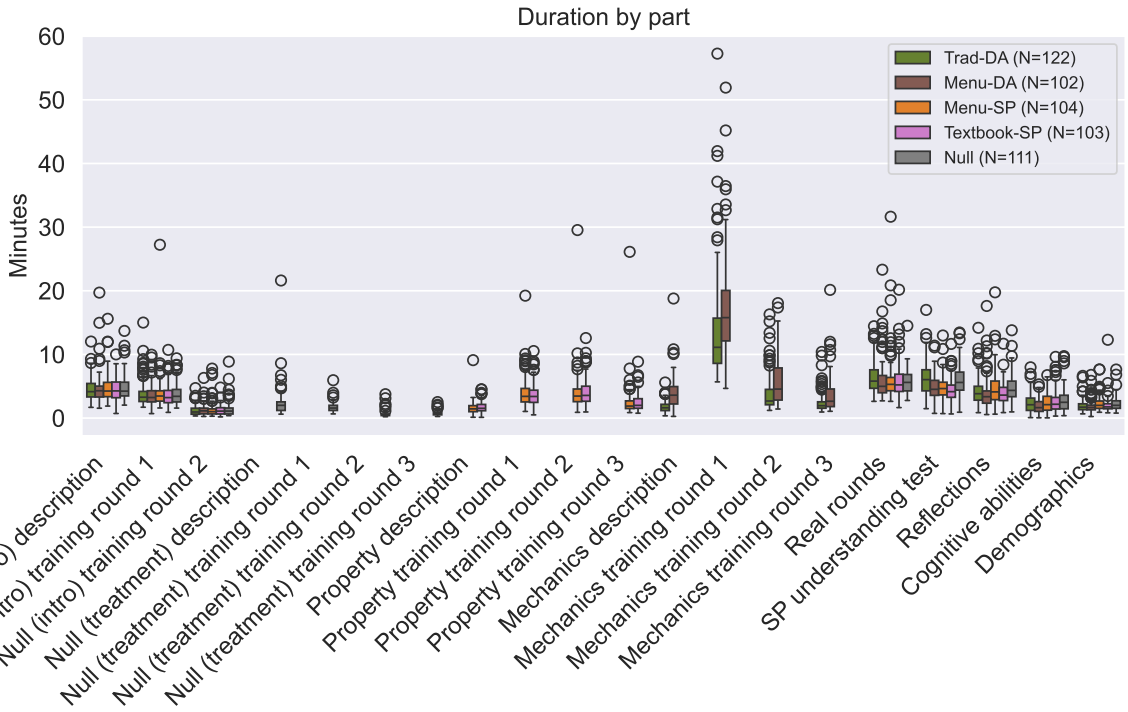
Notes: See Figure B.1.

Figure B.4: Distribution of demographic characteristics in our sample (3/3).



Notes: See Figure B.1.

Figure B.5: Experiment duration by component and by treatment.

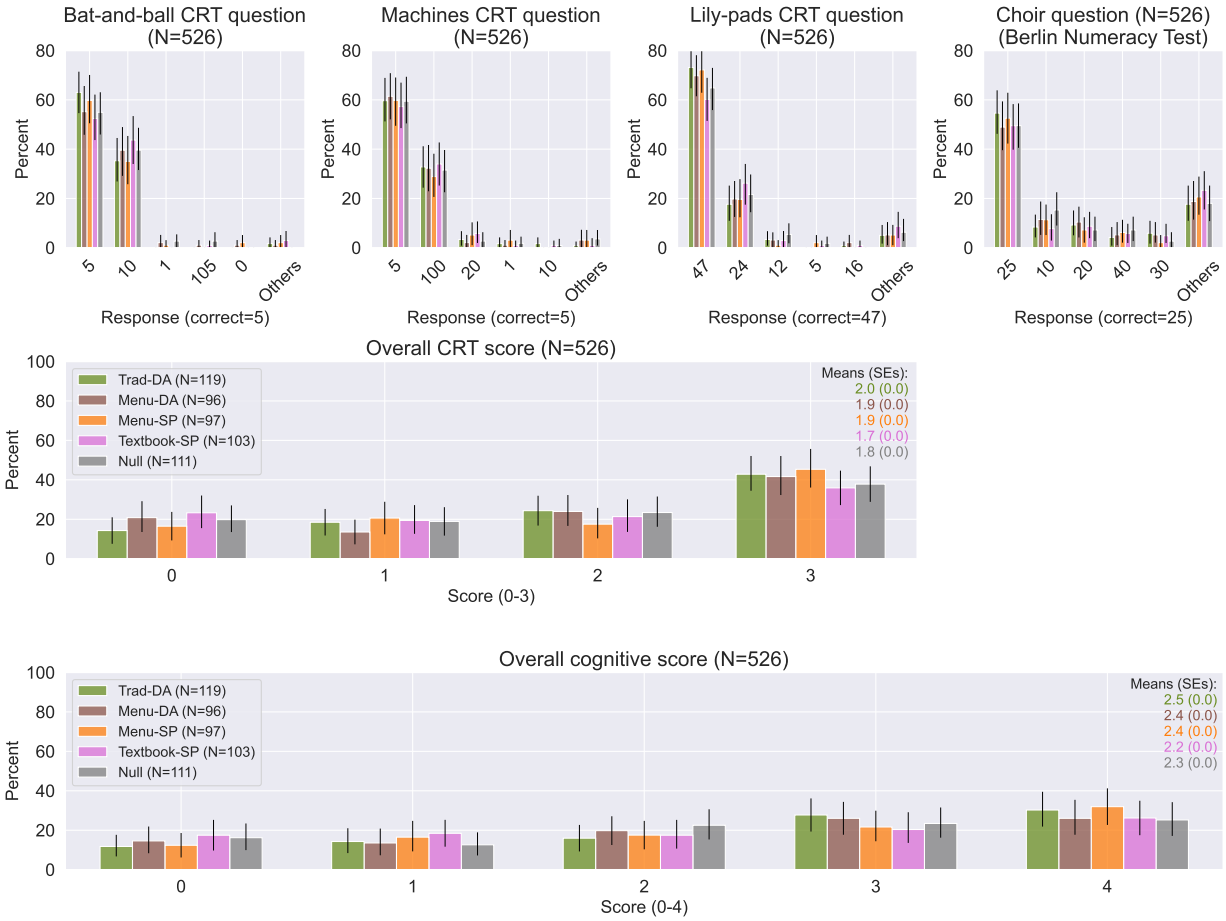


We use three key measures to study these effects: an indicator for whether a participant belongs to the Prolific or Cornell sub-sample, a cognitive score measure elicited at the end of the experiment and an attention score elicited in two attention checks planted within the experiment. As discussed in Section 2.6, the elicited cognitive score varies from 0 to 4. Figure B.6 shows its distribution, as well the distribution of responses to the individual underlying questions, across treatments.[34] The two attention checks were planted in the training questions following the Null description at the beginning of the experiment and in the middle of the SP understanding test, towards the end of the experiment. They include questions that explicitly instruct what response to submit, but are designed to look similarly to adjacent content, such that participants who skim quickly through the text are more likely to miss the explicit instructions and to answer incorrectly. Figure B.7 shows the performance in the attention checks and the total attention scores.

To test the effects of these measures on our main findings, we examine three sample splits, into (1) Prolific vs. Cornell (Figure B.8), (2) the bottom (roughly) half of cognitive scores (0–2) vs. the top half (3–4) (Figure B.9), (3) bottom attention scores (0–1) vs. a top (perfect) attention score (2) (Figure B.10). Each of these figures replicates the mean plots

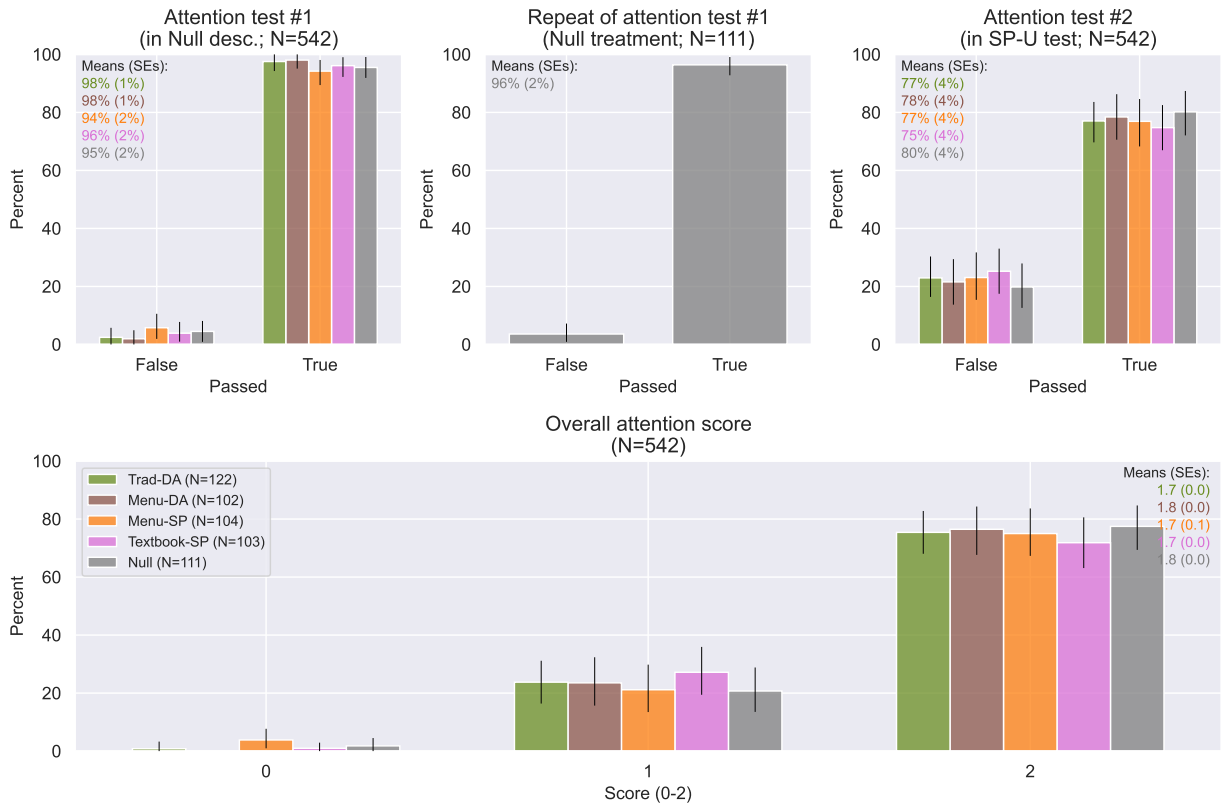
[34]Due to a bug in the cognitive score elicitation interface in the first runs of the experiment, 16 observations lack this data and are omitted from this analysis.

Figure B.6: Distribution of cognitive score and its sub-measures across treatments.



Notes: The panels describe the distribution of responses to the four questions included in the total cognitive score, in order of response frequency, and the distribution of the overall scores. For screenshots of the questions, see Appendix B.

Figure B.7: Distribution of attention score and its sub-measures across treatments.



Notes: The panels describe the distribution of performance in the two attention checks which appeared during the experiment, and the success rate in an additional attention test given to Null treatment participants only.

of Figure 7, Figure 8, and Figure 9 on the left, and in addition shows the conditional mean values when splitting the sample according to the criteria above. On the right, each figure replicates the % SP-U vs. % SF relationship from Figure 11 conditional on the splits.

First, we find a noticeable increase in all means for Cornell participants, top cognitive scorers and top attention scorers relative to the pooled averages, and a similar decrease for Prolific participants and bottom cognitive or attention scorers. These gradients provide additional support to our basic interpretation that % TR and % SP-U reflect understanding of relevant content and are thus positively affected by these mediators (in contrast, e.g., to reflecting other, less relevant features of the descriptions, such as length, general appearance, or language style).

Second, directional differences between % TR and % SP-U rates across treatments remain similar across the sample splits. This finding suggests that even among participants with high abilities, our SP Property descriptions are not easily understood, and Menu-SP is more easily understood than Textbook-SP (i.e., requires lower ability/attention levels to get to the same % TR or % SP-U score).

Third, the particular step-function relation between % SP-U and % SF from Figure 11 seems consistent across the sample splits. However, since % SP-U is lower for the bottom categories of these splits, they include less participants with high levels of % SP-U, above 75%. Therefore the effect of Menu-SP in pushing participants to the high mode of both SP-U and SF play is less pronounced compared to the other treatments.

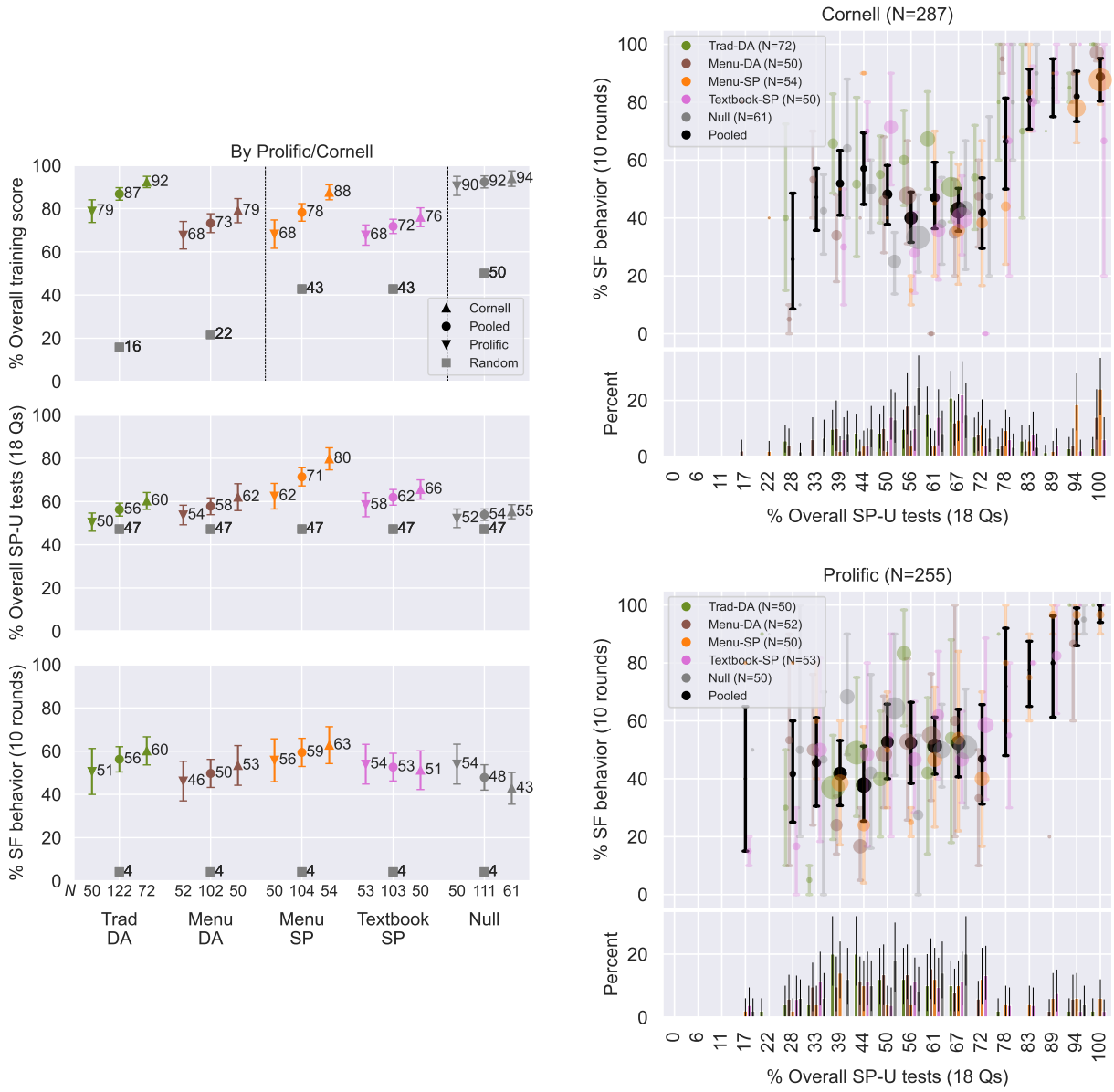
B.4 Robustness of Main Findings to Adding Controls

In this appendix we test whether our four main findings are robust to adding a rich set of controls as explanatory variables.

First, Table B.1 shows the treatment effects on the three outcome variables % TR, % SP-U and % SF using OLS regressions on treatment indicators. The columns without controls repeat the mean values shown in Figure 7, Figure 8 and Figure 9, in terms of difference from the Null treatment mean. The control variables include (1) a set of demographic indicators (described in Section B.1), (2) indicators for cognitive and attention scores, (3) date indicators for all days in which session of the experiment took place. (1) and (2) include a category for missing values in case no response was recorded for a question. Overall, adding controls does not change the treatment-means of our three main outcome variables by much.

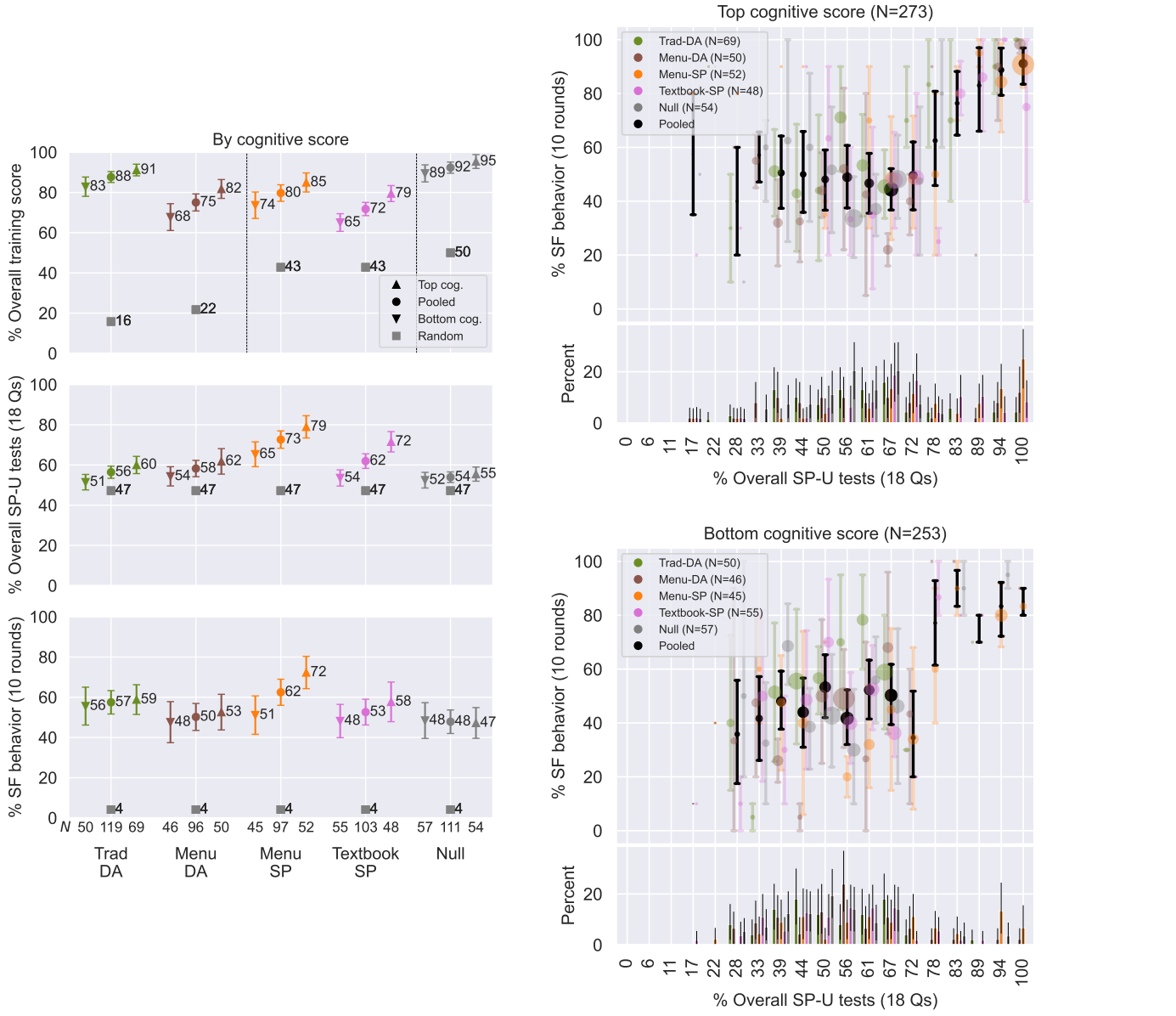
Next, Table B.2 shows mean % SF when conditioning on some % SP-U level (no participants got % SP-U below 17%). In the leftmost column these numbers simply replicate the data from Figure 11. The table shows that adding controls and treatment indicators does

Figure B.8: Main results by Prolific vs. Cornell sub-sample.



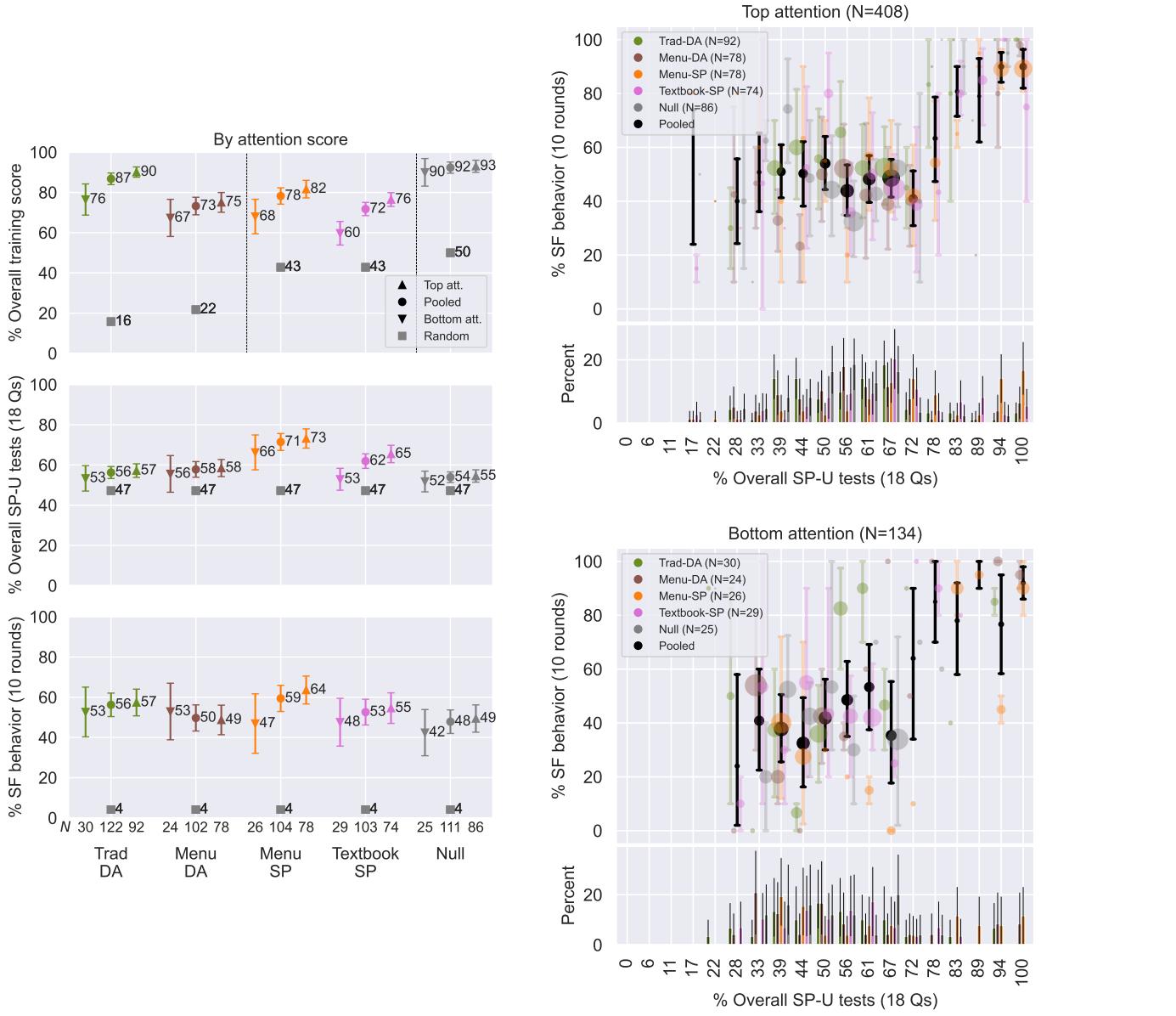
Notes: Left column: Each panel displays mean rates of an outcome variable by treatment. Each panel includes the overall mean, the mean among the sample collected at Cornell ($N = 287$) and among the sample collected at Prolific ($N = 255$; see Section 3.1). "Random": the expectation of each measure from uniformly random answers.
Right column: Replications of Figure 11 for the Cornell and Prolific sub-samples.

Figure B.9: Main results by top vs. bottom cognitive scores.



Notes: See Figure B.8; the sample split is according to top (3–4) vs. bottom (0–2) scores in the cognitive-ability test conducted at the end of the experiment.

Figure B.10: Main results by top vs. bottom attention scores.



Notes: See Figure B.8; the sample split is according to top (2) vs. bottom (0–1) scores in the attention checks planted within the experiment.

Table B.1: Treatment effects on outcome variables without controls and with controls.

	% TR	% TR	% SP-U	% SP-U	% SF	% SF
Trad-DA	−0.06	−0.09	0.02	0.00	0.08	0.08
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.05)
Menu-DA	−0.19	−0.18	0.04	0.05	0.02	0.03
	(0.03)	(0.02)	(0.02)	(0.03)	(0.04)	(0.05)
Menu-SP	−0.14	−0.13	0.18	0.18	0.12	0.14
	(0.03)	(0.02)	(0.03)	(0.02)	(0.04)	(0.05)
Textbook-SP	−0.21	−0.20	0.08	0.11	0.05	0.04
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.05)
Constant	0.92	0.64	0.54	0.73	0.48	0.28
	(0.01)	(0.12)	(0.01)	(0.11)	(0.03)	(0.23)
Controls		X		X		X
R^2	0.16	0.60	0.10	0.50	0.02	0.31
N	542	542	542	542	542	542

Note: Estimated coefficients using OLS regressions of the 3 outcome variables on treatment indicators (with Null treatment the omitted category), without and with including a full set of controls in the regression.

not change this relation by much, both at each individual % SP-U level above 75% and when aggregating all the levels above 75%.

B.5 Detailed Performance in Training Questions

In this appendix we show the details of participants' performance in training questions.

B.5.1 Null Training Common to All Participants

First, Figure B.11 shows the distribution of participants' performance in the training round which follow the Null description, which is common to all participants. These are the only training question which are *not* a part of % TR. There are four true/false questions about statements—shortly summarized above each top-row panel. In this figure and the next, instead of only showing the binary score of 1 or 0 for each training question, we show more details on the number of attempts it took to answer the question correctly. Recall that participants cannot advance beyond a question until getting it correctly and receiving a feedback of why it is correct. In all questions except for a few DA Mechanics questions, participants get a score of 1 (and a resultant monetary understanding bonus) only conditional on answering correctly at first attempt. The overall score in the round is displayed in the bottom row of the figure.

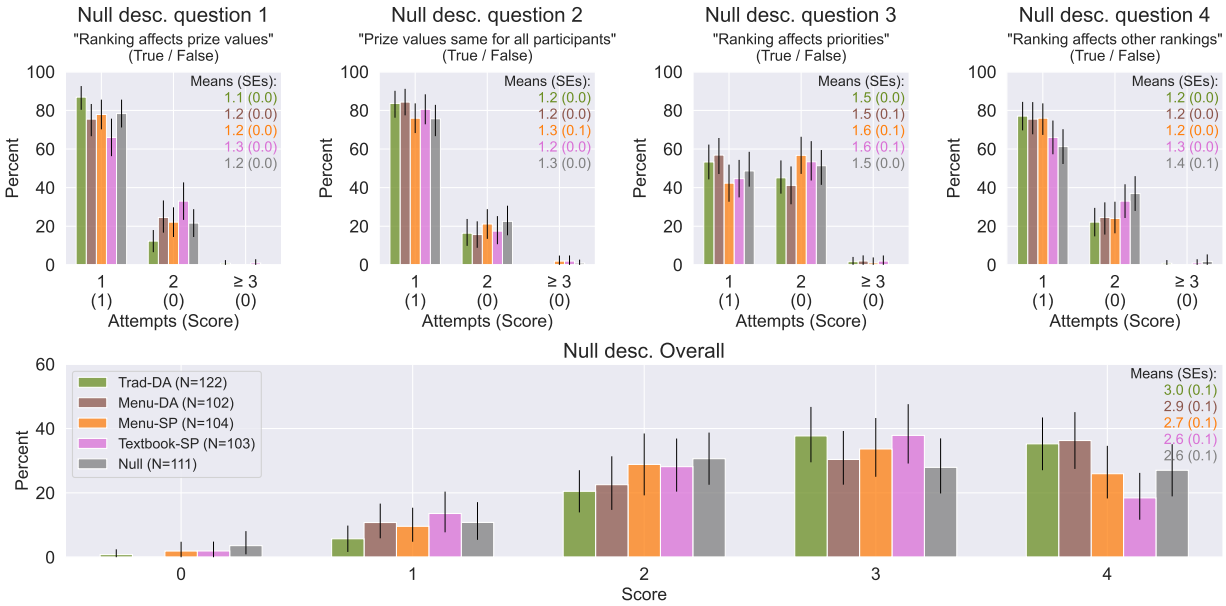
Table B.2: Relation between % SF and % SP-U without controls and with controls.

Dependent variable: % SF

	(1)	(2)	(3)	(4)
% SP-U = 17%	0.48	0.67		
	(0.13)	(0.20)		
% SP-U = 22%	0.65	0.63		
	(0.18)	(0.15)		
% SP-U = 28%	0.36	0.36		
	(0.07)	(0.08)		
% SP-U = 33%	0.46	0.52		
	(0.06)	(0.07)		
% SP-U = 39%	0.46	0.45		
	(0.04)	(0.06)		
% SP-U = 44%	0.44	0.37		
	(0.05)	(0.05)		
% SP-U = 50%	0.50	0.54		
	(0.04)	(0.05)		
% SP-U = 56%	0.45	0.42		
	(0.04)	(0.05)		
% SP-U = 61%	0.49	0.46		
	(0.04)	(0.05)		
% SP-U = 67%	0.46	0.47		
	(0.03)	(0.04)		
% SP-U = 72%	0.44	0.47		
	(0.05)	(0.05)		
% SP-U = 78%	0.68	0.74		
	(0.07)	(0.09)		
% SP-U = 83%	0.80	0.85		
	(0.05)	(0.07)		
% SP-U = 89%	0.82	0.81		
	(0.07)	(0.10)		
% SP-U = 94%	0.87	0.84		
	(0.04)	(0.06)		
% SP-U = 100%	0.90	0.92		
	(0.03)	(0.05)		
% SP-U $<$ 75%			0.46	0.46
			(0.01)	(0.02)
% SP-U $\geq$ 75%			0.83	0.83
			(0.02)	(0.03)
Controls		X		X
Treatment		X		X
R^2	0.21	0.44	0.19	0.42
N	542	542	542	542

Note: Estimated coefficients using OLS regressions % SF on indicators for all possible % SP-U levels, without and with including a full set of controls and treatment indicators in the regression.

Figure B.11: Null training questions common to all participants.



Note: *Top row*: Performance at each of the four Null training round questions, by the number of attempts it took to answer the question correctly. Parentheses below amount of attempts show the score conditional on this amount. *Bottom row*: Overall score. All panels include the mean values (SEs) in the upper right corner.

B.5.2 DA Mechanics Training

Figure B.12, Figure B.13, Figure B.14 and Figure B.15 show participants' performance in the DA Mechanics training questions, including the allocation problems they needed to solve using the GUI, as well as their usage of an additional walkthrough video available in training round 2.

B.5.3 SP Property Training

Figure B.16 and Figure B.17 show participants' performance in the the SP Property training questions.

B.5.4 Null Treatment Training

Figure B.18 show participants' performance in the Null treatment's training questions, which exactly repeated the training questions from the Null training round common to all participants.

Figure B.12: DA Mechanics training questions: round 1.

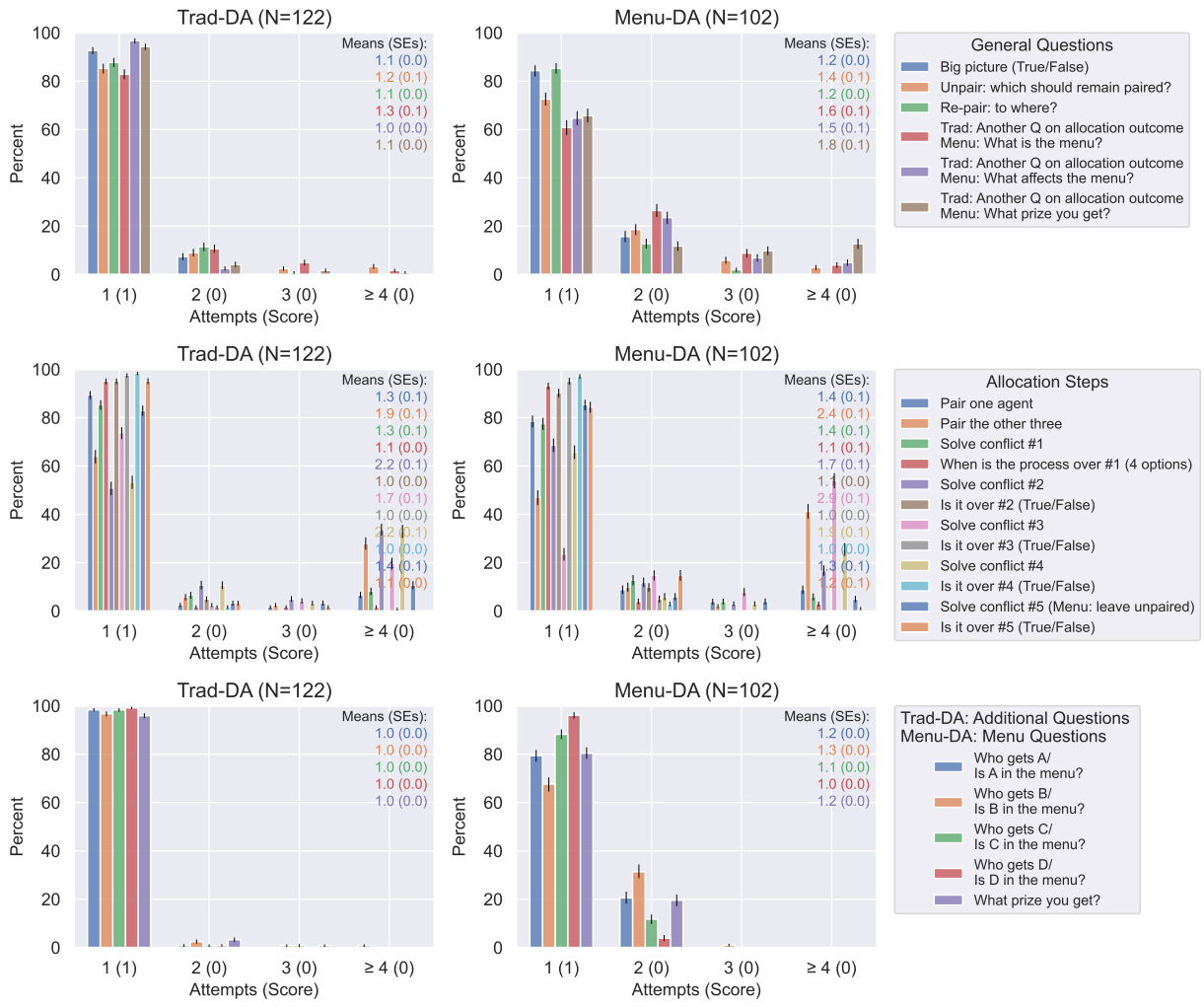


Figure B.13: DA Mechanics training questions: rounds 2–3.

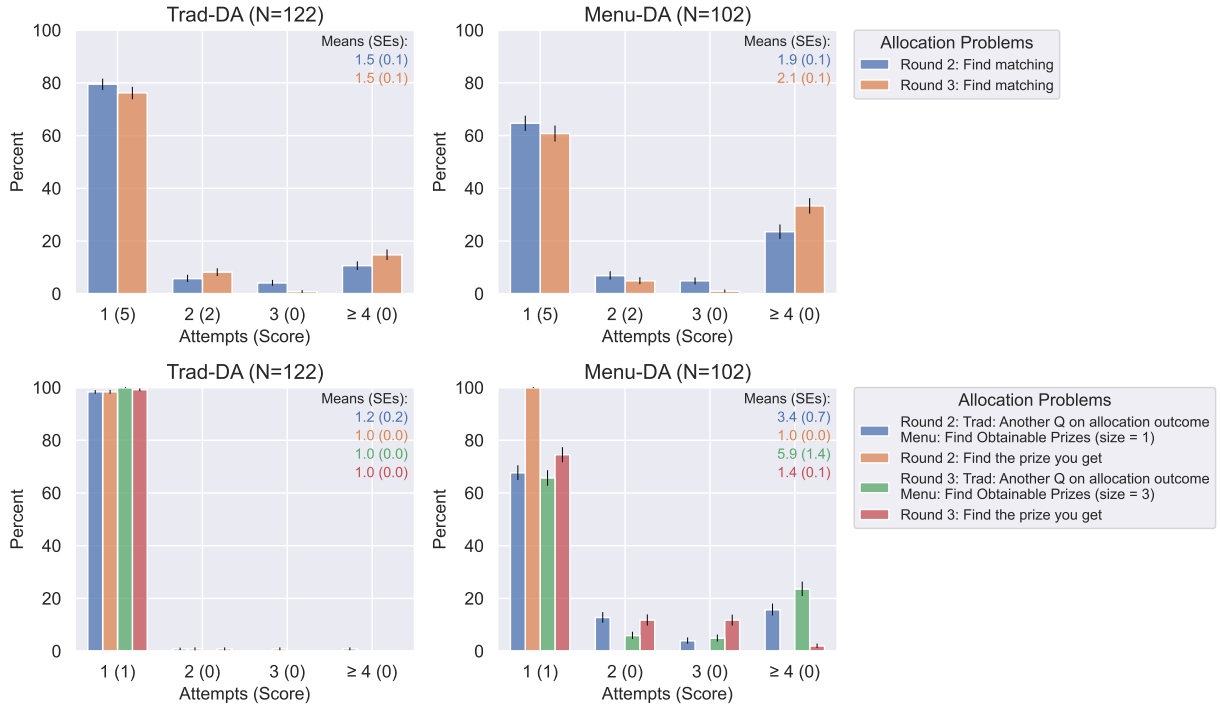


Figure B.14: DA Mechanics walkthrough-video usage (round 2).

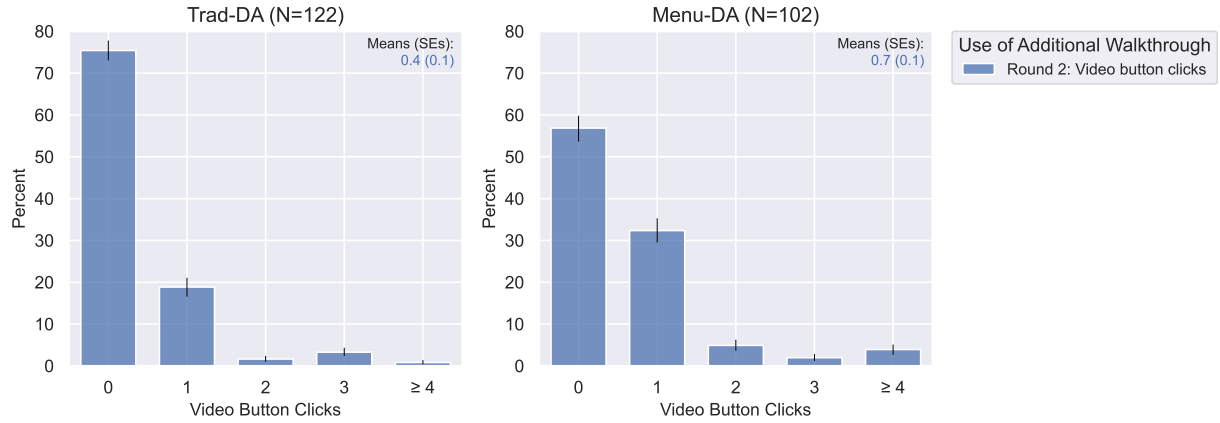


Figure B.15: DA Mechanics training scores.

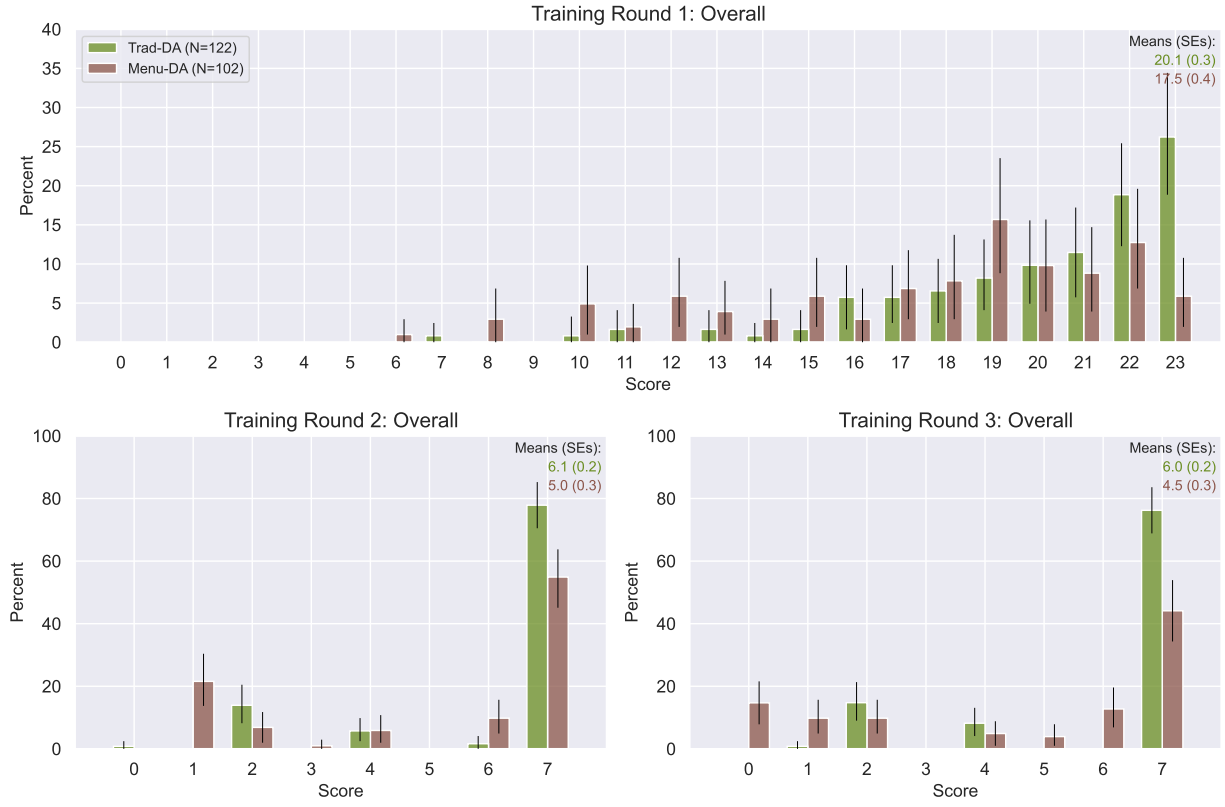


Figure B.16: SP Property training questions.

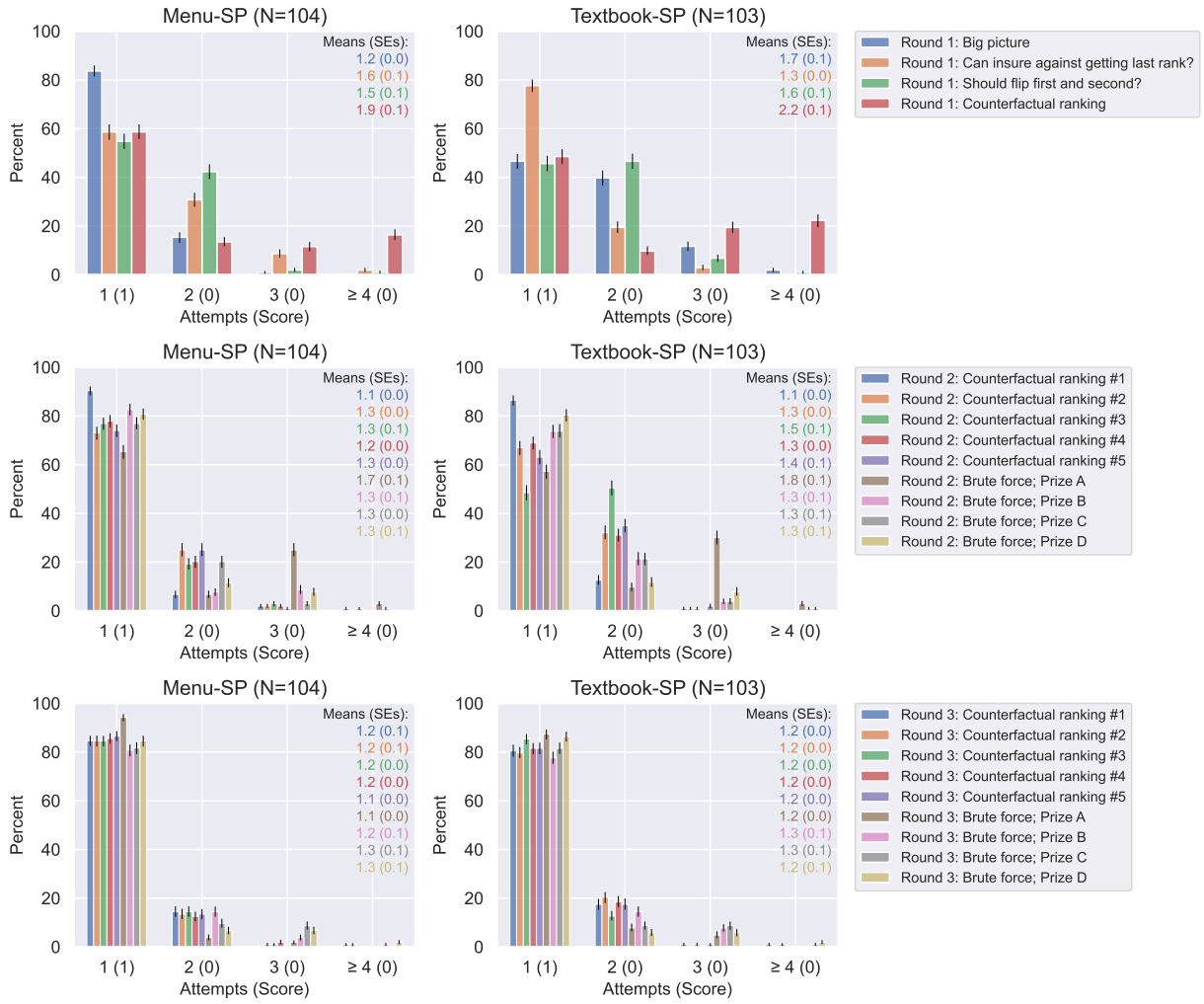


Figure B.17: SP Property training scores.

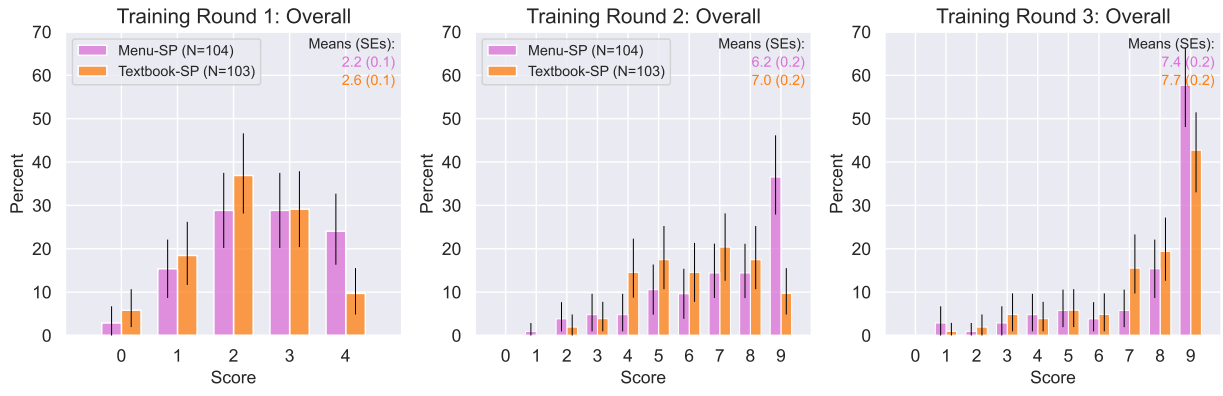


Figure B.18: Null treatment training questions.

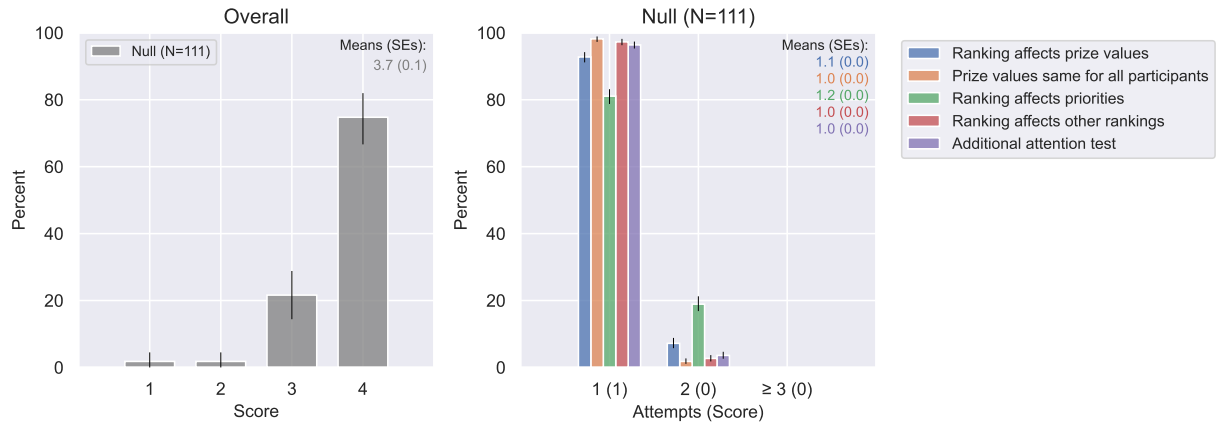
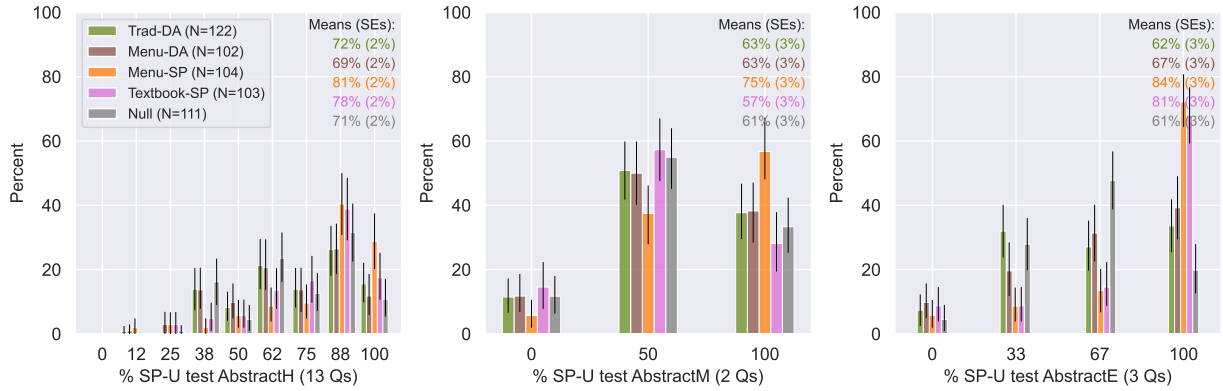


Figure B.19: The three tests composing the % Abstract SP-U sub-measure separately.



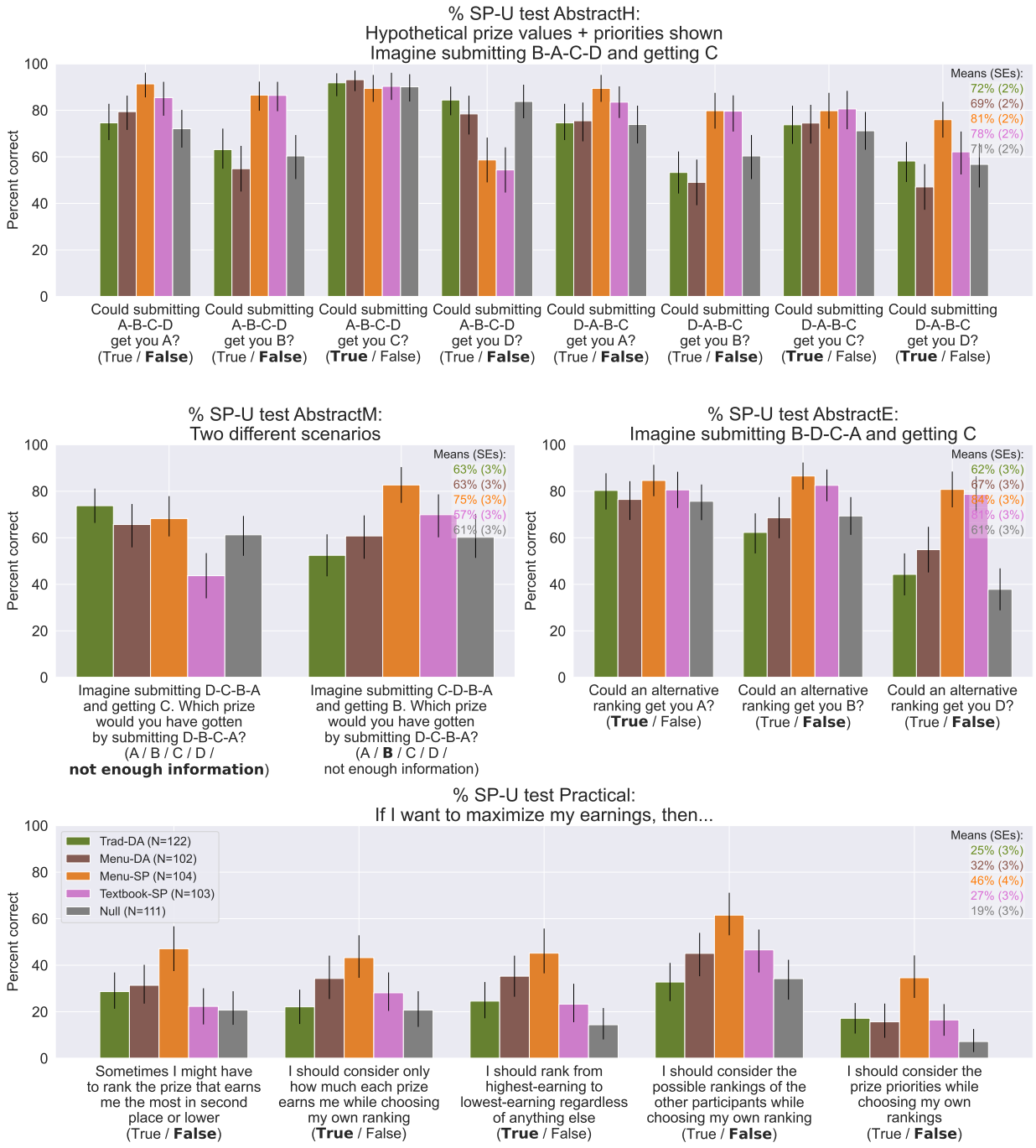
Note: Each panel shows the distribution of score in one of the three Abstract tests across treatments.

B.6 Joint Distribution of % SP-U and its Sub-Measures

First, Figure B.19 separately shows the distribution of the three test scores composing the Abstract sub-measure of % SP-U, termed AbstractHard, AbstractMedium and AbstractEasy, and Figure B.20 shows the success rates at each individual SP-U question of all four tests.

Next, Figure B.21 shows the joint distributions of % SP-U and its sub-measures, Abstract and Practical. The relation is fairly similar across treatments, hence the following discussion focuses on the pooled distributions in each row's 1 – 3 leftmost panel. The top row of the figure suggests strong dependence between the two modes of the bimodal Practical score, and the Abstract score: The high Practical mode is highly predictive of a high Abstract score—among participants with Practical $\geq 75\%$, the average Abstract score is 83%—while in the other direction, the Abstract score is not as predictive regarding Practical—even

Figure B.20: Success rates in all SP-U questions.



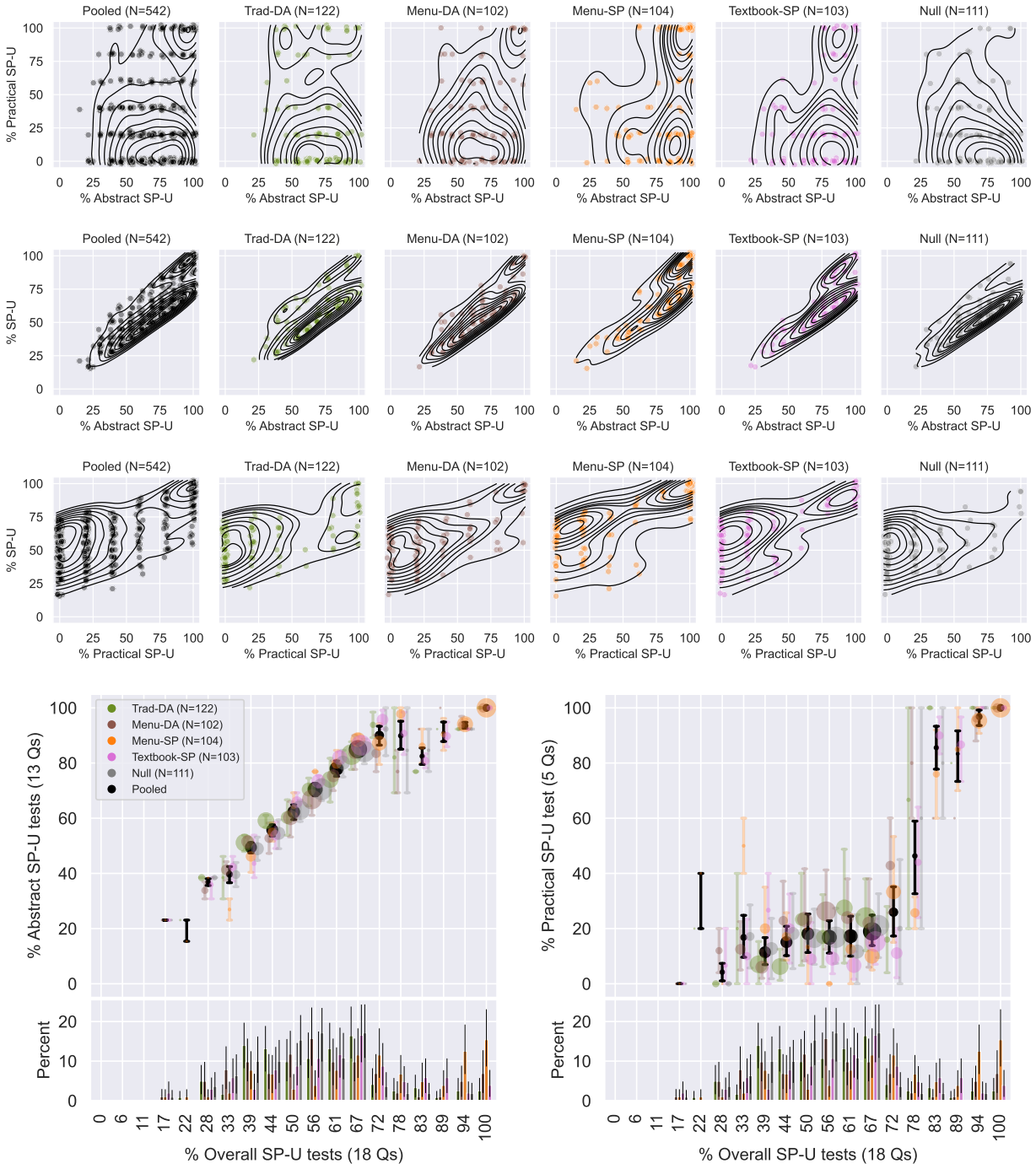
Note: Each panel shows the qusetion-level success rates in one of the four SP-U tests across treatments.

among participants with Abstract $\geq 90\%$, the average Practical score is 47%.

The above is also reflected by an approximately one-to-one relationship between SP-U and Practical shown in the third row, where despite the SP-U measure being dominated by Abstract questions (13 question relative to just 5 Practical questions), high Practical scores only correspond to the high end of SP-U. Among participants with SP-U $\geq 75\%$, the average Practical score is 85% (and unsurprisingly, the average Abstract score among these participants is also high, 93%).

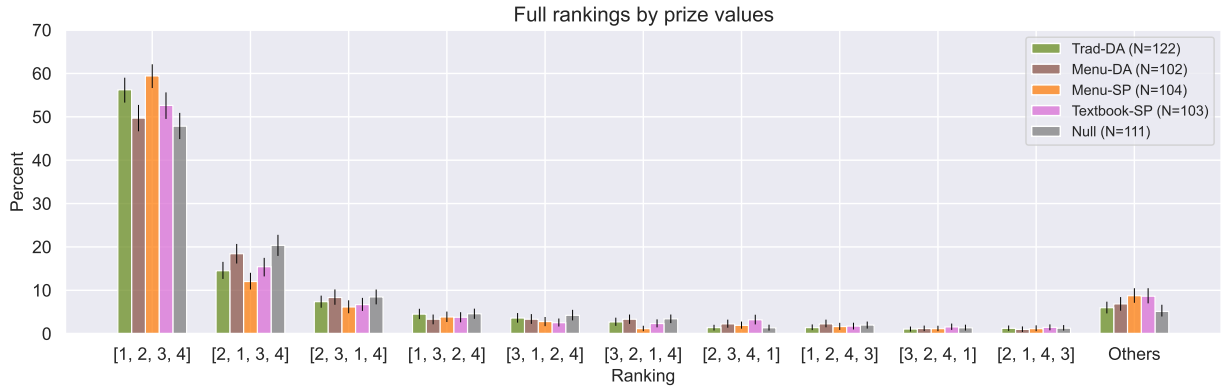
Overall, the % SP-U histogram in Figure 8 can be understood as having two main ranges. The range of SP-U $< 75\%$ is governed mostly by the Abstract score, where Practical is approximately fixed in its low mode. The range of SP-U $\geq 75\%$ is governed mostly by the Practical score, where the bimodality observed in this range reflects Practical's bimodality. This is empasized by the bottom rows of the figure, showing the means of Abstract and Practical conditional on different total % SP-U scores.

Figure B.21: Joint distributions of % SP-U and its Abstract and Practical sub-measures.



Note: Top row: joint distribution of the Abstract and Practical % SP-U sub-measures. Second row: joint distribution of the Abstract % SP-U sub-measure, concerning the abstract logical properties of strategyproofness, and the overall % SP-U measure. Third row row: joint distribution of the Practical % SP-U sub-measure, concerning how participants could maximize earnings, and the overall % SP-U measure. All rows 1 – 3 include a panel with the overall distribution in the full sample, and five additional panels focusing on each treatment's sub-sample. Each panel contains a jittered scatter plot of the two measures and estimated contours smoothing the two-dimensional distribution. Mid-bottom left: histogram of % SP-U at the bottom and the conditional means of the Abstract sub-measure given all possible % SP-U scores at the top (see Figure 11, which is similar in structure, for more details). Mid-bottom right: a similar structure with the Practical sub-measure instead of Abstract.

Figure B.22: Distribution of most frequently submitted rankings.

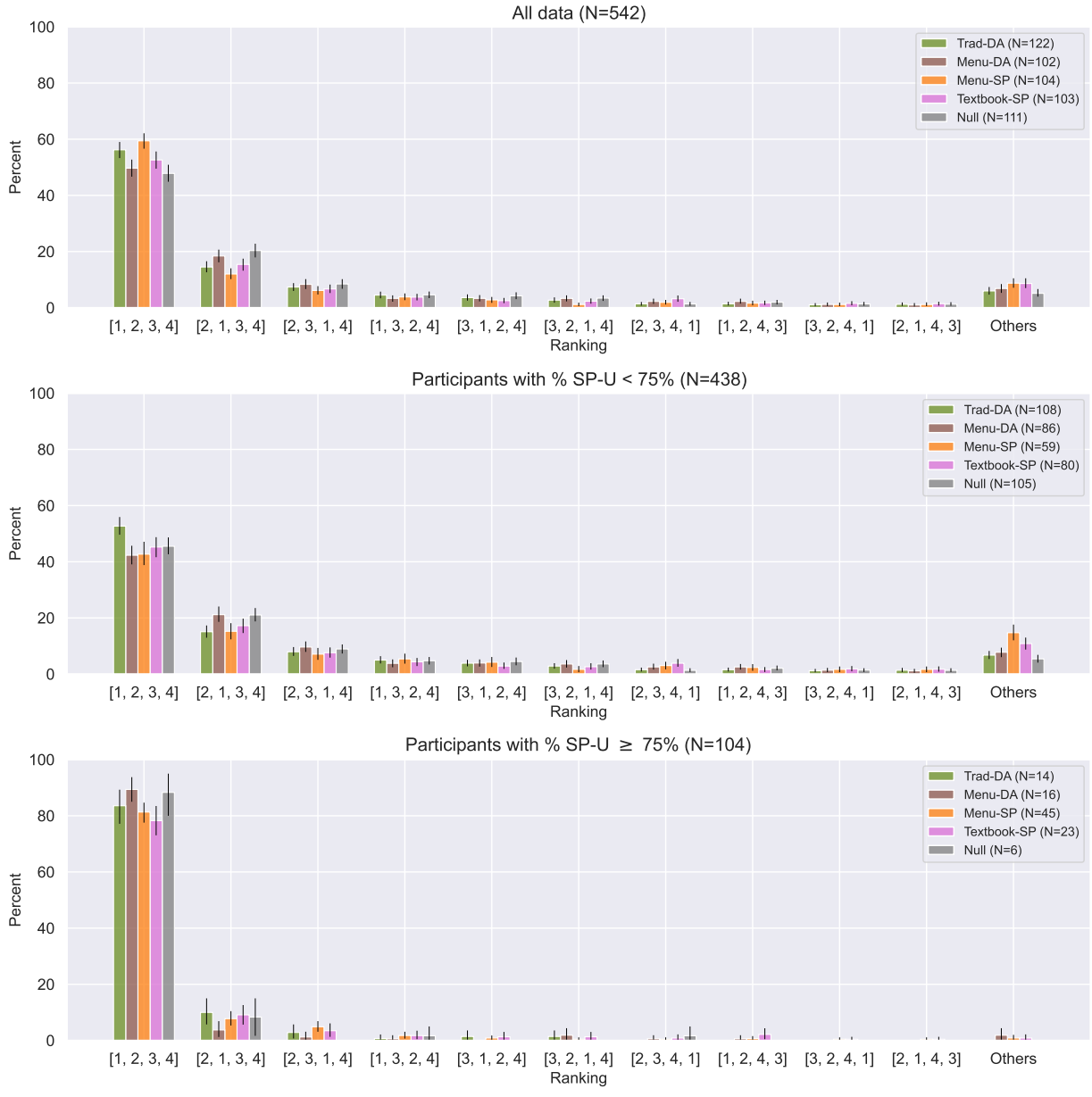


Note: Each $k \in \{1, 2, 3, 4\}$ denotes the subject's kth highest-earning prize, and rankings are written from highest to lowest. For example, ranking $[1, 3, 2, 4]$ flips the second and third highest-earning prize relative to the SF ranking ($[1, 2, 3, 4]$).

B.7 Ranking Patterns Beyond % SF

Figure B.22 shows the distribution of the most common ranking patterns in the real rounds of playing DA. The SF behavior—ranking $[1, 2, 3, 4]$ where numbers reflect ordinal rank of prize value from highest- to lowest-earning—is the most common strategy across all treatments. Additionally, the most common NSF ranking is $[2, 1, 3, 4]$, i.e., a ranking that flips the highest- and second-highest-earning prizes. We see little variation in types of NSF behavior across treatments. Figure B.23 shows that participants with low SP understanding (below 75%), who play NSF in higher rates, still mostly use the $[2, 1, 3, 4]$ strategy among NSF ones.

Figure B.23: Distribution of submitted rankings by SP understanding score.



Note: The categorization of rankings is the same as in Figure B.22.

Next, Figure B.24 explores whether beyond treatment, ranking behavior depends on key round parameters. The figure shows the frequency of four different strategies or groups of strategies as a function of round parameters depending on prize priorities and prize monetary values. The strategies we test are (1) any NSF, (2) any NSF where the highest-earning prize is not ranked first, and (3) any NSF strategy consistent with ranking according to expectations-based reference-dependent (EBRD) preferences under some beliefs (according to a characterization given by Meisner and von Wangenheim (2023)).[35] Since the most common NSF behavior relative to SF ranking is flipping highest- and second-highest-earning prizes, the round parameters we explore include the participant's priority of getting the highest-earning prize, the difference between the participant's priority of getting that prize vs. their priority of getting other prizes, and on the difference between the monetary values of the highest- and second-highest-earning prizes. Overall, we do not find any strong pattern conditional on these parameters in our data. Figure B.25 and Figure B.26 show that the weak patterns we find are somewhat stronger for participants with a low % SP-U score, below 75%, and are non-distinguishable from zero for participants with % SP-U above 75%.

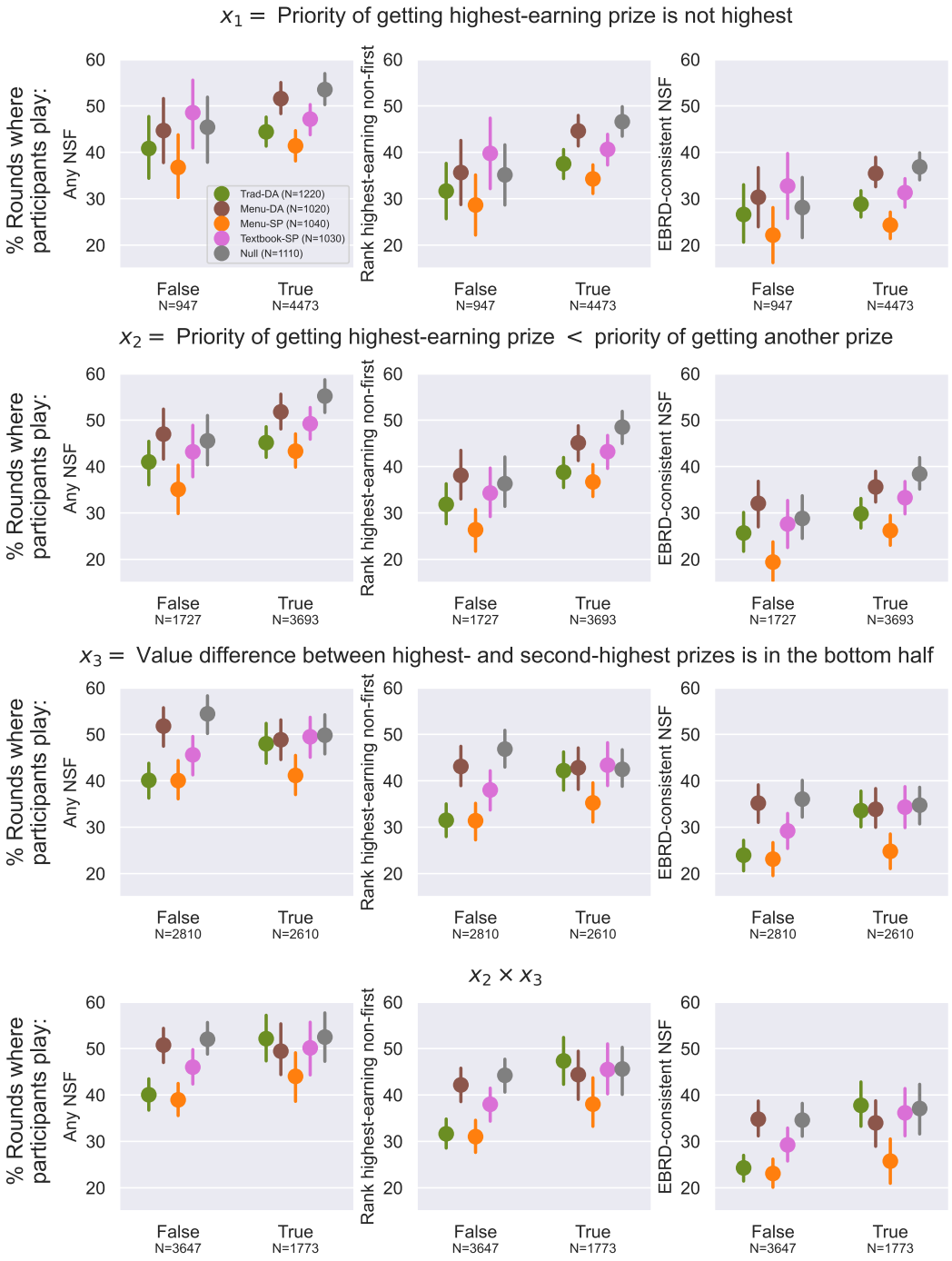
Next, we investigate whether ranking behavior changed over the 10 rounds that each participant played. Table B.3 shows the difference in participants ranking behavior in rounds 1-5 vs. rounds 6-10, and also shows the regression coefficient between round number and whether or not the participant played SF that round. Rates of SF play also seems stable across the 10 rounds, perhaps except in Menu-SP, where there is a slight trend towards more SF play in later rounds, with an average increase of a 1.6% (SE = 0.5%) per round in the fraction of participants playing SF in that round.

B.8 Relationship Between % TR and % SP-U

Figure B.27 expands Figure 10 by adding the joint distribution of the SP Property and Null treatments. The distribution among the pooled data is also added for completeness (however, recall that training scores are not comparable outside the treatment groups of DA Mechanics, SP Property and Null).

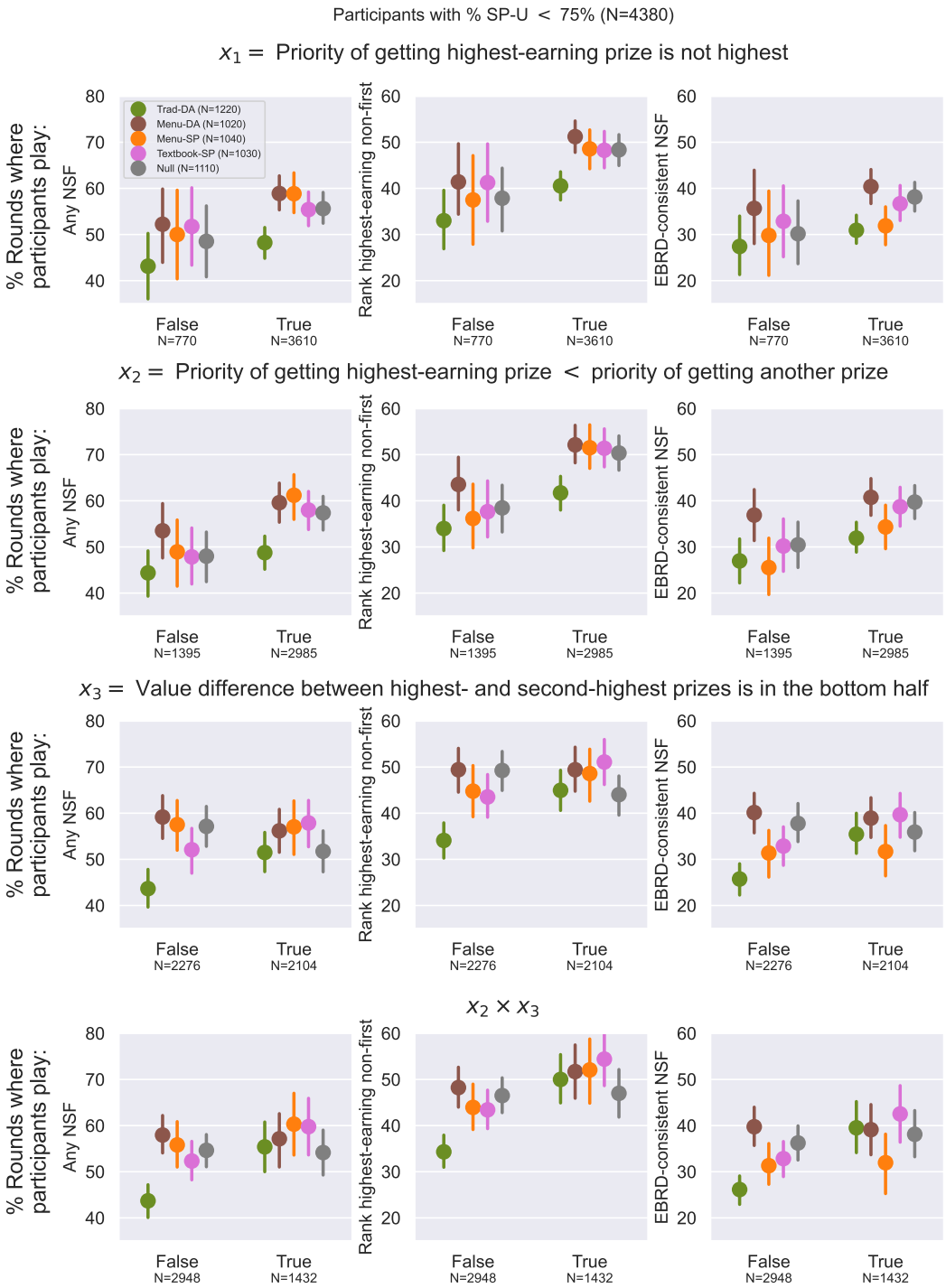
[35] EBRD is a model shown to explain NSF patterns well in some settings (see Dreyfuss et al., 2022a,b). Unlike these papers, we cannot make sharp EBRD predictions in our setttings since it was impossible for participants to calculate their probabilities of getting different prizes (since the computerized participants rankings' distribution were unknown). We instead use the property "top-choice monotonicity" from Meisner and von Wangenheim (2023), which classifies which strategies can EBRD-consistent under some conditions, vs. strategies which are never EBRD-consistent.

Figure B.24: The frequency of different strategies conditional on round parameters.



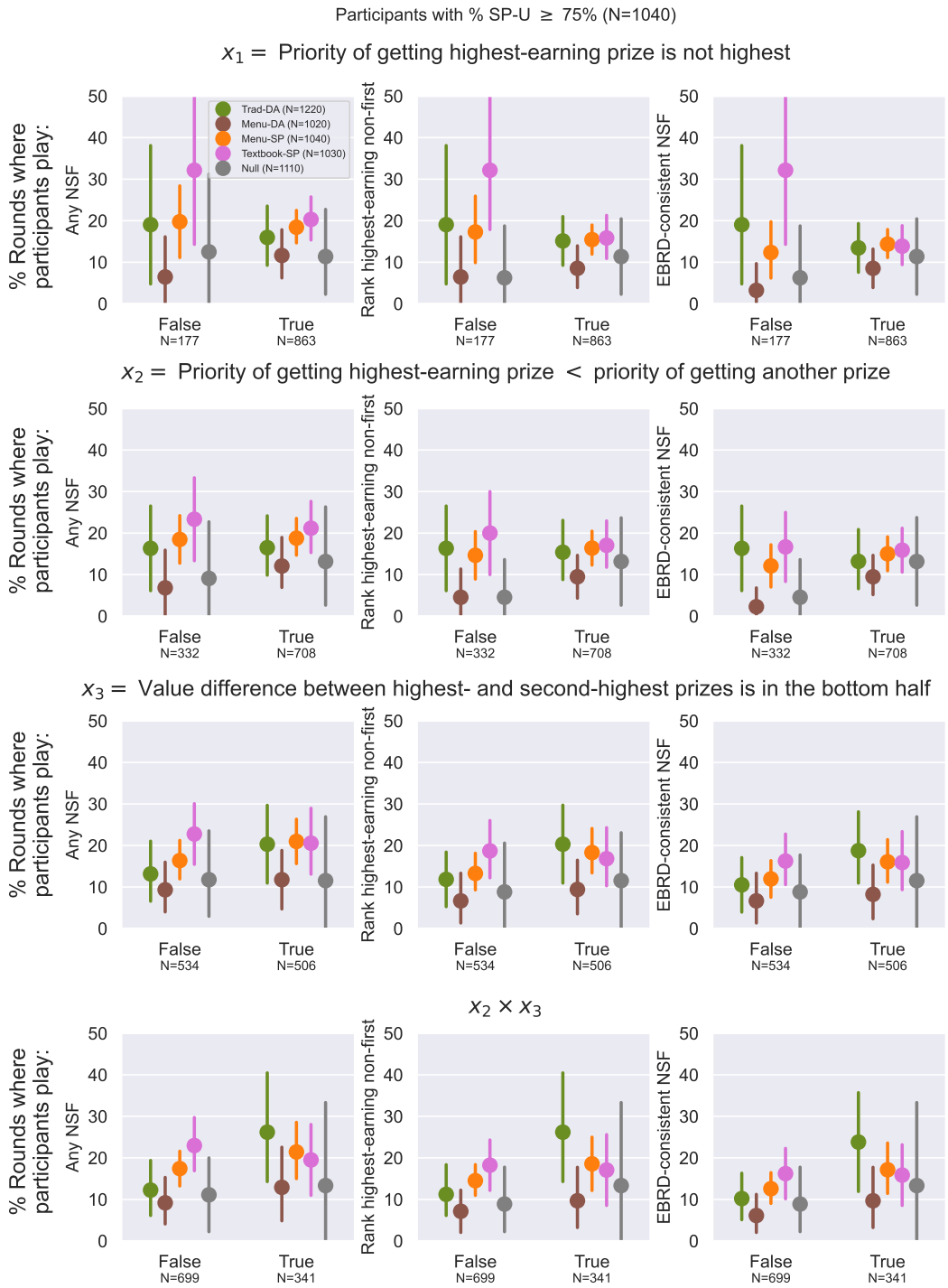
Note: Each row in the grid investigates a different binary classification of rounds according to their prize priorities and prize monetary values (resulting in a binary variable x_i coded for each round). Each column investigates the frequency of a group of strategies conditional on the row's classification. In each panel the five treatments are shown separately. Error bars: 95% bootstrapped confidence intervals.

Figure B.25: The frequency of different strategies conditional on round parameters: participants with % SP-U $<$ 75%.



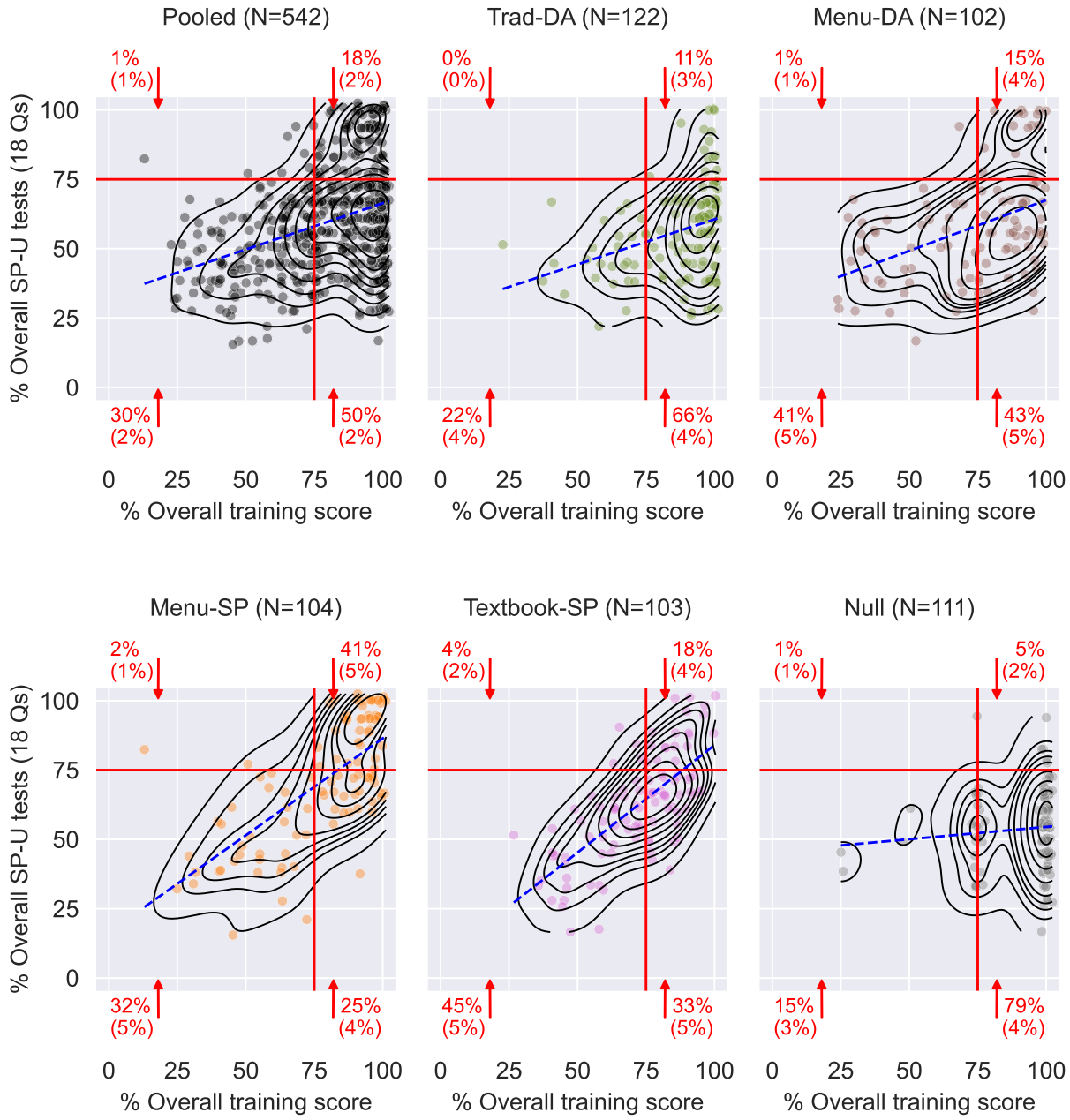
Note: Same as Figure B.24

Figure B.26: The frequency of different strategies conditional on round parameters: participants with % SP-U $\geq$ 75%.



Note: Same as Figure B.24

Figure B.27: Full joint distribution of % TR and % SP-U.



Note: See under Figure 10.

Table B.3: SF play by treatment in first five rounds (1-5) vs. in the last five rounds (6-10), and regression on round number.

	Trad-DA		Menu-DA		Menu-SP		Textbook-SP		Null	
	1-5	6-10	1-5	6-10	1-5	6-10	1-5	6-10	1-5	6-10
% SF	57	56	49	50	56	63	54	52	47	48
	(3)	(3)	(3)	(4)	(4)	(4)	(3)	(4)	(3)	(3)
% All SF	26	26	22	28	28	36	23	25	20	21
	(4)	(4)	(4)	(4)	(4)	(5)	(4)	(4)	(4)	(4)
Regression:	0.2 (0.5)		0.1 (0.5)		1.6 (0.5)		0.1 (0.5)		0.4 (0.5)	

Note: "% All SF" is the fraction of participants with straightforward play in all relevant rounds (1-5 or 6-10). Standard errors are in parentheses. "Regression" gives the regression coefficient with round number as the independent variable and whether or not the participant played SF in that round as the dependent variable.

B.9 Relationship Between % SP-U and its Sub-Measures and % SF

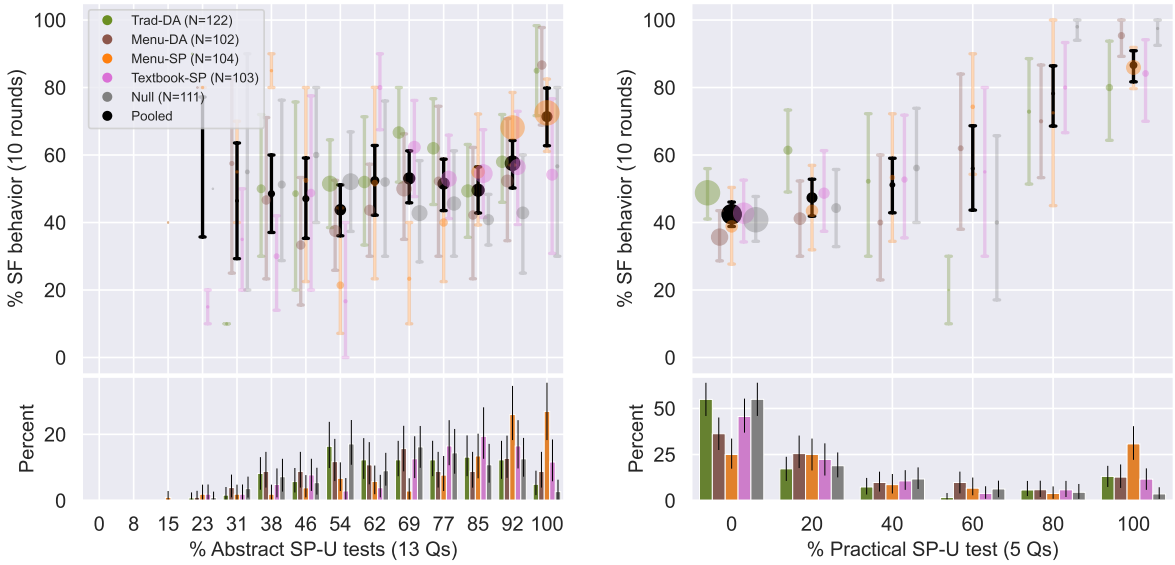
In this appendix we investigate how the different sub-measures of % SP-U contribute to its overall relation with % SF shown in Figure 11.

First, Figure B.28 shows two versions of Figure 11 focusing on % Abstract and % Practical *separately*. While % Abstract is able to explain some of Figure 11's increase in % SF as % SP-U increases, % Practical seems to explain the full increase. In contrast to the sharp increase of % SF in the high % SP-U end in Figure 11, the % SF increase as a function of % Practical alone is more gradual. The % SF estimates at the mid % Practical range are noisy and statistically indistinguishable from % SF rates of low-end % Practical scores.

Second, we investigate the extent to which % Abstract and % Practical are *jointly* required to explain the increase in % SF rates. Table B.4 summarizes OLS regressions showing the average relation between SF behavior, % Abstract and % Practical, and total % SP-U. Table B.5 add a breakdown of % Abstract into its three sub-measures % AbstractHard, % AbstractMedium, and % AbstractEasy. The tables suggest that % Practical is much more strongly related to % SF than % Abstract or its sub-measures, and that the relation between % Abstract alone and % SF is mostly due to the correlation between % Abstract and % Practical.

Despite that % Abstract has an average insignificant relation to % SF when considered jointly with % Practical, it may contribute to the overall relation in more specific cases. Fig-

Figure B.28: Relationship between % SF and the % Abstract and % Practical sub-measures of % SP-U.



Note: *Left Panel*: Histogram of % Abstract at the bottom and the conditional means of % SF sub-measure given all possible % Abstract scores at the top (see Figure 11, which is similar in structure, for more details). *Right Panel*: a similar structure with the Practical sub-measure instead of Abstract.

ure B.29 non-parametrically investigates % SF rates at different combinations of % Abstract and % Practical. Pooling all treatments on the top-left panel, the figure suggests that % Abstract is able to separate between low-SF and high-SF participants at a mid level of % Practical, between the two modes of its bimodal distribution. For % Practical between 40% and 60%, the average % SF difference between % Abstract $< 90\%$ and % Abstract $\geq 90\%$ is 31% (SE = 8%). Adding a full set of controls, this difference becomes 29% (SE = 11%). Cautiously note, however, that this pattern is estimated on a small sample of 82 participants who has these mid-level % Practical scores, and there is insufficient statistical power to test whether it persists within treatments.

Table B.4: Separate vs. joint relation of SP-U tests with SF behavior: Abstract and Practical

Dependent variable: % SF

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
% Abstract	0.26	0.07			0.11	0.01		
	(0.07)	(0.09)			(0.07)	(0.09)		
% Practical			0.43	0.41	0.41	0.41		
			(0.03)	(0.04)	(0.03)	(0.04)		
% SP-U							0.62	0.60
							(0.06)	(0.10)
Controls		X		X		X		X
Treatment		X		X		X		X
R^2	0.03	0.31	0.22	0.44	0.22	0.44	0.13	0.37
N	542	542	542	542	542	542	542	542

Note: Estimated coefficients using OLS regressions of % SF on % SP-U separated into its different measures and their relevant combinations. Each regression is shown without and with including a full set of controls in the regression, described in Section B.4. Treatment indicators are also included in the controlled regressions.

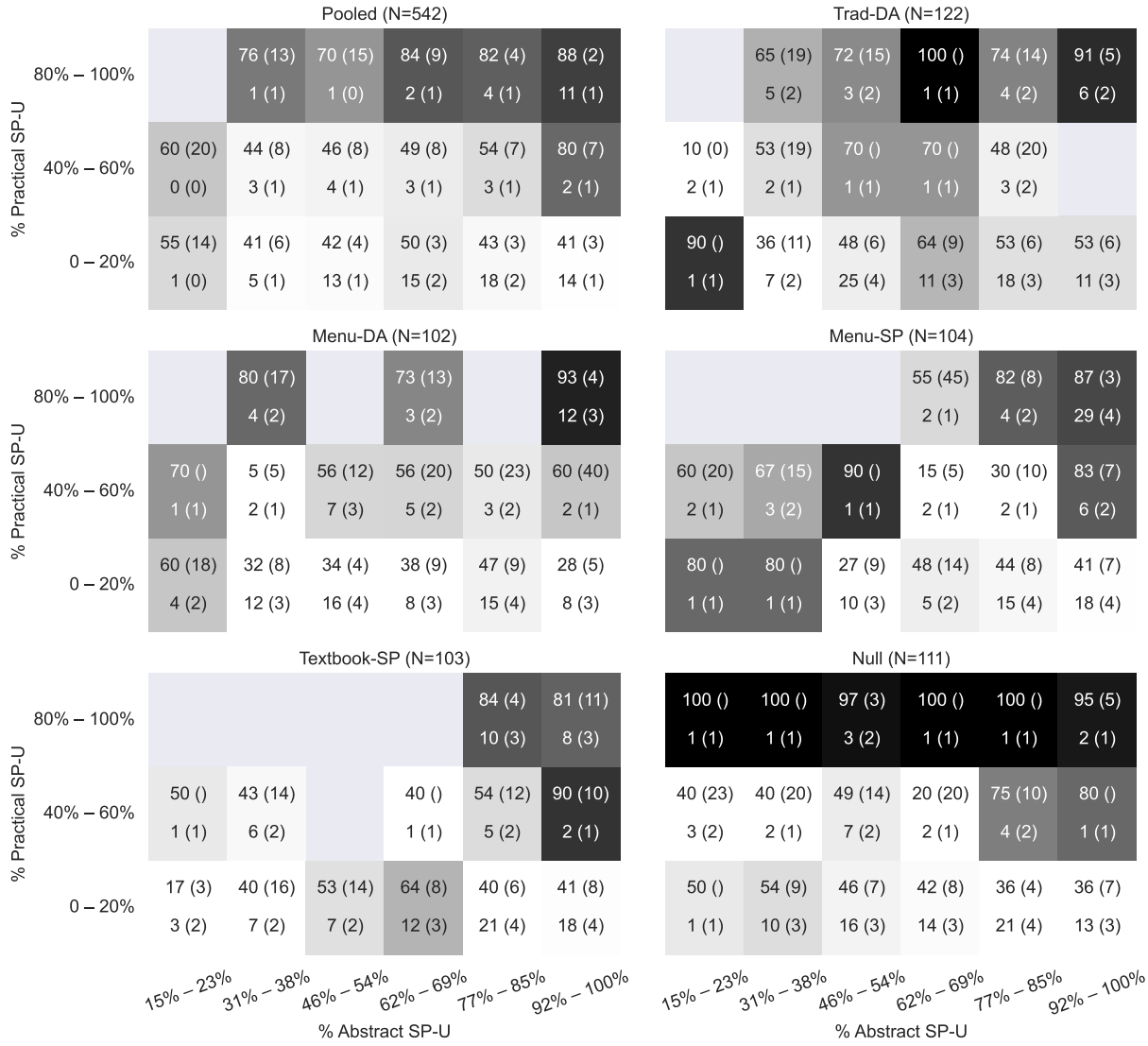
Table B.5: Separate vs. joint relation of SP-U tests with SF behavior: All 4 tests

Dependent variable: % SF

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
% AbstractH	0.15	−0.03							−0.01	−0.06		
	(0.07)	(0.08)							(0.07)	(0.08)		
% AbstractM			0.15	0.04					0.06	−0.01		
			(0.04)	(0.06)					(0.05)	(0.06)		
% AbstractE					0.17	0.12			0.05	0.07		
					(0.04)	(0.06)			(0.05)	(0.06)		
% Practical							0.43	0.41	0.41	0.40		
							(0.03)	(0.04)	(0.03)	(0.04)		
% SP-U											0.62	0.60
											(0.06)	(0.10)
Controls		X		X		X		X		X		X
Treatment		X		X		X		X		X		X
R^2	0.01	0.31	0.02	0.31	0.03	0.31	0.22	0.44	0.22	0.45	0.13	0.37
N	542	542	542	542	542	542	542	542	542	542	542	542

Note: See Table B.5.

Figure B.29: % SF rates and fractions of participants at different combinations of % Abstract and % Practical.

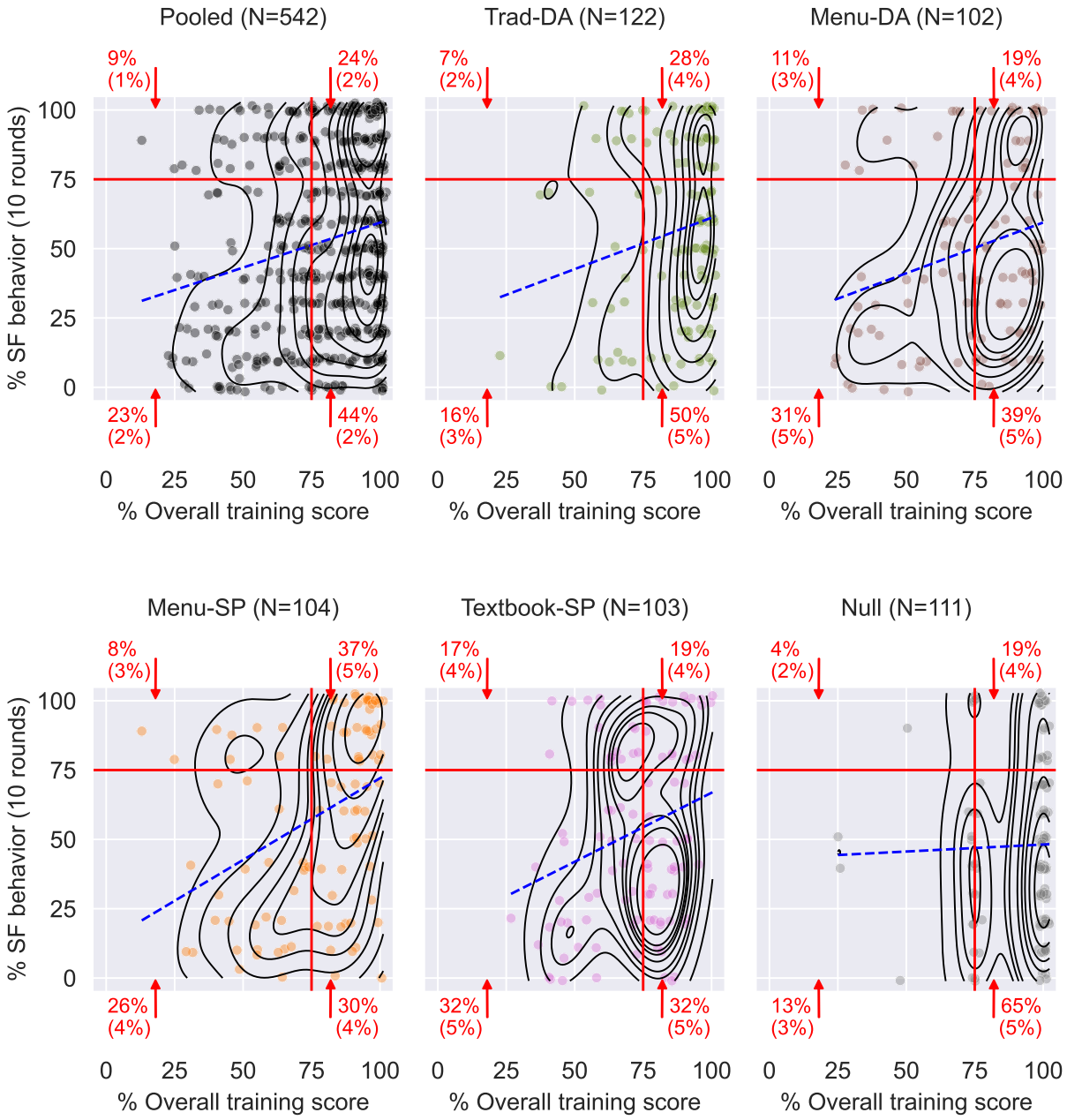


Note: In each cell, the top row indicates % SF (SE) and the bottom row indicates % participants at this cell, out of the whole panel (SE). A darker cell color corresponds to a higher % SF rate. Each panel focuses on a different treatment, or on the pooled sample (top left). % Abstract scores are presented in 6 bins, each bin including 2 adjacent scores out of the total 12 scores which have non-zero numbers of participants. Similarly, % Practical scores are presented in 3 bins summarizing the 6 possible scores in this test. Standard errors are shown in parentheses, where no SE at top row indicates a single-observation cell.

B.10 Relationship Between % TR and % SF

Figure B.30 shows the full joint distribution of % TR and % SF. The distribution among the pooled data is also added for completeness (however, recall that training scores are not comparable outside the treatment groups of DA Mechanics, SP Property and Null).

Figure B.30: Full joint distribution of % TR and % SF.



Note: See under Figure 10.

C Select Experiment Materials

In this appendix, we discuss our pre-registration, describe the full procedures of running the experiment, and provide some experimental materials for easy reference. Specifically, we describe the distribution of DA round parameters and include screenshots of the main description text and the incentivized rounds of DA.

For Appendix A, which contains screenshots of every screen of every treatment, see the supplementary material on the authors' websites.

C.1 Comparison of Analysis and Findings With Our Pre-registration

Our main pre-registred hypothesis was that different descriptions would convey strategyproofness better or worse, and thus influence % SP-U.

Across treatments, we mainly hypothesized regarding the different performance of Trad-DA and Menu-DA, and had weaker conjectures regarding the SP Property treatments.[36] First, we hypothesized that Menu-DA would make it more complicated to understand DA mechanics than Trad-DA. Second, we hypothesized that Menu-DA would outperform Trad-DA in % SP-U, since in Menu-DA strategyproofness follows from a one-sentence proof, but the same is not true for Trad-DA. Third, we hypothesized that the correlation between % TR and % SP-U in Menu-DA would be positive and stronger than that correlation for Trad-DA (and specifically, that conditional on a high percentile of % TR, Menu-DA would outperform Trad-DA in % SP-U).

We did not have clear hypotheses on how % SF would compare across treatments and on whether % SP-U and % SF behavior would correlate, since this requires assumption on participants preference.

In contrast to these hypotheses, we find little difference in our main outcome measures between Trad-DA and Menu-DA, other than a difference in % TR. However, we see significant differences between Menu-SP and Textbook-SP (and the other treatments); most significantly, Menu-SP leads to the highest rates of % SP-U in our experiment. We also see little difference in the correlations between % TR and % SP-U across Trad-DA and Menu-DA.

Our other findings are more ex-post and are less strongly related to the pre-registered hypotheses, e.g., Section 1's finding (4) on the relation between % SP-U on % SF (exhibited

[36] The treatment names we use in the paper are different than those we used in the pre-registration: Traditional DA Mechanics (Trad-DA) was named "Traditional Mechanics" ("Trad-Mech"), Menu DA Mechanics (Menu-DA) was "Menu Mechanics" ("Menu-Mech"), Menu SP Property (Menu-SP) was "Menu Property" ("Menu-Prop") and Textbook SP Property (Textbook-SP) was "Traditional Property" ("Trad-Prop"). The Null treatment was named the same.

by Figure 11).

C.2 Experiment Procedures

In this appendix, we describe the full protocol used to run the experiment on Prolific and in the Cornell Business Simulation Lab (BSL), and provide more details on the session run in each of the platforms.

C.2.1 Prolific

Protocol. Participants were recruited using the page shown in Figure C.1. Only Prolific participants who passed the pre-screening criteria of living in the US and having at least 99% past approval rate and at least 50 past completed tasks were able to see the page. Participants who wished to participate could click a button (not shown in the screenshot) that redirected them to the video conference session where the experiment took place (see below).

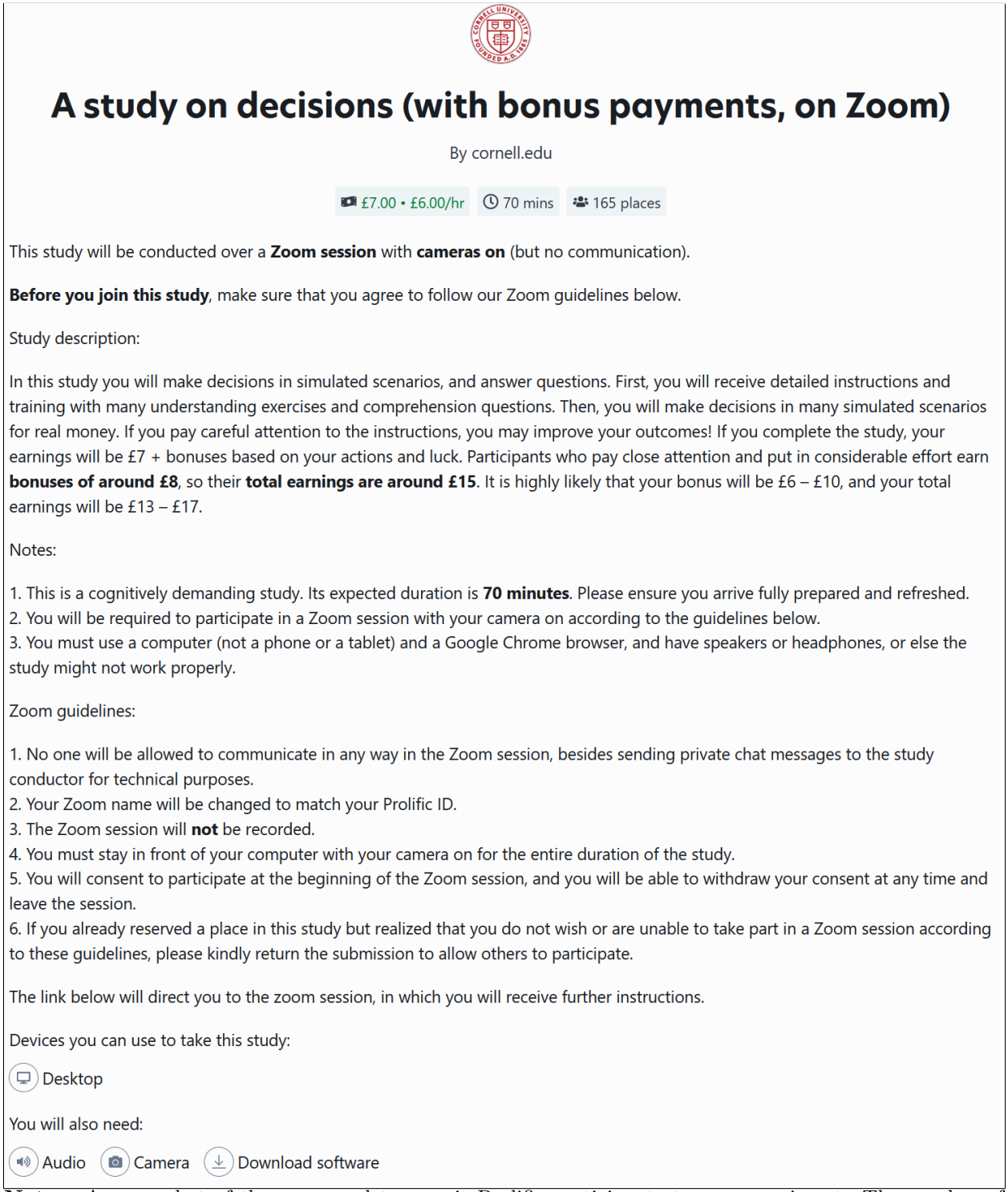
In the recruitment text, participants were explicitly informed of the length of the study and on it being cognitively demanding, and were asked to prepare accordingly. They were informed on the general nature of the experiment and on their expected payment (based on pilot runs).

Participants had to run the experiment using a PC, rather than a tablet or a phone, use a specific browser (due to technical reasons of software stability) and verify they have functional camera and audio output in their computer.

Importantly, they were required to join a video-conference session (using Zoom) with the experiment conductor in order to participate, and were asked to follow specific guidelines in the zoom session: (1) avoid any kind of communication with each other, (2) use their Prolific ID instead of their name, (3) stay visible in front of their camera but stay muted and not communicate through messages (as mentioned next, in addition to asking participants to avoid such behavior, Zoom settings were changed to technically prevent as much of it as possible).

The experiment Zoom sessions were run according to the following protocol. Early preparations prior to the session included creating a Zoom room for the session, and changing Zoom settings to disable the chat function of participants with one another, enable a waiting room, disable any kind of screen sharing, emojis, ability to unmute onself, captioning and recording. The Zoom room password (and link given to participants) was changed between sessions to prevent past participants from returning using their link.

Figure C.1: Prolific recruitment text.



A study on decisions (with bonus payments, on Zoom)

By cornell.edu

£7.00 • £6.00/hr | 70 mins | 165 places

This study will be conducted over a **Zoom session** with **cameras on** (but no communication).

Before you join this study, make sure that you agree to follow our Zoom guidelines below.

Study description:

In this study you will make decisions in simulated scenarios, and answer questions. First, you will receive detailed instructions and training with many understanding exercises and comprehension questions. Then, you will make decisions in many simulated scenarios for real money. If you pay careful attention to the instructions, you may improve your outcomes! If you complete the study, your earnings will be £7 + bonuses based on your actions and luck. Participants who pay close attention and put in considerable effort earn **bonuses of around £8**, so their **total earnings are around £15**. It is highly likely that your bonus will be £6 – £10, and your total earnings will be £13 – £17.

Notes:

1. This is a cognitively demanding study. Its expected duration is **70 minutes**. Please ensure you arrive fully prepared and refreshed.
2. You will be required to participate in a Zoom session with your camera on according to the guidelines below.
3. You must use a computer (not a phone or a tablet) and a Google Chrome browser, and have speakers or headphones, or else the study might not work properly.

Zoom guidelines:

1. No one will be allowed to communicate in any way in the Zoom session, besides sending private chat messages to the study conductor for technical purposes.
2. Your Zoom name will be changed to match your Prolific ID.
3. The Zoom session will **not** be recorded.
4. You must stay in front of your computer with your camera on for the entire duration of the study.
5. You will consent to participate at the beginning of the Zoom session, and you will be able to withdraw your consent at any time and leave the session.
6. If you already reserved a place in this study but realized that you do not wish or are unable to take part in a Zoom session according to these guidelines, please kindly return the submission to allow others to participate.

The link below will direct you to the zoom session, in which you will receive further instructions.

Devices you can use to take this study:

Desktop

You will also need:

Audio | Camera | Download software

Notes: A screenshot of the page used to recruit Prolific participants to our experiment. The number of places mentioned at the top ("165 places") is session-specific.

In addition, early preparations of the oTree environment and server running the experiment were done before each session, and included resetting all data and runtime variables.

The Zoom session was opened shortly prior to posting the recruitment message on Prolific. The session conductor was named "CONDUCTOR," and was required to enter the session using another device as "BACKUP," in case the main device disconnects.

Participants typically joined the session gradually over a few hours, but the session's protocol was designed to create an experience as similar as possible for all participants despite their different timings of joining the session. As participants entered the Zoom waiting room, they were first renamed as "Pending approval #" (where # was 1, 2, 3 and so on, in the participants' order of connection). After renaming, a participant was let in the meeting and was sent the following message: "Dear participant, please turn on your camera. Please enter the following link to read the general guidelines for this study: https://jescstudies.my.canva.site/" (in case of noting that the camera is already on, the message read "Dear participant, please enter...").

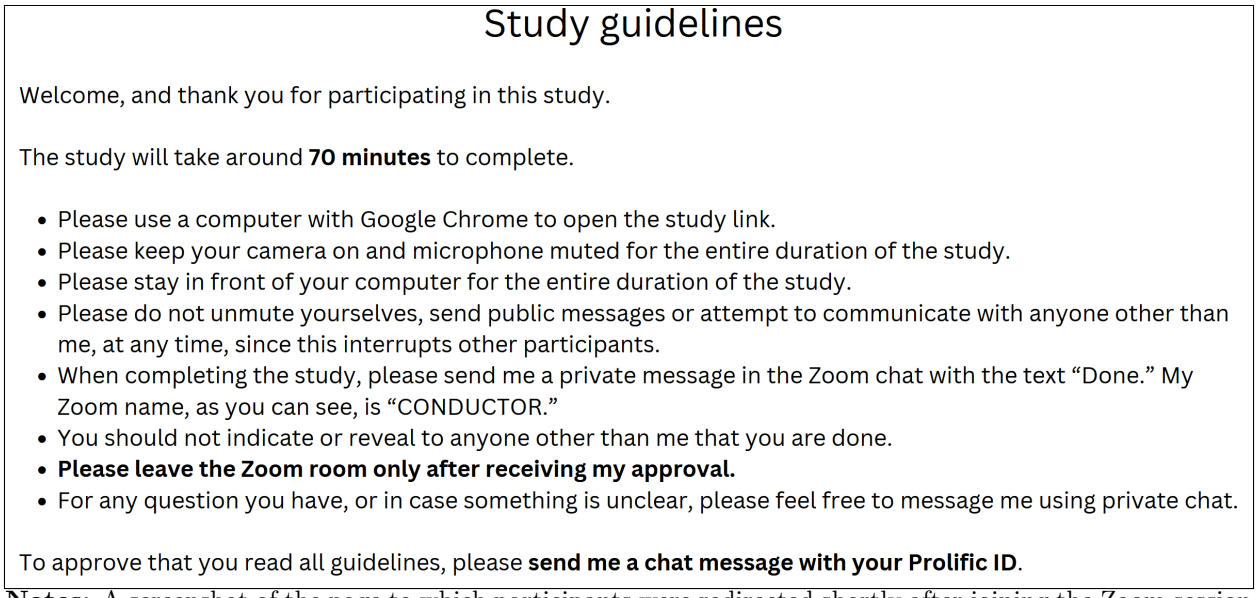
The participant then saw the slide shown in Figure C.2. The slide contained a repetition of the guidelines for the zoom session, including instructions on how and when to communicate with the session conductor to signal that they completed the experiment or to ask for assistance. Participants were asked to leave the session only upon getting approval from the conductor. Following the guidelines slide, each participant sent the conductor a message with their Prolific ID and was then sent the link to the experiment in the following message: "Thank you. Remember to stay in front of your computer with your camera on for the entire duration of the study, and send me "Done" when you complete it. The study link is: [LINK]."

As participants began the experiment, they were renamed to to their Prolific ID to help keep track of their identities inside the session, and were documented in a spreadsheet with their ID and time of starting the experiment. Typically, there was no communication with participants from this point until the point when they completed the experiment. After informing on completion to the conductor, they received the following message: "Thank you for completing the study. You can leave the meeting," and their end time was documented in the spreadsheet.

Participants who asked questions about the experiment itself were answered "Unfortunately, I can't help you with the study itself. Please try to answer to the best of your understanding." In case of a technical difficulty, the conductor provided assistance in trying to solve the problem.

After the session has ended, participants were typically paid and their completion was approved in the Prolific platform within a day.

Figure C.2: Study guidelines read by Prolific participants once entering the Zoom session.



Study guidelines

Welcome, and thank you for participating in this study.

The study will take around **70 minutes** to complete.

- Please use a computer with Google Chrome to open the study link.
- Please keep your camera on and microphone muted for the entire duration of the study.
- Please stay in front of your computer for the entire duration of the study.
- Please do not unmute yourselves, send public messages or attempt to communicate with anyone other than me, at any time, since this interrupts other participants.
- When completing the study, please send me a private message in the Zoom chat with the text "Done." My Zoom name, as you can see, is "CONDUCTOR."
- You should not indicate or reveal to anyone other than me that you are done.
- **Please leave the Zoom room only after receiving my approval.**
- For any question you have, or in case something is unclear, please feel free to message me using private chat.

To approve that you read all guidelines, please **send me a chat message with your Prolific ID**.

Notes: A screenshot of the page to which participants were redirected shortly after joining the Zoom session and before starting the experiment.

Sessions. 291 Prolific participants participated in a total of 3 Zoom sessions on August 3, 7 and 8, 2023.

The Aug 3 session was a final pilot including 38 participants. We opted (and pre-registered) to include it in our data, as at that point the experiment no longer changed except for two small bugs which were fixed toward the final two sessions.

The session on Aug 7 included 72 participants, and the session on Aug 8 included 176 participants.

C.2.2 Cornell Business Simulation Lab (BSL)

Protocol. Participants were recruited using the page shown in Figure C.1. Only students were able to register to the experiment, and could choose a future physical lab session out of a few possibilities, typically at standard work hours.

In the recruitment text, participants were explicitly informed of the length of the study and on it being cognitively demanding, and were asked to prepare accordingly. They were informed on the general nature of the experiment and on their expected payment (based on pilot runs).

In contrast to Prolific sessions to which participants could join at any time in during some time window, BSL sessions required all participants start at the same time. Sessions at a same day were scheduled at least 1.5 hours apart to enable enough time for completing the experiment (in the rare case where a participant did not finish before another session began,

Figure C.3: Cornell BSL recruitment text.

Study Name	Learning and Decisions Study
Study Type	 **Standard (lab) study** This is a standard lab study. To participate, sign up, and go to the specified location at the chosen time.
Pay	[$17/1 Credit] + avg. $8 bonus
Duration	50 minutes
Abstract	($17 or 1 credit + BONUS) Make decisions and answer questions for money
Description	** This is a cognitively demanding study. Its expected duration is 50 minutes. Please ensure you arrive fully prepared and refreshed. ** In this study you will make decisions in simulated scenarios, and answer questions. First, you will receive detailed instructions and training with many understanding exercises and comprehension questions. Then, you will make decisions in many simulated scenarios for real money. If you pay careful attention to the instructions, you may improve your outcomes! If you complete the study, your earnings will be $17 + bonuses based on your actions and luck. Participants who pay close attention and put in considerable effort earn bonuses of around $8, so their total earnings are around $25. It is highly likely that your bonus will be $6 – $10, and your total earnings will be $23 – $27. Note: You will be paid through Amazon gift cards. You may choose to earn one credit instead of the fixed participation fee of $17 (in which case any earnings beyond $17, due to bonuses, will still be paid through Amazon gift cards).

Notes: A screenshot of the page used by the Cornell Business Simulation Lab (BSL) to recruit Cornell students to our experiment.

she could stay in the lab and complete the experiment while the next session took place). Sessions were planned to include at most 20 participants each, but in practice registration was sparse and most sessions included less than 10 participants.

Participants used lab computers to access the experiment, but the experiment was run on a remote server. Prior lab preparations included opening the experiment link in the browser in all lab stations.

Each session was run using the following protocol. The session conductor read participants the following text aloud:

> Welcome and thank you for participating in this study.
>
> There is no deception in this study—everything you will read on the screen during this study is true.
>
> Please wear the headphones in your station during the entire study and follow the instructions on your screen.
>
> If you are done in less than 50 minutes you can do whatever you would like while seated in your station, but you still have to stay in your station until the 50 minutes are over. You should not indicate or reveal to anyone that you are done.
>
> If you are done and 50 minutes have already passed, please raise your hand quietly and we will come to your station to dismiss you.
>
> In case you have any question or problem, please also raise your hand quietly and we will reach out to help.

As the text mentions, participants were required to stay in the lab for at least 50 minutes, to avoid potential effects of participants randomized into shorter treatments leaving before participants who were randomized to longer treatments. If a participant joined the session late but by no more than 10 minutes, they were allowed to join and the above text was read to them again privately.

Participants were typically paid within a few days.

Sessions. 296 Prolific participants participated over 32 days including each 1–3 sessions. This large number of small sessions was required to reach our pre-registered sample size since sign-up rates were lower than expected.

It should be mentioned that in one of the days a session including 16 participants was set incorrectly, such that instead of randomizing participants into treatments they were all allocated into the Trad-DA treatment. We do not count this session towards the 50 participants-per-treatment threshold defined by our pre-registration, and instead add these

participants on top of at least 50 which were collected in correctly run sessions. In our robustness analysis we control for the day of the session and do not find it to change our results.

C.3 Randomization of DA Setting Components

Participants' earnings in our DA setting depend on their chosen ranking of prizes and on three randomly determined components: prize values for the human participant, prize priorities ("preferences of prizes"), and the ranking of the three computerized participants. These three components' values are randomly drawn from a joint distribution for each round, where each (human) participant × round's draw is independent from others. Each participant × round's randomization is performed sequentially, as follows:

1. Prize values (for the human participant) are randomly and independently determined:
    - The highest value is drawn from a uniform distribution over $\{0.90, 0.91, ...0.99\}$.
    - The second highest value is drawn from a uniform distribution over $\{0.50, 0.51, ...0.89\}$.
    - The third highest value is drawn from a uniform distribution over $\{0.10, 0.11, ...0.49\}$.
    - The lowest value is drawn from a uniform distribution over $\{0, 0.01, ...0.09\}$.
    - Each prize among Prize A, B, C and D gets one of the above values, with equal probabilities.
    - All numbers defining the support of the probability are the same across the two experiment runs in Prolific and in Cornell, but the currency is GBP (£) in Prolific and USD ($) in Cornell.
2. For each prize, priorities are sequentially determined from highest to lowest priority:
    - For all prizes except the highest-valued one from (1), each participant has an equal chance of being chosen at each priority step.
    - For the highest-valued prize, the human participant's chances are different from the computerized participants': at each priority step, each computerized participant has a r_1-times higher chance of being picked to that priority than human's chance.
    - We use a value of $r_1 = 1.7 > 1$, such that the chances of having a high priority for getting the highest prize are "rigged" against the human participant. (Note that a value of $r_1 = 1$ would generate symmetry across the human and computerized

participants, and a value $r_1 < 1$ would rig the chances in favor of the human participant).

3. For each computerized participant, their ranking is sequentially determined from top to bottom, where the chances to pick a prize at each step decay exponentially with the prize value from (1):

   - First, the participant's first (highest) rank is determined. The chance to pick each prize for this rank is proportional to r_2^{P-1}, where r_2 is a fixed parameter (see below) and P is the position of that prize when sorted from highest to lowest (i.e., for the highest-valued prize $P = 1$ and for the lowest-valued one $P = 4$).
   - Then, the participant's second rank is similarly determined, after sorting the remaining three prizes from highest to lowest, and so on until the fourth rank.
   - The r_2 parameter sets the correlation between the computerized participants' ranking and the human's participant straightforward (SF) ranking. We use a parameter value of $r_2 = 0.5$. (Note that $r_2 = 0$ would generate a perfect correlation of all rankings with the human participant's SF ranking, $r_2 = 1$ would generate zero correlation and uniformly random rankings for the computrized participants.)

The joint distribution of these elements is designed to achieve a few goals. First and most importantly, it was tested to induce sufficient variation in participants' submitted rankings—a necessary condition to identify of our treatments' effect on ranking behavior—and successfully replicates non-straightforward ranking patterns observed in previous studies (e.g., Li 2017, Dreyfuss et al. 2022b, Dreyfuss et al. 2022a).

Second, it is designed to make the setting feel "hard" and competitive, in order to encourage participants to think more carefully on their strategy. This is done by (1) giving the participant a lower probability of having first priority for getting highest-earning prize than to the computerized participants (recall that all these priorities are shown in a table each round), (2) by correlating the computerized participants' rankings with the participant's straightforward ranking,[37]and (3) setting the distribution's parameters such that there are similar likelihoods of winning all four prizes given straightforward behavior—hence making them all important to consider.

Third, the distribution we use typically induces large differences between prize values, such that conditional on NSF behavior being costly, i.e., earning-decreasing, the cost is

[37]Using a simpler, uniform distribution over both prize priorities and computerized participants' rankings instead results in an "easy" setting on average, with a 0.25 probability for obtaining the highest priority for getting the highest-earning prize.

significant. In most rounds of our experiment, the most common NSF ranking pattern of flipping the ranking of the highest- and second-highest-earning prizes is not costly. However, participants are neither informed on this nor can easily calculate this, as they do not know the computerized participants' rankings or their distribution. Such NSF ranking pattern can only be costly in situations where ranking the highest-earning prize first would win it, and making such situations frequent would interfere with our goal of creating competitive settings, where it is unlikely to win the highest-earning prize. Evidence from Dreyfuss et al. (2022a)—that NSF behavior is most prevalent when the probability of winning the highest reward is low, i.e., when such behavior is unlikely to affect the participant's final outcome and hence is not costly—suggests that increasing the expected cost of NSF behavior would also interfere with maintaining overall sufficient rates of such behavior. In order to promote better identification, we opt for high rates of non-costly NSF behavior over lower rates of costly NSF behavior.

C.4 Select Screenshots of the Experiment

We now give screenshots of some key elements of our experiment. (As mentioned, our Full Experimental Materials Appendix (Appendix A) can be found on the authors' websites.) Note that in all the description screens, successive parts of the screen appeared one-by-one as the participant advanced.

Figure C.4 and Figure C.5 show the description provided in the Traditional DA Mechanics treatment. Figure C.6 and Figure C.7 show Menu DA Mechanics. We remark that participants are not necessarily expected to understand the details of this description immediately; as discussed in Section 2.3, they receive substantial additional coaching throughout the training rounds.

Figure C.8 and Figure C.9 show the main description texts of our two SP Property treatments. Similarly to the DA Mechanics treatments, participants receive additional coaching on this description as they complete the training rounds, as discussed in Section 2.3.

Figure C.10, Figure C.11, and Figure C.12 show the full text of the Null treatment in the main description section of the experiment. Note that this text is identical to the Null description which was the initial screens of all treatments (excluding the small bit of text at the beginning noting that information will be repeated).

Figure C.4: Traditional DA Mechanics description main text, part 1.

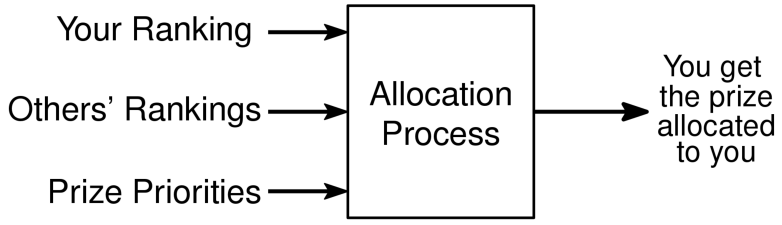
How the Allocation Process Works

You will now learn the full technical details of the allocation process.

These details are important to learn: *You may be able to apply your knowledge of them to make better decisions in the upcoming real rounds of this study.*
Some details may seem confusing at first. This is quite natural! But don't worry, we will show you step-by-step examples. Things will become clearer along the way.

Overview of allocation process

The following image illustrates how your own ranking, the rankings of the other participants and the prize priorities affect the prize you get:



Details of allocation process

The allocation process is a multi-step process, as follows (it may look complicated, but don't worry, we will rehearse this in a moment):

*In the first step, each participant is paired to their **highest**-rank prize.*

In the next step, possible conflicts are detected and solved.
*If two (or more) participants are paired to the same prize, this is a **conflict**.*
Each conflict is solved in two steps:

- ***Unpair**: only the participant highest in that prize's priorities remains paired to that prize. The others get unpaired.*
- ***Re-pair**: all unpaired participants can only get re-paired to prizes that they were not paired with before. Each unpaired participant is re-paired to their **highest-rank** prize among the prizes they **were not yet paired with**.*

Figure C.5: Traditional DA Mechanics description main text, part 2.

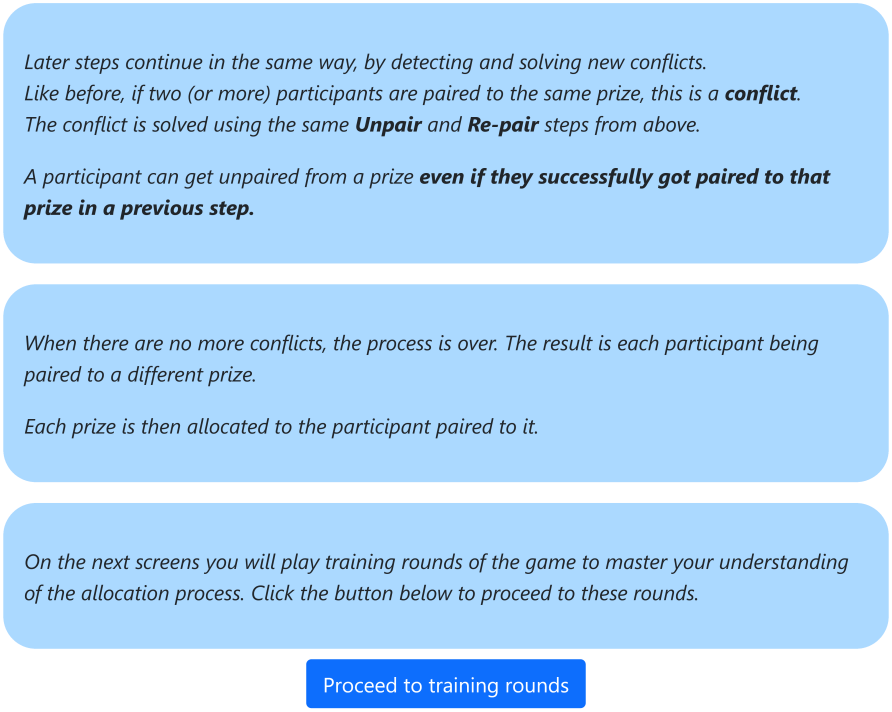


Figure C.6: Menu DA Mechanics description main text, part 1.

How the Allocation Process Works

You will now learn the full technical details of the allocation process.

These details are important to learn: *You may be able to apply your knowledge of them to make better decisions in the upcoming real rounds of this study.*
Some details may seem confusing at first. This is quite natural! But don't worry, we will show you step-by-step examples. Things will become clearer along the way.

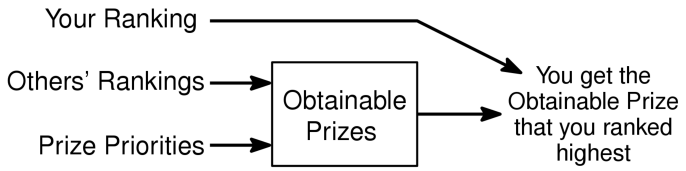
Overview of allocation process

The prize you get is determined using an allocation process with two main steps:

1. *The computer determines some group of* ***Obtainable Prizes*** *that you might receive. Your own ranking does not influence the Obtainable Prizes. Instead, they are determined using only the prize priorities and the rankings of the other participants.*
2. *You get the Obtainable Prize that you* ***ranked highest*** *(in the ranking you submitted).*

The important principle*: Your own ranking does* ***not*** *influence what the Obtainable Prizes are, but it* ***does*** *determine what you get from among the Obtainable Prizes—you get the Obtainable Prize that you ranked the* ***highest.***

The following image illustrates how your own ranking, the rankings of the other participants and the prize priorities affect the prize you get:



Details of allocation process

Priorities and rankings → Temporary allocation → Obtainable Prizes

The allocation process begins with a multi-step process. This process determines a "temporary allocation" of prizes to all participants ***except for you****, and then determines your Obtainable Prizes based on this temporary allocation. This process* ***does not involve your own submitted ranking****, and works as follows (it may look complicated, but don't worry, we will rehearse this in a moment):*

In the first step, each prize is paired to its ***highest****-priority participant, among all participants* ***except for you.***

Figure C.7: Menu DA Mechanics description main text, part 2.

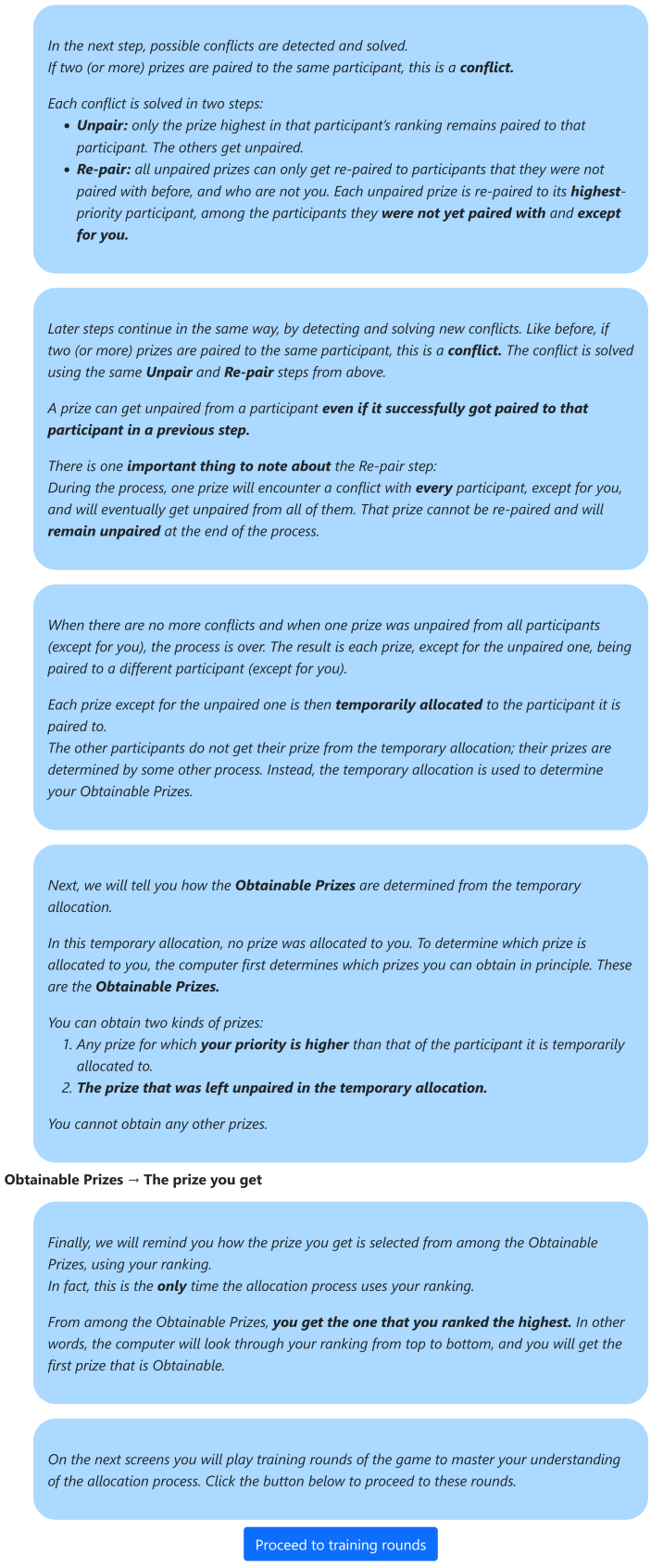


Figure C.8: Menu SP Property description main text.

The Key Principle of Your Ranking

We will now tell you a general important principle of how your own ranking affects the allocation process.

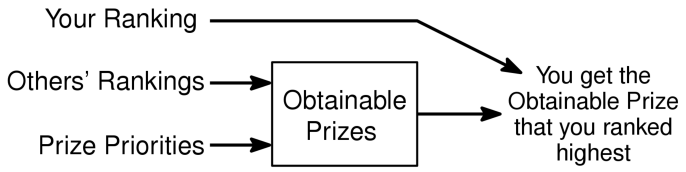
This principle is important to learn: *You may be able to apply your knowledge of it to choose your rankings in the upcoming real rounds of this study.*

The prize you get is determined using an allocation process with two main steps:

1. *The computer determines some group of **Obtainable Prizes** that you might receive. Your own ranking does not influence the Obtainable Prizes. Instead, they are determined using only the prize priorities and the rankings of the other participants.*
2. *You get the Obtainable Prize that you **ranked highest** (in the ranking you submitted).*

***The important principle**: Your own ranking does **not** influence what the Obtainable Prizes are, but it **does** determine what you get from among the Obtainable Prizes—you get the Obtainable Prize that you ranked the **highest**.*

The following image illustrates how your own ranking, the rankings of the other participants and the prize priorities affect the prize you get:



For example, imagine that your Obtainable Prizes are C and D. If you submit the ranking A–B–C–D, you will get Prize C, which is the one you ranked highest among the Obtainable Prizes. No ranking you could possibly submit would get you Prize A or Prize B, since the Obtainable Prizes are C and D, and since your own ranking cannot influence the Obtainable Prizes.

On the next screens you will play training rounds of the game to master your understanding of this principle. Click the button below to proceed to these rounds.

Proceed to training rounds

Figure C.9: Textbook SP Property description main text.

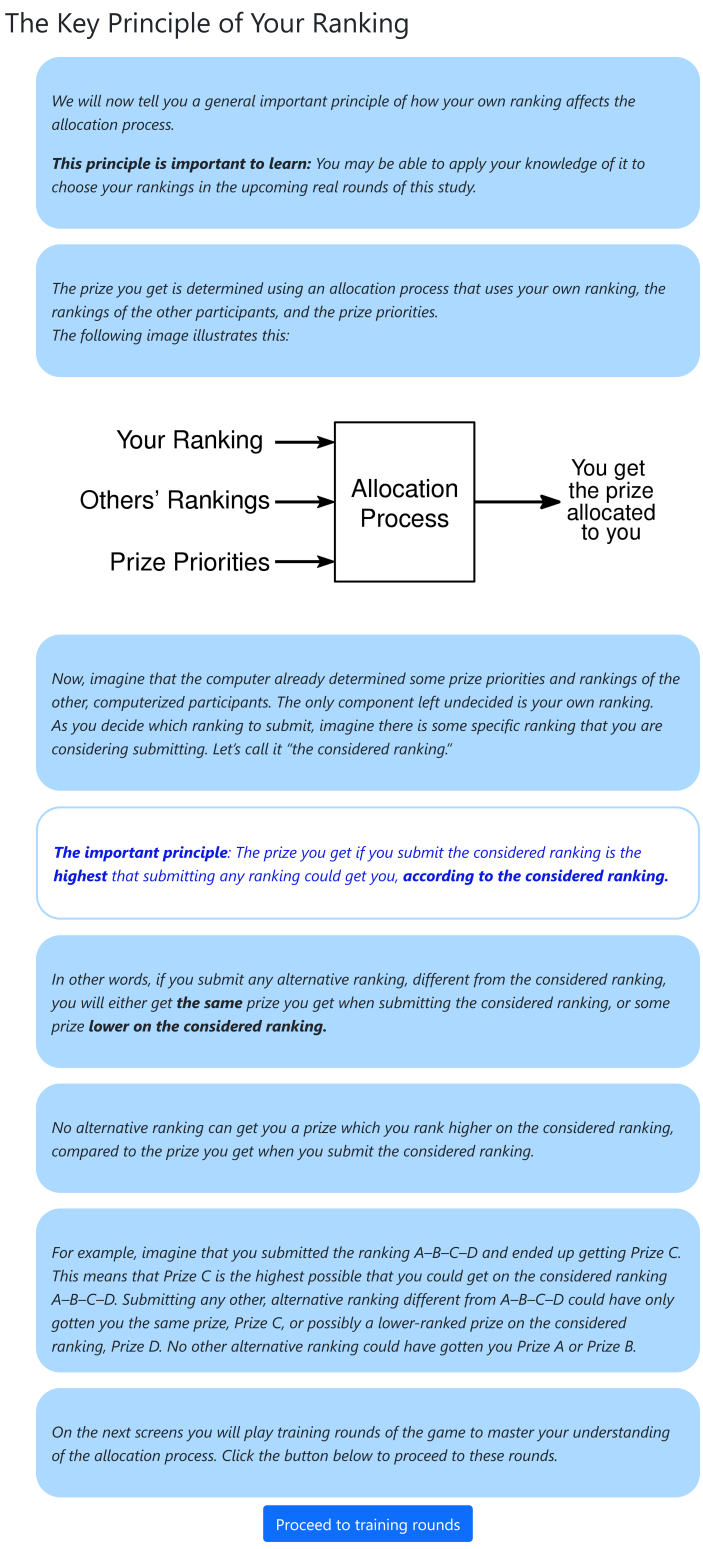


Figure C.10: Null treatment description main text, part 1.

The Basics

*We will now remind you of the general details of the game and allocation process. This page includes the **exact same text** you read at the beginning of the study. Please read it again to make sure you understand.*

In this study, you and three computerized participants, Ruth, Shirley, and Theresa, are going to play a game for four prizes. Each prize is worth money, but might be worth a different amount of money for each participant.

*You and the computerized participants will each rank the four prizes in any order you wish. Then, an **allocation process** will use these rankings to allocate the prizes—one prize for each participant.*

The allocation process attempts to give each participant a prize that they ranked higher rather than a prize that they ranked lower. However, this is not always possible, since the allocation process must take into account the rankings of all participants.

*You will now learn about the game and allocation process. The text in blue bubbles, such as this text, provides explanations about the game and allocation process. **Make sure you read it carefully**.*

A round of the game has three steps.

Step 1: Round Information

*In Step 1, you first see the **prizes** you can get in this round and how much money they are worth to **you**.*

In this round, the **prizes** are:

Prize	A	B	C	D
Money worth	67p	7p	95p	45p

Figure C.11: Null treatment description main text, part 2.

*In the table above, under each prize A, B, C or D, you can see how much money it would add to **your** earnings.*

*Each prize might be worth a **different** amount of money for each participant, and each participant can only see the money amounts relevant to **themselves**. However, the prizes that earn **you** a large amount of money are also likely to earn the **other participants** a large amount of money. There is more likely to be **competition** for the high-earning prizes.*

*The money worth of prizes for you and for the other participants can be different in different rounds of the game, and they were **determined beforehand**. You and the other participants **cannot affect the money worth of prizes**.*

*Still in Step 1, you see next what we call the **prize priorities**.*

*All four participants have some **priority** for getting each of the four prizes.*

These priorities can affect the allocation of prizes.
The higher your priority is for getting some prize, the more likely you generally are to get that prize at the end of the process.

In this round, the **prize priorities** for you and for the other participants are:

Prize	A	B	C	D
1st priority (highest)	Ruth	Shirley	Theresa	You
2nd priority	Shirley	You	Shirley	Theresa
3rd priority	You	Theresa	You	Ruth
4th priority (lowest)	Theresa	Ruth	Ruth	Shirley

Each column shows the priorities of all participants for getting some prize, written from highest to lowest

*The prize priorities can be different in different rounds of the game, and they were **determined beforehand**.*
*You and the other participants **cannot affect the prize priorities**.*

Figure C.12: Null treatment description main text, part 3.