NBER WORKING PAPER SERIES

DESCRIBING DEFERRED ACCEPTANCE AND STRATEGYPROOFNESS TO PARTICIPANTS: EXPERIMENTAL ANALYSIS

Yannai A. Gonczarowski
Ori Heffetz
Guy Ishai
Clayton Thomas

Working Paper 33020
http://www.nber.org/papers/w33020

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2024

The authors thank Keren-Or Barashi Gortler, Itamar Bellaiche, Yehonatan Caspi, Yael Cohen, Gabriela Cohen-Hadid, Ayala Goldfarb, Michael Khalfin, Ido Leshkowitz, Josef Mccrum, Shenhav Or, Yonatan Rahimi, and Ohad Weschler for excellent research assistance; Eric Budish, Ben Enke, Nicole Immorlica, David Laibson, Markus Mobius, 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, EC 2024, Communicating Clearly to Market Participants Workshop (Stony Brook), and seminar participants at Bar Ilan, Cornell, the Hebrew University, and Microsoft Research 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. Pre-registration of our experiment can be found at https://aspredicted.org/7eq7e.pdf. A one-page abstract of this paper appeared in the proceedings of EC 2024. 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 views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2024 by Yannai A. Gonczarowski, Ori Heffetz, Guy Ishai, and Clayton Thomas. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Describing Deferred Acceptance and Strategyproofness to Participants: Experimental Analysis
Yannai A. Gonczarowski, Ori Heffetz, Guy Ishai, and Clayton Thomas
NBER Working Paper No. 33020
October 2024
JEL No. C78, C91, D47, D91

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.

Yannai A. Gonczarowski
Harvard University
Department of Economics and
Department of Computer Science
Cambridge, MA
yannai@gonch.name

Guy Ishai
The Hebrew University of Jerusalem
Bogen Family Department of Economics
and Federmann Center for the Study of
Rationality
guy.ishai@mail.huji.ac.il

Ori Heffetz
S.C. Johnson Graduate School of Management
Cornell University
324 Sage Hall
Ithaca, NY 14853
and The Hebrew University of Jerusalem and
also NBER
oh33@cornell.edu

Clayton Thomas
Microsoft Research
thomas.clay95@gmail.com

A randomized controlled trials registry entry is available at https://aspredicted.org/7eq7e.pdf

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. It thus promotes ease of participation, reduces inequality across those with different levels of strategic sophistication, and increases robustness of theoretical predictions on play. However, strategyproofness achieves all this only if 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 & Kübler, 2021; Rees-Jones & 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 ranking.[3]

[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, Heffetz, and Rabin (2022) 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 discussion in Section 4.

[3] See, e.g., Guillen and Hing (2014), Rees-Jones and Skowronek (2018), Guillen and Veszteg (2021), and

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 description of (the mechanics of, i.e., 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)

Masuda et al. (2022); for details, see Section 4.

mechanisms where strategyproofness holds via a simple argument. 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 argument: 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 compare participants' reactions to such a description with their reactions 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 the first treatment. 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 mechanism has a menu description, this treatment conveys no more information than that the mechanism is strategyproof.)

[4]Katuščák and Kittsteiner (2024) 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.

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, we distinguish between *advice*—such as recommending that participants play the SF strategy—which all our treatments avoid; and *the strategyproofness property*—a particular property of the mechanism, i.e., of the mapping from reported rankings to allocations—that the SP Property treatments aim to directly teach. These 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 short, 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.

In all five treatments, the training module—which differs by treatment—is followed by an incentivized DA-rounds module—which is *identical* across treatments (in both interface and underlying mechanism). 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 samples: 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 results replicate across the two 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 at 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 in Trad-DA (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 work.

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., "...imagine that you have submitted the ranking B–A–C–D [of four possible prizes A, B, C, D], and got Prize C... If you had instead submitted A–B–C–D... is it possible (or certain) that you would have gotten Prize A? Yes/No"); and (ii) five *practical* questions asking participants about the implications of strategyproofness (e.g., "If I want to maximize my earnings in a given round, then... Sometimes I might have to rank the prize that earns me the most in second place or lower. True/False"). In Trad-DA, mean overall SP-understanding score is 56%, and in Menu-DA it is 58% (both SEs = 2%). Despite strategyproofness holding via a one-sentence argument 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 at 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 an SP Property treatment, and in particular 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 in both the abstract and the practical questions. We also find that scores on the SP-understanding test, particularly the practical ones, 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 overall distributions of SF play. In particular, the rate of SF play has a spread-out, somewhat bimodal distribution 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).[5] 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 75–80%, SF behavior rates are dramatically higher (mostly 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 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 appendices include further analysis. We find that in addition to its relatively low SP-understanding scores, Menu-DA makes DA mechanics substantially more complicated for participants relative to Trad-DA—as reflected by training scores and success at DA-outcome calculations using our GUI. We investigate the joint distribution of participants' performance on the abstract and practical sets of questions of SP understanding and find that in our setting, abstract understanding of strategyproofness is mostly a necessary condition (albeit an insufficient one) to understanding its practical implications. We investigate how participants' behavior relates to SP-understanding in more detail, and find that success at the practical questions is a far better predictor of SF behavior compared to the abstract questions. We also show that our results are robust to controlling for a rich set of demographic characteristics and session-date fixed effects. Our findings also replicate across samples—while, notably, baseline understanding levels, as well as treatment effects, are lower on Prolific than at Cornell.[6]

Overall, 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 at teaching the former,

[5] 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.

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

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 as follows. 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. In particular, we see promise in our new Menu SP Property description, for both future research and real-world applications.

2 Experimental Design

Our experiment studies 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.

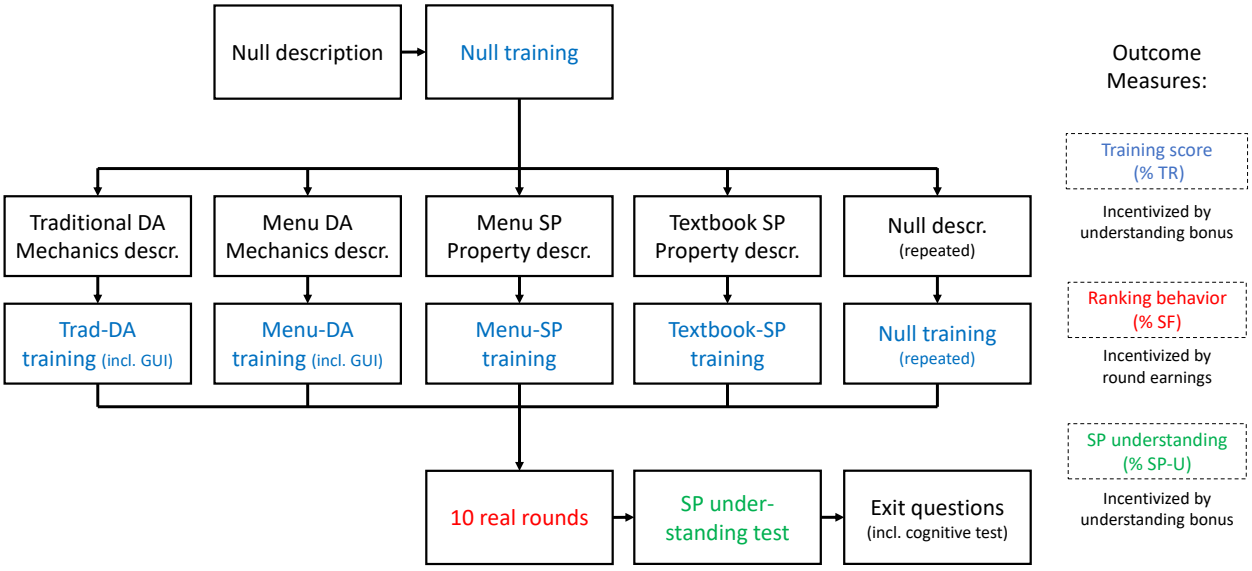
Our experiment (a) introduces the matching environment; (b) describes the matching mechanism—"allocation process"—using one of our five descriptions; (c) attempts to improve, and tests, participants' understanding of the mechanism using novel training questions and/or GUI; (d) elicits behavior in standard, incentivized DA rounds; (e) elicits understanding of strategyproofness using novel tests; and finally (f) elicits reflections, perceptions, cognitive-ability 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 increase their attention, and to incentivize each of our main outcome variables. Participants are shown their cumulative DA-round earnings after each round; their understanding bonus is only 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 the 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 allocation process—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 (D. L. Chen et al., 2016). The rest of this section describes the experiment parts in chronological order. For screenshots of all screens in all treatments, see Appendix D in the Supplementary Materials.

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 allocation process and how the outcomes will earn them money, without any description of how the actual allocation will be determined. We call this the "Null description."

The Null description presents the matching environment as a static setting with four participants: one human participant (labeled "You," to whom we refer below simply as "the participant" when no confusion can arise) and three computerized participants (labeled "Ruth," "Shirley," and "Theresa"). 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 participant is informed that the prizes are worth different amounts of money to the different participants, and in each round the participant is shown how

much each prize earns them. (See Figure 5 in Section 2.4 for an example real round. The participant's prize values are presented immediately under "Step 1: Round Information" in the interface. The rest of the interface shown in the figure is now described.)

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.[7]

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). These practice rounds also include an attention-check question.

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.1 in the Supplementary Materials.

In the 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., Y. Chen and Sönmez, 2006; NMS, 2020). In the 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 partici-

[7]These intuitive descriptions read as follows. For rankings: "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." For priorities: "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."

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 submitted rankings 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 include each prize which your priority of getting is higher than that of the participant paired to it in the temporary allocation. You will receive your highest-ranked Obtainable Prize.
Menu SP Property (Menu-SP): Some set of Obtainable Prizes will be calculated using the submitted rankings 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 ranking L is always *at least as high*, according to ranking L, compared to the prize you would receive submitting another ranking.
Null: [A repeat of the Null description initially shown to all participants in order to convey the basic environment (see Section 2.1).]

pant's ranking, in order to determine a menu of prizes; and (2) the human participant gets their highest-ranked prize out of the menu.

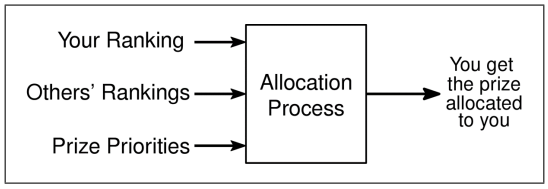
In the Menu-SP description, the participant is informed only that the mechanism is strategyproof, by explaining that *some* menu of prizes will be determined without using the participant's ranking (with no details on how this is done), and the participant will get their highest-ranked prize out of the menu. In the Textbook-SP 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 ranking gets you a prize at least as high on that ranking compared to any different ranking.

Each non-Null description is also accompanied by a chart that conveys either: the fact that the participant's 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.

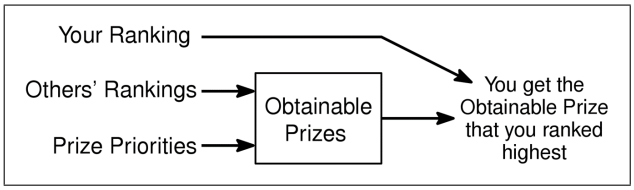
Our Null treatment gives no additional information on the allocation mechanism. Instead, it repeats the Null description discussed in Section 2.1.

Figure 2: Charts illustrating to participants the overall allocation process

(a) Trad-DA and Textbook-SP



(b) Menu-DA and Menu-SP



SP Property vs. ranking advice. 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.

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 rejections, 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 strategyproofness via different (though equivalent) definitions, 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 language as we reasonably can).

2.3 Training Rounds and Training Score (by Treatment)

After reading the description, the participant completes a 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, in non-Null treatments, the participant encounters three rich training rounds after the description screen. These training rounds differ significantly across treatments, but are designed to be as comparable as possible within the DA Mechanics or SP Property treatment groups. Unlike in real rounds, in training rounds the 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.[8] During these screens (and the real-round screens that follow), the participant is able to open a pop-up window to remind them of the text they saw in the description screen.

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 with the help of our specially developed allocation-calculation GUI. (In Menu-DA, they are also asked to calculate their menu.) The first training round includes a detailed step-by-step walk-through, which guides the participant in calculating the allocation, based on the DA Mechanics description screen 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 that round, but the participant must

[8]The most complicated, multi-step questions of the DA Mechanics treatments are counted as 5 regular questions. Thus, the participant earns 5 points (i.e., the equivalent of five correct answers) toward their understanding bonus for answering correctly on their first attempt. In those questions only, the participant also earns some bonus for answering correctly on their second attempt, of 2 points.

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.

Figure 3 shows screenshots of the main components of the DA Mechanics training rounds, including the allocation-calculation GUI. The figure also includes links to the optional GUI training video included in the second training round. In Trad-DA (respectively, Menu-DA), the participant iteratively clicks on the purple boxes signifying participants (prizes) to tentatively pair them with prizes (participants).

As Figure 3 shows, in Trad-DA, participants use the GUI to calculate the outcome of participant-proposing DA. In Menu-DA, the training rounds have three steps. First, participants use the allocation-calculation GUI to calculate the outcome of prize-proposing DA excluding the human participant; second, they use this to calculate the menu; third, they pick their highest-ranked prize from this menu. To somewhat mirror Menu-DA's second and third steps, Trad-DA also includes further questions asking participants to re-state the full allocation and the prize they receive.

In both DA Mechanics descriptions, if the participant does not correctly answer a question involving the GUI correctly within three attempts (in the first round, a question is one step of the calculation; in the second and third rounds, it is the entire GUI calculation), they are given the answer (which, in the second and third rounds, they are able to reference for the remainder of the training round). For questions not involving the GUI, the participant must enter the correct answer, regardless of the number of attempts, in order to proceed.

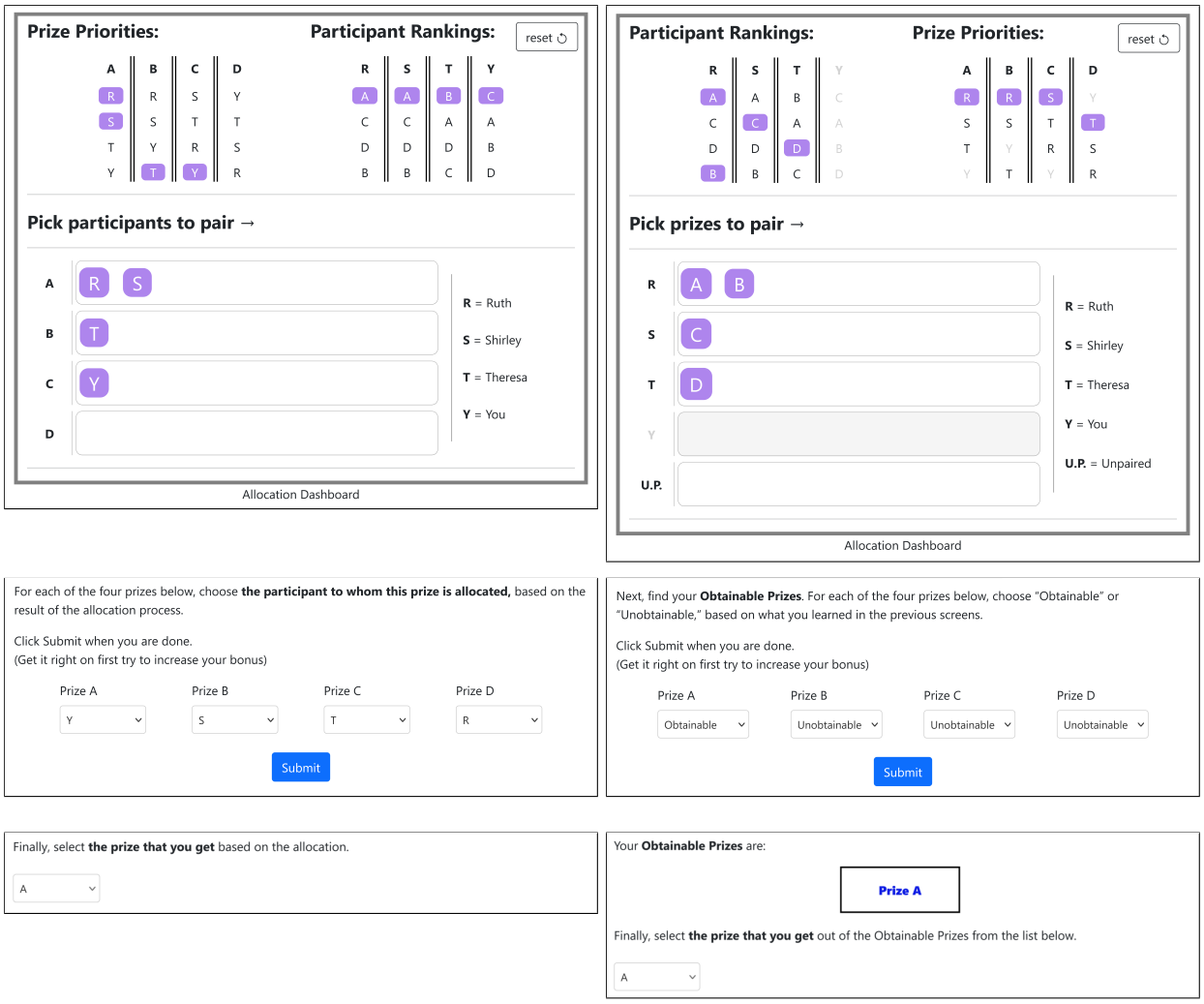
The same DA scenarios (i.e., the human and computerized participants' rankings and the prize priorities) are presented in both Trad-DA and Menu-DA; see Appendix C.2 in the Supplementary Materials for the scenarios used in the first and third training rounds (Figure 3 shows the second-round scenario). These scenarios are chosen so that running each of these algorithms (whether participant- or prize-proposing) would be of comparable complexity.

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, and the participant is informed of the calculated allocation. Then, the participant is asked a series of multiple-choice questions requiring application of the definition of strategyproofness, based on the SP Property description they received. They are asked a few questions aimed to reveal

Figure 3: Main components of DA Mechanics training

(a) Trad-DA allocation-calculation GUI and allocation entry screens

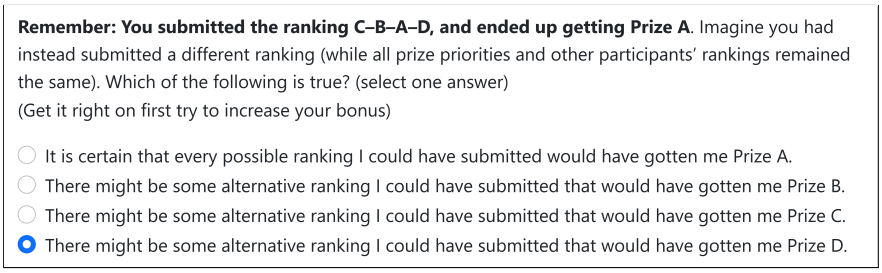
(b) Menu-DA allocation-calculation GUI, menu entry screen, and final allocation entry screen



Notes: The figure shows each treatment's second training round; the allocation-calculation-GUI snapshots at the top are taken from the optional instructional videos available in this round. These training videos are available at https://youtu.be/qK9JL32oxJg (Trad-DA) and https://youtu.be/dYePFVdqm5I (Menu-DA). *Allocation-calculation GUI*: The four prizes are indicated by A, B, C and D, the human participant is indicated by Y ("You") and the three computerized participants are indicated by their initials R, S, and T. In Menu-DA, Y is grayed out to emphasize that the prize-proposing DA is calculated while excluding the human participant. In Trad-DA (Menu-DA), below "Pick participants (prizes) to pair," each row indicates a prize (participant) and the purple boxes in that row indicate the participants (prizes) currently paired to it. The human participant can click the purple boxes to repair prizes (participants) to different participants (prizes). In Menu-DA, the row "U.P." is used to leave one prize unpaired at the end of the prize-proposing DA.

some common misconceptions about strategyproof mechanisms, and many questions about counterfactual outcomes, had they submitted other rankings in the context of the current allocation scenario.[9] See Figure 4 for an example.

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.[10] 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 the two Null training rounds discussed in Section 2.1. The participant's training score is calculated based on their performance in this second repetition of the Null training rounds. To maintain three training rounds in all treatments, the Null treatment also includes a third training round that merely provides additional experience with the setting, and does not include training questions.

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.

[9]The misconceptions which the first questions ask about 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.

[10]The exceptions are: (1) the first training question is different in Menu-SP vs. Textbook-SP, because it 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.

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 Appendix C.3 in the Supplementary Materials). This distribution is designed to encourage an ample fraction of participants—particularly those who fail to understand that the mechanism is strategyproof, or who have non-classical preferences—to carefully consider their ranking and thus play NSF. For example, the distribution is meant to make the setting feel competitive by giving the human participant a lower-than-uniform probability of having a high priority for getting their highest-earning prize, and by making 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 outcomes meaningfully affect a participant's earnings.

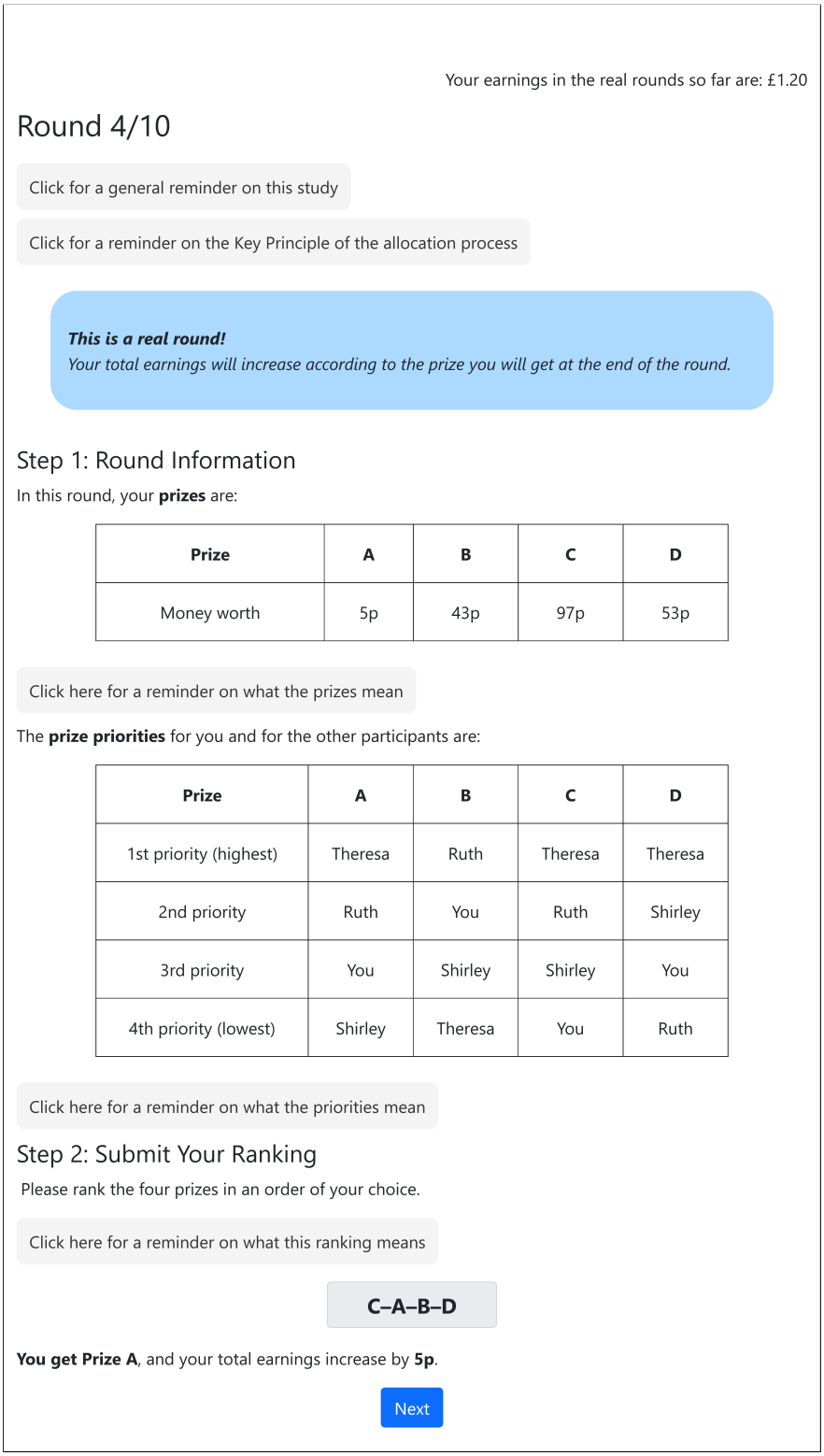
Information and feedback. As discussed in Section 2.1, at the beginning of each round, the participant observes each prize's value for them and all prize priorities. We present the prize priorities both to imitate a real-world context in which the participant has an idea about their likelihood to get 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; see Figure 6 for samples of questions. Their test score is our final main outcome variable, denoted % SP-U. The test consists of four screens, each with 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 screens of the strategyproofness understanding test are on abstract logical properties of strategyproofness, and we refer to their thirteen questions jointly as Abstract; see Figure 6(a) for examples. These questions are similar in nature to the comprehension questions asked during the SP Property training rounds, but broader in scope, and common

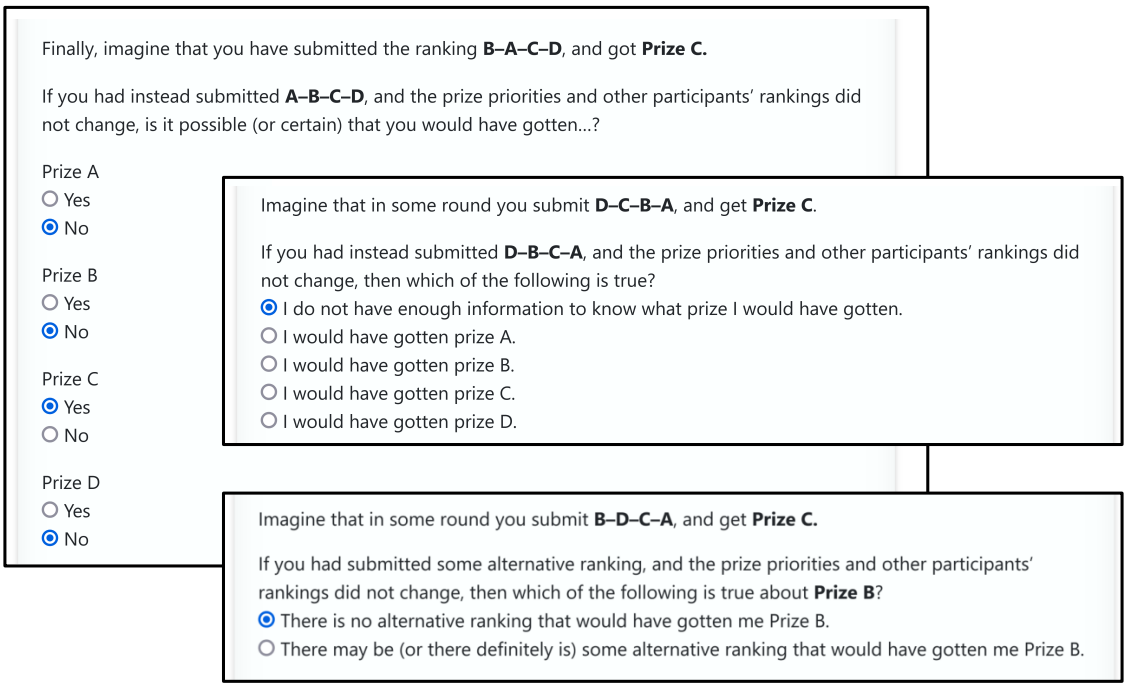
Figure 5: A real round



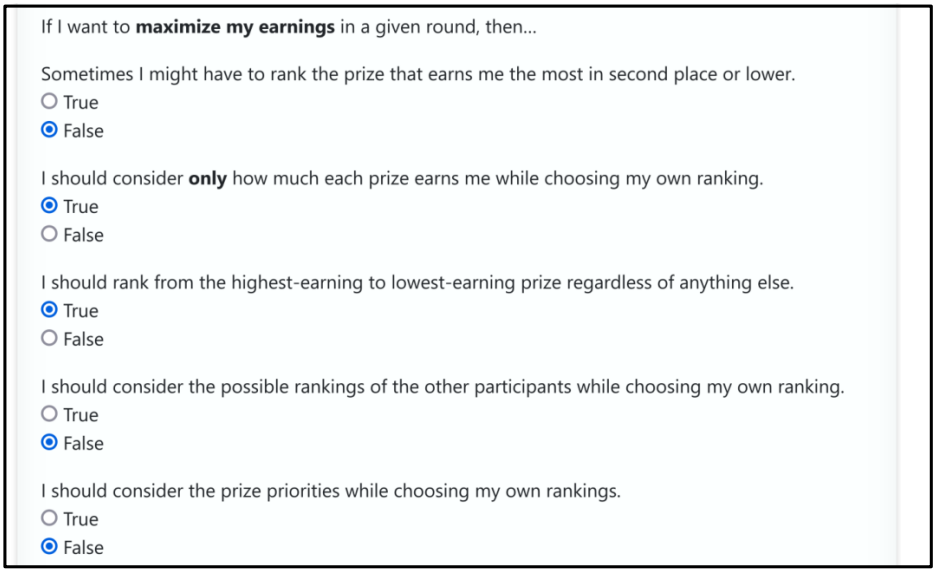
Notes: The participant entered the ranking C-A-B-D and received Prize A.

Figure 6: Samples of the strategyproofness understanding test

(a) Abstract



(b) Practical



Notes: The content of each bordered box appears on a separate screen. *Panel (a)*: questions directly concerning the abstract logical properties of strategyproofness ("Abstract"). The first screen contains another similar question, and an attention-check question. The second and third screens also contain, respectively, one and two other similar questions. *Panel (b)*: questions on practical implications, i.e., on how participants can maximize earnings ("Practical"). This panel shows the entire set of questions of that screen.

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 first screen also contains an attention-check question.

The final screen of the strategyproofness understanding test contains questions on practical implications of strategyproofness, and specifically, how one could maximize their earnings. We refer to these five questions jointly as Practical; see Figure 6(b) for a complete screenshot of this screen.

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 & Skowronek, 2018) and some are new to this study. The cognitive-abilities questionnaire uses the three questions of the Cognitive Reflection Task (CRT; Frederick, 2005), and measures numeracy using one question from the Berlin Numeracy Test (Cokely et al., 2012). We use this questionnaire to give participants a cognitive score between 0 and 4. We also similarly define an attention score outcome based on two attention-check questions planted in the Null training and SP-understanding test, discussed above. For a full list of the demographics and social, economic, and political leanings that we collect, see Appendix B.2 in the Supplementary Materials. The final questionnaire elicits general feedback on the experiment.

3 Results

3.1 Sample

Our data includes a total of 542 participants, collected from two sources. 255 participants were recruited on Prolific—a crowd-sourcing platform designed for scientific use—from August 3 to August 8, 2023. 287 participants were recruited at the Cornell Johnson Business

Simulation Lab (BSL) from September 26 to November 17, 2023.[11] We use these two different sources to increase 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. See Appendix A.1 for information on recruitment (which was as similar as possible across samples), earnings (which are on average £15.0 and $25.5 on Prolific and at Cornell, respectively), attrition levels (which do not vary by treatment), and duration.[12] See Appendix B.2 in the Supplementary Materials for demographic characteristics of our participants.

Our main results are qualitatively similar across Prolific and Cornell; hence, we pool the two samples for our main analysis (Sections 3.2 through 3.4). We report the cross-sample differences in Section 3.5. The two main differences are in baseline levels of all our main outcome variables, and in the magnitude of treatment effects; both are lower on Prolific.

3.2 Main Result #1: We Can Teach DA

We begin by studying participants' training score (% TR), elicited through our incentivized training rounds (which, in DA Mechanics treatments, use our training GUI). Our first main result, and our main result regarding % TR, is that the Traditional DA Mechanics description is effective at teaching participants the mechanics of DA. We begin by showing this result, and then investigate participants' training performance in other treatments.

Training score in Trad-DA. Figure 7 shows participants' average training-score performance. Figure 7(a) shows average overall % TR. In Trad-DA, mean % TR is 87% (SE = 1%), i.e., overall very high performance in the 37 questions and problems that participants face.

The training modules are specific to each treatment, and in particular differ dramatically between DA Mechanics treatments on the one hand and SP Property treatments on the other hand. Hence, in Figure 7, % TR is only meaningfully comparable within treatment pairs (separated by the dotted lines), but it differs somewhat even within pairs.[13] The panel,

[11] 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. See https://aspredicted.org/7eq7e.pdf for our pre-registration, and see Appendix B.1 in the Supplementary Materials for a summary and comparison with our analysis and findings.

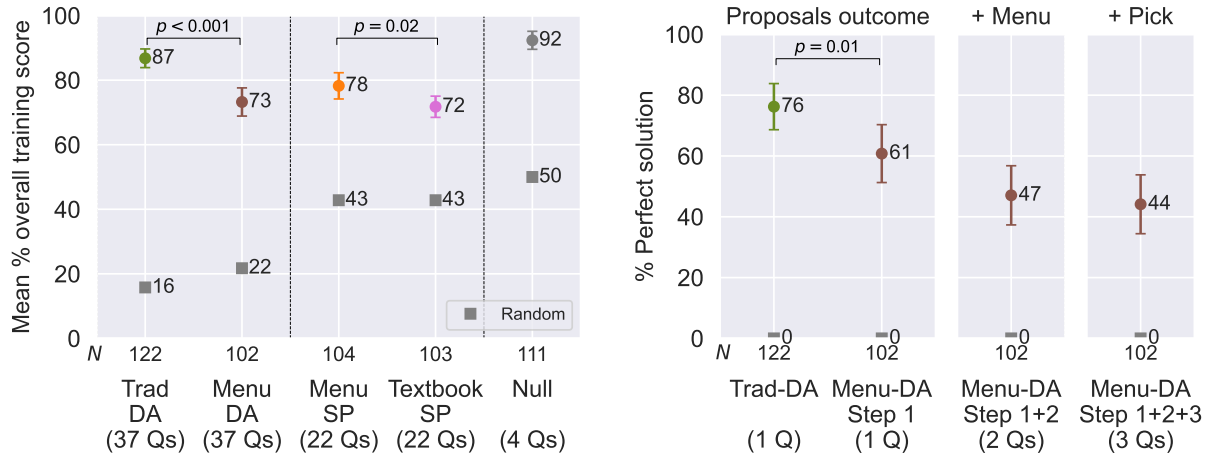
[12] The overall median completion time is 4 minutes higher on Prolific than at Cornell. In addition, median completion times differ across treatments: Trad-DA and Menu-DA take 48 minutes and 53 minutes, respectively, Menu-SP and Textbook-SP take 40 minutes and 39 minutes, respectively, and Null takes 35 minutes.

[13] 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

Figure 7: Training score (% TR) by treatment

(a) Mean overall score

(b) Perfect-solution rates in calculating DA outcome independently



Notes: *Panel (a)*: Mean overall % TR by treatment. *Panel (b)* (relevant to DA Mechanics treatments only): Rates of perfectly calculating (parts of) a DA outcome using our training GUI in the final training round, on the first attempt. "Proposals outcome": the rate of participants perfectly indicating the DA algorithm proposal-sequence outcome in our GUI. In Menu-DA, this is only the first step; "+ Menu": the rate of participants who also perfectly calculate their menu based on it; "+ Pick": the rate of participants who also perfectly indicate that their highest ranked prize is the one picked from the menu. *Both Panels:* Error bars: 95% confidence intervals. p-values: two-sample t-tests. "N": number of observations per treatment. "(# Qs)": the number of questions that the relevant score is based on. "Random": the expected score of answering every question uniformly at random. Averages are not comparable across vertical dotted lines.

along with others like it throughout this section, includes a "Random" benchmark that indicates what results would have looked like had participants answered questions uniformly at random. Participants' scores vary somewhat, but they are much higher than the Random benchmark in all treatments. For the full training-score distributions by treatment, see Appendix A.2, and for more details on question-level performance, see Appendix B.3 in the Supplementary Materials.

Figure 7(b) shows DA Mechanics participants' performance when focusing only on the final, most-challenging, and most externally valid training problem, which asks participants to correctly calculate a DA allocation outcome completely by themselves. Participants use our interactive GUI to perform the entire proposal sequence, which consists of 10 total proposals when executed correctly. In Trad-DA, 76% (SE = 4%) of participants succeed in solving this problem perfectly and on their first attempt (shown in the mini-figure labeled "Proposals outcome"). This fraction increases to 84% (SE = 3%) when including second-attempt successes.

much closer across the Menu-SP and Textbook-SP treatments.

These findings suggest that with sufficient care and coaching, many participants can be taught (the mechanical step-by-step method of calculating) DA using the traditional description of DA. At the same time, given that a quarter of our participants still do not accurately calculate DA on first attempt, our findings cast doubt on the assumption—often implicit in existing work—that most participants perfectly understand DA.

Training score in Menu-DA. Figure 7 also shows that training scores are lower in Menu-DA compared to Trad-DA, suggesting that of the two, Trad-DA is the simpler description. Figure 7(a) shows that % TR in Menu-DA is 73% (SE = 2%), compared to 87% in Trad-DA. Figure 7(b) then focuses on the hardest training problem, which has three steps in Menu-DA. The outcome variables indicating whether participants perfectly solve these steps together on the first attempt are labeled "Proposals outcome" (Step 1), "+ Menu" (both Steps 1 and 2), and "+ Pick" (Steps 1, 2, and 3). Participants in Menu-DA find (all steps of) the training more challenging than in Trad-DA. 44% (SE = 5%) of participants are able to perfectly complete the entire three-step process on their first attempt (which increases to 56% when including participants who succeed on the first or second attempt in all steps of the process; SE = 5%).

Training score in SP Property treatments and Null. Finally, Figure 7(a) also shows that training scores in the two SP Property treatments are fairly high, and are more comparable in mean levels compared to the wider gap between Trad-DA and Menu-DA. Null training scores are near perfect, suggesting that repeating the understanding questions on the basic components of our setting is indeed simple and straightforward for most participants.

3.3 Main Results #2 and #3: Understanding DA $\neq$ Understanding SP, But We Can Teach SP

Perhaps our most important and novel outcome variable is % SP-U—participants' score on the strategyproofness-understanding test. This incentivized eighteen-question test, identical across treatments, is conducted after training and playing ten real rounds of DA, and underlies our second and third main results.

Our second main result is that understanding the mechanics of DA does not imply understanding of its SP property, as indicated by low % SP-U levels in Trad-DA (comparable to those in the Null treatment). Our third main result is that SP Property descriptions, and particularly Menu-SP, are more successful in teaching SP to participants. This subsection presents these results, as well as a more detailed investigation that separates the full % SP-U

measure into its two main components: Abstract and Practical.

3.3.1 Overall Measure of Strategyproofness Understanding

Figure 8(a) shows average % SP-U—the percent of the eighteen questions that a participant answered correctly—across treatments. The figure first shows that both DA Mechanics treatments are indistinguishable from each other in % SP-U, and are essentially indistinguishable from the baseline of 54% (SE = 1%) Null treatment. Trad-DA's high training scores but low strategyproofness-understanding scores underlies our second main result. In particular, in spite of three quarters of Trad-DA participants learning to fully calculate the final DA allocation on their own with no mistakes (see Section 3.2), they do not understand that the mechanism is strategyproof more than Null participants, who know essentially nothing about the mechanism!

In contrast, Menu-SP, and to a much lesser extent Textbook-SP, lead to a noticeable increase in % SP-U, with mean levels 71% and 62% (SE = 2% in both treatments), respectively—underlying our third main result.[14] We speculate that the difference in % SP-U between Menu-DA and Menu-SP—where the former is merely a more detailed version of the latter—suggests that the additional complex details in Menu-DA make the important features of menu descriptions insufficiently salient.

Figure 8(b) presents these results in more detail, reporting the distribution of % SP-U across treatments. A noticeably larger fraction of participants get perfect or near-perfect SP-U scores in the Menu-SP treatment, with much smaller differences across other treatments (including Null). For example, 28% (SE = 4%) of Menu-SP participants get at least seventeen of the eighteen questions correct. This figure drops to 12% (SE = 3%) in Menu-DA, 6% (SE = 2%) in Textbook-SP, 5% (SE = 2%) in Trad-DA, and 2% (SE = 1%) in Null.

3.3.2 Sub-Measures of Strategyproofness Understanding

Figure 9 shows participants' performance separately on the Abstract (consisting of 13 questions) and Practical (5 questions) parts of the SP-U test.[15] Figures 9(a) and 9(b) support our second main result—by showing that Trad-DA participants score low (statistically, as low as Null) on both sub-measures—and our third main result—by showing that Menu-SP

[14] There are ten possible comparisons of % SP-U across pairs of treatments. Applying an arguably over-conservative Bonferroni correction to address multiple comparisons would multiply all p-values in Figure 8(a) by 10, maintaining the difference between Menu-SP and Textbook-SP with $p < 0.01$, and the difference between Menu-SP and any other treatment with $p < 0.001$.

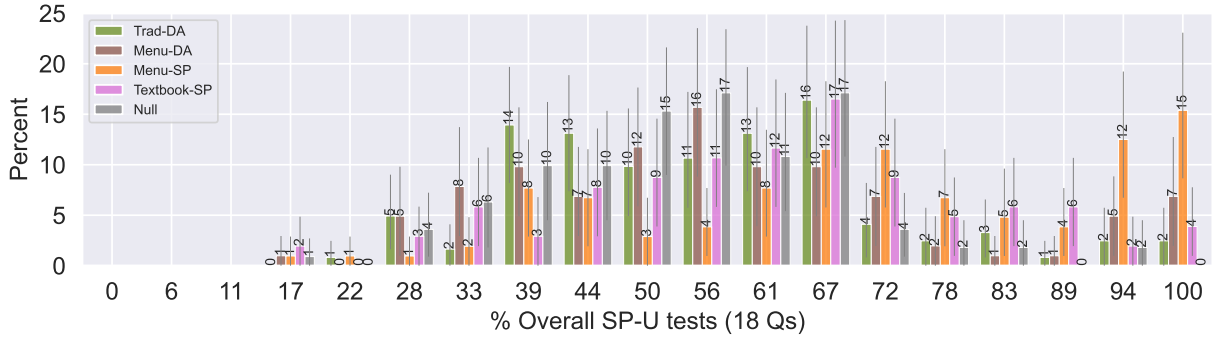
[15] See Appendix A.3 for participants' performance at the question level, along with a summary of each question.

Figure 8: Strategyproofness understanding test score (% SP-U) by treatment

(a) Mean overall score



(b) Distribution of overall score



Notes: *Panel (a):* Mean of % SP-U by treatment. p-values: two-sample t-tests. "N": number of observations per treatment. "Random": the expected score of answering every question uniformly at random. *Panel (b):* Distribution of % SP-U by treatment. *Both Panels:* Error bars: 95% confidence intervals. "(# Qs)": the number of questions that the test consists of.

participants score high compared to all other treatments on both sub-measures. Remarkably, Figure 9(b) also shows that with the exception of Menu-SP, participants' Practical SP-U scores in all other treatments are significantly worse than Random. We return to this point momentarily.

Figure 9(c) shows the Abstract SP-U score distributions within each treatment. They seem very roughly bell-shaped, with a mode that moves right across treatments, from Trad-DA and Null at 54% correct (i.e., seven out of thirteen questions) to Menu-SP at 100% correct (with 27% of participants getting this score, and another 26% getting the next possible highest score).

In contrast, Figure 9(d), which shows the Practical SP-U score distributions, suggests that they are bimodal, with many participants scoring close to zero while a group of others get a perfect 100% score (i.e., five out of five questions). This bimodality 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).

Additionally, Figure 9(d) shows 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 score zero—i.e., answer according to an "opposite-SP" perception of DA—a rate similar to that among Null participants, who are told almost nothing about the mechanism. In contrast, in Menu-SP, many fewer participants have this "opposite-SP" perception while many more participants compared to other treatments correctly perceive strategyproofness.

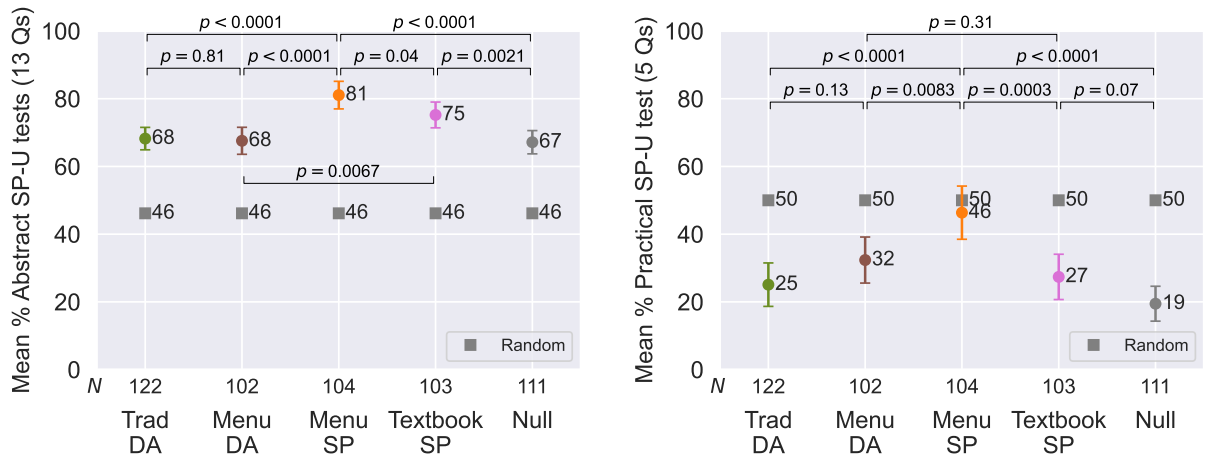
In Appendix A.3, we further investigate the joint distribution of Abstract and Practical, and how the two sum up to % SP-U. Our main finding is that the variation in % SP-U is largely driven by the different sub-measures for different ranges of the % SP-U distribution. Among observations with % SP-U $< 75\%$, Practical is typically low—approximately at its "opposite-SP" perception mode (with an average of 17%; SE $< 1\%$)—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% in Figure 8(b). 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 85% (SE $< 1\%$).[16] This pattern suggests a potentially more general feature of participants' understanding of SP—that reaching an intermediate level of SP understanding does not mean

[16]The specific value of (approximately) 75% seems largely due to the relative weight of the sub-measures in the overall measure—13/18 questions for Abstract, and 5/18 for Practical.

Figure 9: Abstract and Practical sub-measures of strategyproofness understanding test score (% SP-U) by treatment

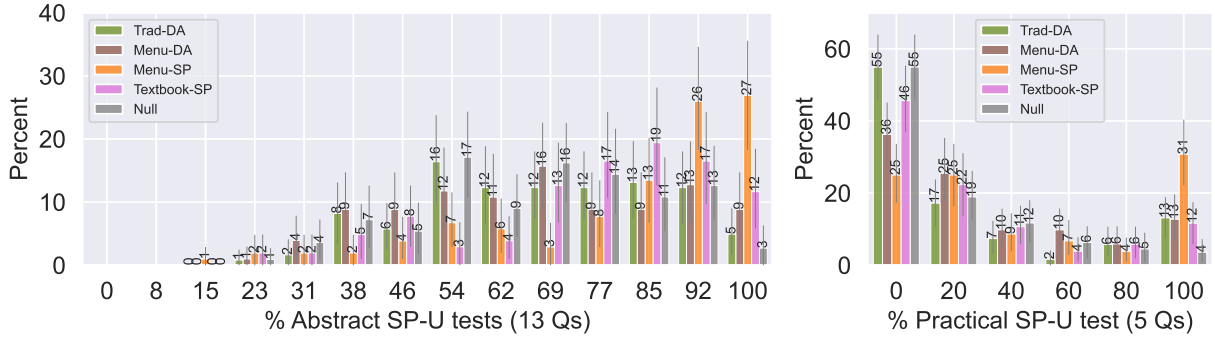
(a) Mean Abstract sub-measure score

(b) Mean Practical sub-measure score



(c) Distribution of Abstract sub-measure score

(d) Distribution of Practical sub-measure score



Notes: *Panel (a):* Mean of Abstract sub-measure of % SP-U, i.e., the first three screens of the test (13 questions), which ask about abstract logical properties of strategyproofness. *Panel (b):* Mean of Practical sub-measure of % SP-U, i.e., the last screen of the test (5 questions), which ask about practical implications of strategyproofness (specifically, how participants could maximize earnings). p-values: two-sample t-tests. "N": number of observations per treatment. "Random": the expected score of answering every question uniformly at random. *All Panels:* Error bars: 95% confidence intervals. "(# Qs)": the number of questions that the relevant part of the strategyproofness understanding test consists of.

"partially" understanding how to maximize earnings; rather, (some) participants realize how to maximize their earnings only when they understand SP quite well.

Overall, the joint distribution of Abstract and Practical suggests that in our setting, in order for participants to understand the practical implications of SP on earnings maximization, they need to understand its abstract logical properties. However, understanding the abstract logical properties is far from implying understanding of the practical implications. Even in Menu-SP, where many participants score perfectly or nearly perfectly on Abstract, a majority of participants are still closer to perfect opposite-SP than to perfect SP perception in Practical.

3.4 Main Result #4: Strategyproofness Understanding Predicts Straightforward Behavior

Our final main outcome measure is the rate of straightforward behavior (% SF) in standard incentivized DA rounds. We first find that our treatments' effects on % SF are overall small. However, when investigating the relation between % SP-U and % SF, we find our fourth main result: Participants with high-enough SP-understanding play SF at very high rates.

3.4.1 Mean Levels of Straightforward Behavior

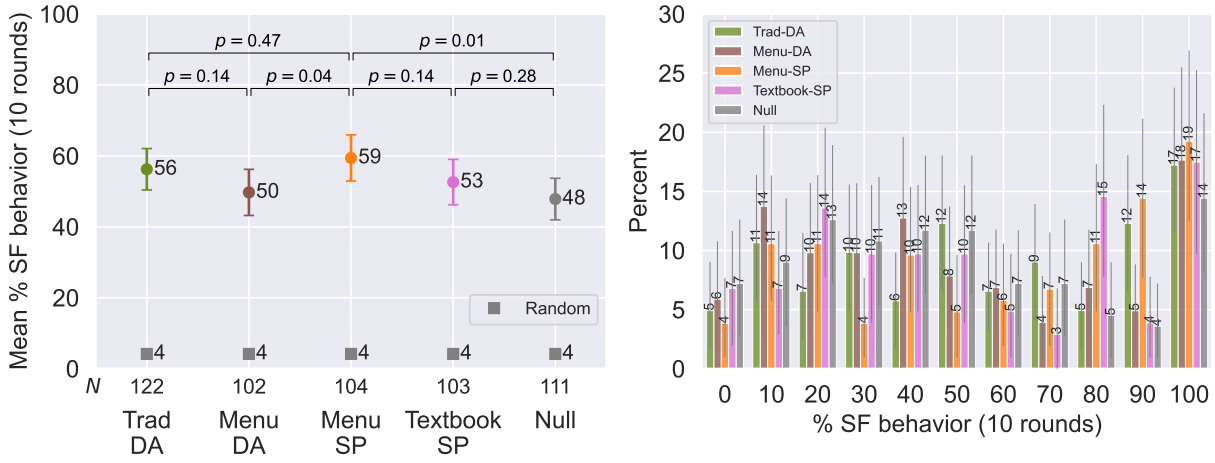
Figure 10 shows means and distributions of % SF across treatments. % SF is never particularly high on average. Its baseline rate is 48% in Null. In Menu-SP, in spite of the significant increase in % SP-U we observe in Section 3.3, % SF is still only 59%, and only slightly above Trad-DA's 56%. Menu-DA and Textbook-SP remain even lower, at 50% and 53% respectively—statistically indistinguishable from Null. (SE = 3% for % SF in all treatments.) Turning to distributions, % SF seems widely spread in all treatments, and again (at least somewhat) bimodal, with a plateau around 20-40%, and a peak at 100%.

In Appendix A.4, we look for patterns of % SF over the 10 played rounds, as well as patterns in participants' ranking behavior beyond % SF, i.e., beyond the binary classification of SF vs. NSF play. By far the most common NSF ranking flips 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 seems to depend only weakly on parameters of the specific round of DA.[17] We do not find strong trends of % SF over the 10 rounds in most

[17]For example, we find that across all treatments, participants, and rounds, the fraction of times a participant *does not* rank 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 this comparison across treatments. We also find little variation in ranking patterns conditional

Figure 10: Straightforward ranking behavior (% SF) by treatment

(a) Mean % SF **(b)** Distribution of % SF



Notes: *Panel (a)*: p-values: two-sample t-tests. "N": number of observations per treatment. "Random": the expected % SF from ranking the prizes uniformly at random. *Both Panels*: Error bars: 95% confidence intervals.

treatments. The exception is Menu-SP, where there is an average increase of 1.6% (SE = 0.5%) per round in % SF, suggesting potential effects of learning, experience, fatigue, and other dynamic phenomena.

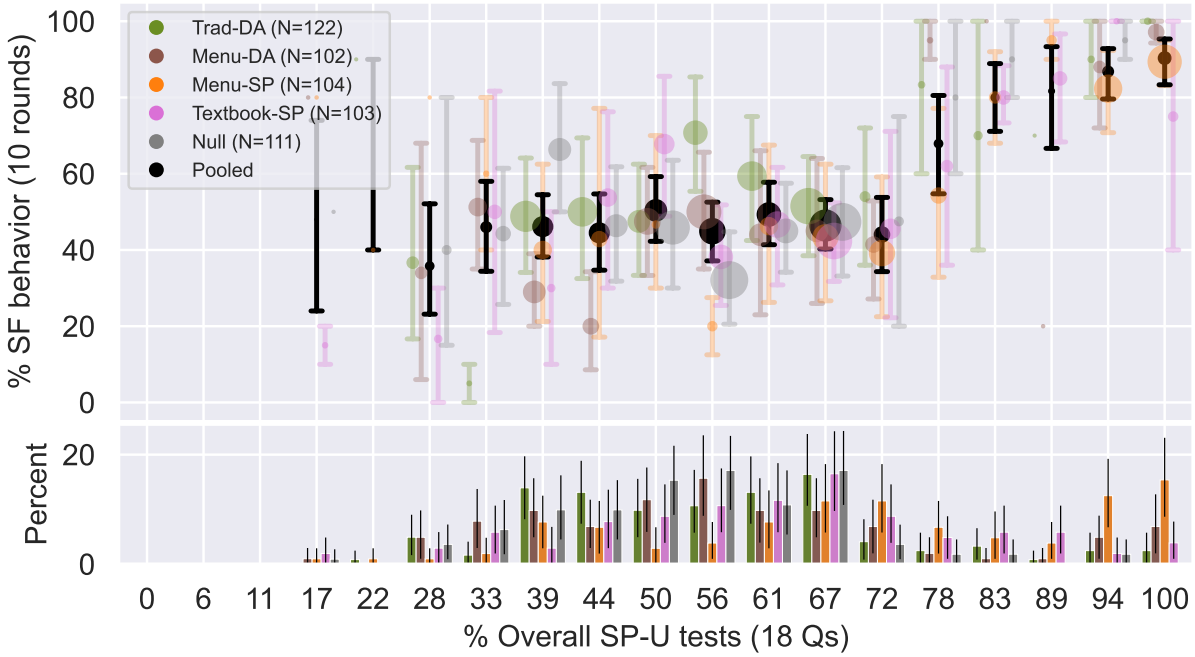
3.4.2 Relation Between Strategyproofness Understanding and Straightforward Behavior

Finally, we investigate the joint distribution of % SP-U and % SF. While not part of our pre-registered analysis plan, we find strong results, robust across treatments and sub-samples. Figure 11 shows mean % SF conditioned on (i.e., binned by) % SP-U. At its bottom, the figure reports the distribution of % SP-U scores by treatment (replicating Figure 8(b)). At its top, it reports mean % SF among participants with given % SP-U scores, both by treatment (in semitransparent colors) and pooled (black). The diameter of circle markers is proportional to bin size (in the histogram at the bottom).

Figure 11 shows our fourth main result: That those who grasp strategyproofness and its implications sufficiently well play SF significantly more than those who do not. Indeed, we see that mean % SF "jumps" to noticeably high levels once participants score sufficiently high on % SP-U, with a threshold at roughly 75–80%. While more noisy at the single-treatment level, this relation seems remarkably global. That it is not treatment-specific is consistent with the notion that better understanding of SP may increase SF behavior of

on the difference in monetary value between the highest-earning prize and the second-highest-earning prize.

Figure 11: Relationship between % SP-U and % SF by treatment



Notes: *Top:* Mean % SF by treatment and by % SP-U score. *Bottom:* Distribution of % SP-U by treatment (replicating Figure 8(b)). *Both top and bottom*: Circular markers (top) are proportional in diameter to corresponding fraction of participants (bottom). Error bars: 95% confidence intervals.

some participants, regardless of the description or training that improved SP understanding. Moreover, 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 high levels of both % SP-U and % SF.

This strong relation between % SP-U and % SF, and the strong variation in % SP-U across treatments from Section 3.3, may initially appear somewhat at odds with our findings of small overall differences in mean % SF across treatments. Figure 11 suggests at least a partial explanation. First, it shows that Pooled across all treatments, the mean % SF among participants with % SP-U less than 75% is 46%, while for participants above 75% it is 83%, a difference of 36% percentage points. Second, it also shows that the fraction of participants with % SP-U above 75% ranges from 43% in Menu-SP to 5% in Null, a difference of 38% percentage points. Thus, one would expect a ballpark difference in % SF of at most $38\% \times 36\% = 14\%$ across treatments. This is consistent with the (noisy) range of mean % SF values observed across treatments in Figure 10(a). As the p-values reported in that figure suggest, our sample size is slightly under-powered to detect most differences within this range.

In Appendix A.5, we investigate the non-conditional joint distribution of % SP-U and %

SF. We observe that Menu-SP is most effective at moving participants into a region where they both understand SP well and play SF at high rates. For instance, 32% (SE = 5%) of participants in Menu-SP have both high % SP-U and high % SF, namely, above 75% in both metrics. In all other treatments this fraction is 17% (SE = 4%) or less. This suggests that participants in Menu-SP tend to not only move towards the higher mode of understanding of strategyproofness, but may additionally act on this understanding by moving towards the higher mode of % SF.

In Appendix A.5, we additionally investigate how the Abstract and Practical sub-measures of % SP-U differently contribute to the overall relation between % SP-U and % SF. 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 vs. low SF play based on high vs. low sub-measure score. However, we find that including Abstract may increase the sharpness of this separation, especially among participants who score in the middle range of Practical.[18]

3.5 Additional Results and Robustness

In this subsection, we investigate differences in our main results across the Prolific and Cornell samples, test robustness to controlling for a set of additional variables, and summarize additional findings on the relationship between our three outcome variables.

3.5.1 Variation Across Samples

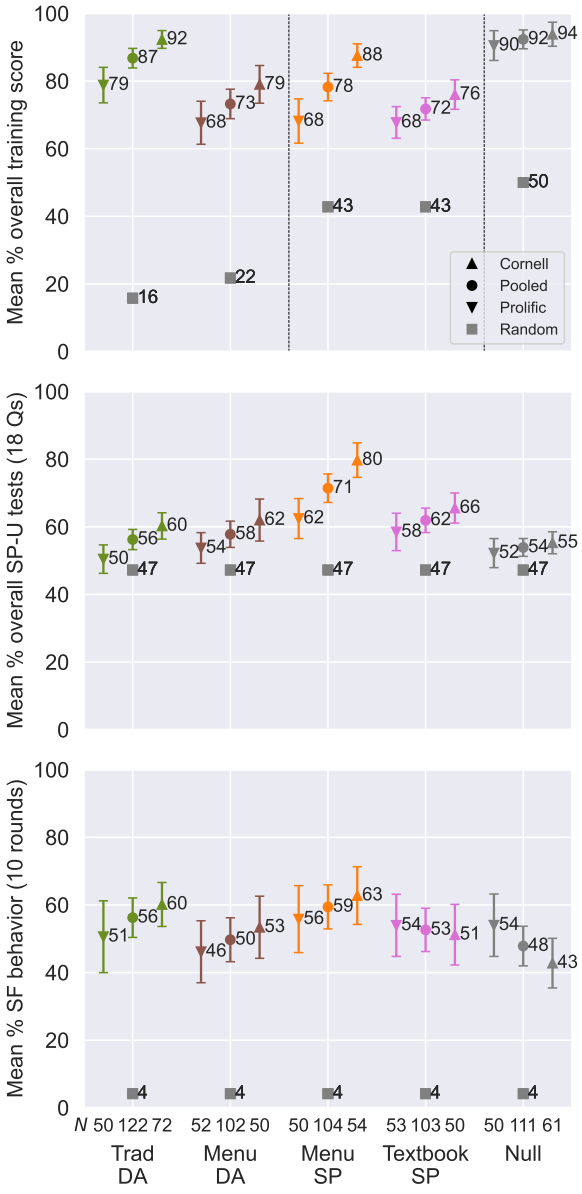
Figure 12 reproduces the main figures of this section—Figures 7(a), 8(a), 10(a), and 11—split by sample (Cornell vs. Prolific). As seen most clearly in Figure 12(a), the baseline levels of outcome variables in non-Null treatments are lower on Prolific (▼) than at Cornell (▲), particularly % TR (Figure 12(a)'s top mini-figure) and % SP-U (middle mini-figure). And, as seen in both panels, Menu-SP's effect of moving participants to both high % SP-U and high % SF is less pronounced (in accordance with fewer participants achieving high % SP-U in the Prolific sample).

We draw three conclusions. First, these results support the idea that our novel metrics, % TR and % SP-U, are meaningfully measuring understanding levels. In non-Null treatments, those levels should in principle increase with education, cognitive ability, attention, and ability to focus in a physical lab (vs. remote participation); we indeed find % TR and

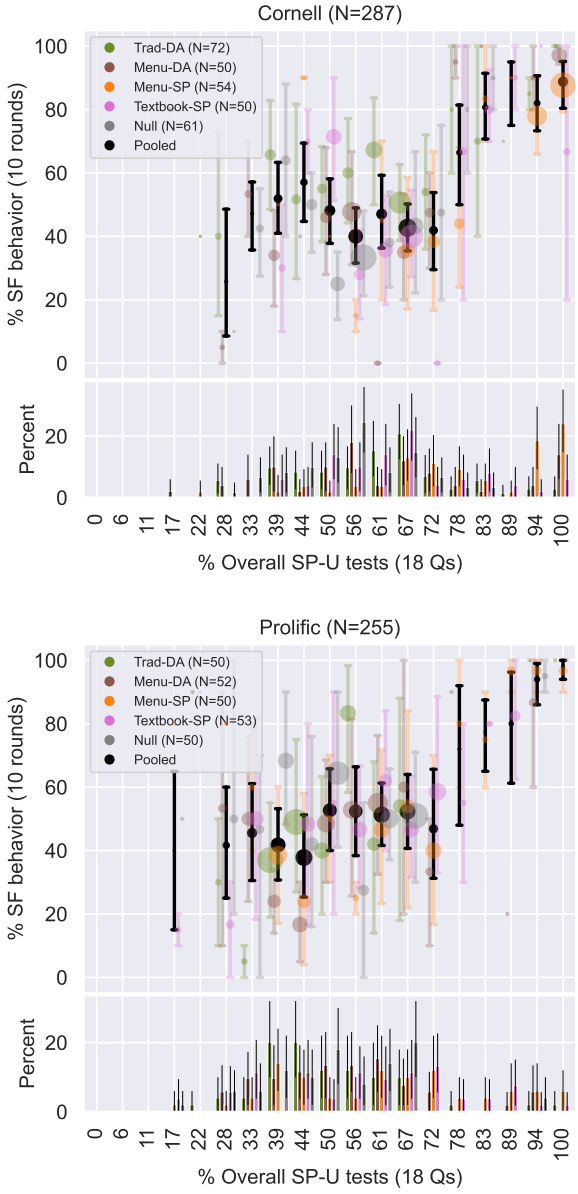
[18]For example, among participants whose Practical scores are 40% or 60% (two or three correct answers out of five), the average SF rate moves from 49% (SE = 4%) to 80% (SE = 7%) when Abstract changes from below 90% to above 90%.

Figure 12: Main results within the Prolific and Cornell samples

(a) Mean outcome variables by sample



(b) Relation between % SF and % SP-U by sample



Notes: *Panel (a)*: The three mini-figures report the mean rates of % TR (top), % SP-U (middle), and % SF (bottom), by treatment. Each mini-figure reports the overall mean and the means among Cornell participants ($N = 287$, ▲) and among Prolific participants ($N = 255$, ▼; see Section 3.1). "N": number of observations per treatment. "Random": the expected score of answering every question uniformly at random. *Panel (b)*: Figure 11 by sample (top: Cornell; bottom: Prolific).

% SP-U to increase within each non-Null treatment for Cornell vs. Prolific. Our new metrics thus pass an important sanity check.

Second, these results suggest that despite the sample differences, cross-treatment differences are at least qualitatively similar across the two samples. This conclusion applies to both the average cross-treatment differences in mean outcome variables (Figure 12(a)) and the sharp change in % SF when % SP-U increases from low to high levels (Figure 12(b)).

Third, importantly, these cross-sample differences highlight a potential limitation of Menu-SP: Its effect in improving SP understanding appears stronger among those of our participants who are more educated, more highly cognitively skilled, more attentive, or participate onsite (vs. remotely). Each of these factors may imply that real-world implementations of Menu-SP may have significant distributional consequences, to be investigated in future work.

In Appendix A.6, we conduct a similar comparison of participants with high vs. low cognitive scores, and of high vs. low attention scores, and see similar results to the Cornell vs. Prolific comparison (for more details on the cognitive and attention scores, see Appendix B.4 in the Supplementary Materials).

3.5.2 Robustness to Controls

In Appendix A.7, we use the demographic control variables (elicited at the end of the experiment) to conduct two additional robustness tests to our main results. First, we verify that the mean cross-treatment differences of % TR, % SP-U, and % SF (shown in Figures 7, 8, and 10) change little quantitatively when controlling for demographic characteristics, cognitive and attention scores, and session-date fixed effects. Second, we find that the relation of % SP-U and % SF (shown in Figure 11) maintains its step-function shape when controlling for the above variables, across treatments.

3.5.3 Other Relations Between Outcome Variables

Section 3.4.2 focused on the relation between % SP-U and % SF. Among the three pairwise relations between our three main outcome variables, this is the only one that significantly contributes to our understanding of our data beyond that conveyed by mean levels.

For completeness, in Appendix B.5 in the Supplementary Materials, we investigate the other joint relationships between our main outcome variables. We make three main observations. First, average relations between % TR, % SP-U and % SF using linear regressions and correlation coefficients give further evidence that the relation between % SP-U and % SF is fairly consistent across treatments. Second, we investigate the relation between % TR and

% SP-U in more detail. In both DA Mechanics treatments, we find within-treatment support for our second main result, that teaching participants (the mechanics of) DA does not imply teaching them strategyproofness, even for those participants who perform very well in % TR. In the SP Property treatments, we find a strong monotone relationship between % TR and % SP-U, suggesting that the SP Property treatments and the standardized SP-U tests are well-aligned in the way they measure strategyproofness understanding. Finally, we investigate the relation between % TR and % SF, which looks like a noisy combination of the two other relations. This is consistent with the idea that % TR is mostly related to % SF through its relation with % SP-U.

4 Related Literature

Our paper sits most directly within the literature on behavioral mechanism and market design. Excellent reviews can be found in Hakimov and Kübler (2021) and Rees-Jones and Shorrer (2023). Early works within this literature compare behavior across different social choice rules, e.g., Kagel and Levin (1993), Y. Chen and Sönmez (2006), Pais and Pintér (2008), and Featherstone and Niederle (2016), among others. More recent papers (surveyed below, e.g., Li, 2017) often consider a fixed social choice rule, and vary some feature of how the rule is implemented. Our paper considers a fixed implementation—i.e., we keep the rounds of play identical across treatments—and we vary the framing and information provided to participants. We now discuss, and compare our paper with, the most directly related prior work and themes in the literature.

Menu descriptions. As discussed, our Menu DA Mechanics treatment is directly based on the mechanism design theory paper GHT (2023), which also indirectly inspired our Menu SP Property treatment.

Other than GHT, the most closely related paper to ours is the experimental paper by Katuščák and Kittsteiner (2024). Like our study, they conduct an experiment with a menu description of a matching mechanism, Top Trading Cycles (henceforth TTC).[19] Beyond differing in the mechanism used, their experiment has only behavior as an outcome variable, while our focus is on the interplay between training score, SP understanding, and behavior. Moreover, they do not conduct training, and do not test SP understanding separately from behavior. They find higher rates of SF play in their menu treatment of TTC, especially for

[19]Phrased in the language of our paper, one treatment of Katuščák and Kittsteiner (2024) 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.

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.[20] Our training questions and allocation-calculation GUI extend these papers' descriptions and their training tasks, and formalizes them into a computerized interface with hand-held coaching and feedback. Recently, Serizawa et al. (2024) conduct 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. Prior work that explores strategic advice in matching mechanisms includes Guillen and Hing (2014) and 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). These papers generally find that advice is effective at increasing SF behavior, but does not completely eliminate NSF behavior. Our SP Property treatments describe the strategyproofness property, and avoid providing any explicit strategic advice to choose the SF ranking.

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

[20] These exceptions include Guillen and Hing (2014), Bó and Hakimov (2020), and Guillen and Veszteg (2021). Each of them reports 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 (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 submitted ranking is considered by the algorithm.

participants submit (reversed) straightforward rankings. Danz et al. (2022) and Guillen and Hing (2014) show that, in different strategyproof environments, providing participants with more information on the environment's mechanics can *decrease* straightforward reporting. Related to such results, we find that Menu-DA is less effective at teaching SP relative to Menu-SP—despite merely providing more information.

Behavioral mechanism design more broadly. Other papers propose behavioral models of behavior or confusion in strategyproof mechanisms, and investigate these models empirically or theoretically. Li (2017) theoretically and experimentally studies failures to grasp that a mechanism is strategyproof due to limited contingent-reasoning skills, and defines and proposes switching to obviously strategyproof (OSP) implementations of strategyproof choice rules. Ashlagi and Gonczarowski (2018) and Thomas (2021) show that, except under highly restrictive sets of priorities, DA does not have an OSP implementation.[21] Pycia and Troyan (2023) theoretically study a hierarchy of strategic simplicity notions strengthening OSP.

Dreyfuss, Glicksohn, et al. (2022), Dreyfuss, Heffetz, and Rabin (2022), and Meisner and von Wangenheim (2023) study loss aversion in matching mechanisms and DA in particular, and find that it can explain NSF behavior patterns well. 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 our work focuses on participants' misunderstandings of strategyproofness, we do not provide explicit information on probabilities of getting different prizes, 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 from some set of still-available options, and find that such mechanism can increase SF play.

Our work is also indirectly inspired by a vast literature of empirical papers that study real-world implementations of matching mechanisms. See Pathak (2017) for a review.

5 Conclusion

Describing designed markets, such as those using DA, to participants is a challenging task. One reason is that calculating matches in DA and in various other popular mechanisms requires a detailed combinatorial algorithm. We view the encouraging training scores in all our treatments as a successful indication that both the mechanics of DA, as well as the strate-

[21] While classifying OSP matching 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.

gyproofness property, can be taught to participants with careful instruction. Building on the tools we have developed, such as our experimental flow, training GUI, and strategyproofness understanding tests, future work may further refine methods designed to help participants better understand 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, Glicksohn, et al., 2022; Dreyfuss, Heffetz, & Rabin, 2022; Meisner & 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: Above a certain sufficiently high level of strategyproofness understanding, participants indeed play SF at rather high rates.

Our main results suggest that the traditional description of DA can be taught to participants without teaching them strategyproofness. At the same time, participants *can* (to some extent) be taught strategyproofness via our novel Menu-SP description. While it was our Menu-*DA* description—which we based on the main theoretical result in (Gonczarowski et al., 2023)—that originally inspired us to embark on our experimental investigation into how participants respond to different descriptions of DA, our results suggest that mechanical descriptions may remain too complicated to effectively convey strategyproofness, even in their menu form. In contrast, our Menu-SP treatment—a stripped-down version of the full Menu-DA that focuses on the main principle of menu descriptions—is found to be much more effective in our data. Its simplicity may be the key to success. It simply informs participants of a concrete property satisfied by the matching algorithm: Each participant's allocation will always be their highest-ranked choice among some set of possibilities that their ranking cannot influence.

Our study suggests many promising directions. Future work may develop SP Property descriptions further, explore the interaction between conveying SP Property and providing concrete ranking advice, investigate real-world deployments, and, importantly, pay careful attention to potential redistributional consequences. Future work may also study broadly different new approaches to conveying other important properties of mechanisms, such as those related to different notions of fairness.

References

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

Bó, I., & Hakimov, R. (2020). Iterative Versus Standard Deferred Acceptance: Experimental Evidence. *The Economic Journal*, *130*(626), 356–392.

Bó, I., & Hakimov, R. (2023). Pick-an-object mechanisms. *Management Science*.

Breitmoser, Y., & Schweighofer-Kodritsch, S. (2022). Obviousness around the clock. *Experimental Economics*, *25*, 483–513.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*(100), 88–97.

Chen, Y., & Sönmez, T. (2006). School choice: An experimental study. *Journal of Economic Theory*, *127*(1), 202–231.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25–47.

Danz, D., Vesterlund, L., & Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, *112*(9), 2851–2883.

Dreyfuss, B., Glicksohn, O., Heffetz, O., & Romm, A. (2022). *Deferred acceptance with news utility* [Working paper].

Dreyfuss, B., Heffetz, O., & Rabin, M. (2022). Expectations-based loss aversion may help explain seemingly dominated choices in strategy-proof mechanisms. *Forthcoming in American Economic Journal: Microeconomics*.

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

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

Gale, D., & Shapley, L. S. (1962). College admissions and the stability of marriage. *American Mathematical Monthly*, *69*, 9–14.

Gonczarowski, Y. A., Heffetz, O., & Thomas, C. (2023). Strategyproofness-exposing mechanism descriptions [Extended abstract appeared in EC 2023.]. *arXiv preprint arXiv:2209.13148*.

Guillen, P., & Hakimov, R. (2018). The effectiveness of top-down advice in strategy-proof mechanisms: A field experiment. *European Economic Review*.

Guillen, P., & Hing, A. (2014). Lying through their teeth: Third party advice and truth telling in a strategy proof mechanism. *European Economic Review*.

Guillen, P., & Veszteg, R. F. (2021). Strategy-proofness in experimental matching markets. *Experimental Economics*, *24*, 650–668.

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

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

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

Katuščák, P., & Kittsteiner, T. (2024). *Strategy-proofness made simpler* [Mimeo].

Li, S. (2017). Obviously strategy-proof mechanisms. *American Economic Review*.

Masuda, T., Mikami, R., Sakai, T., Serizawa, S., & Wakayama, T. (2022). The net effect of advice on strategy-proof mechanisms: An experiment for the Vickrey auction. *Experimental Economics*, *25*, 902–951.

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

NMS. (2020). The matching algorithm - explained [Video produced by National Matching Services]. https://www.youtube.com/watch?v=kVTwXNawpbk

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

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

Pycia, M., & Troyan, P. (2023). A theory of simplicity in games and mechanism design. *Econometrica*.

Rees-Jones, A., & Shorrer, R. (2023). Behavioral economics in education market design: A forward-looking review. *Journal of Political Economy Microeconomics*, *1*(3), 557–613.

Rees-Jones, A., & Skowronek, S. (2018). An experimental investigation of preference misrepresentation in the residency match. *Proceedings of the National Academy of Sciences*.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. John Wiley & Sons.

Serizawa, S., Shimada, N., & Tse, T. T. K. (2024). *Toward an understanding of insincere bidding in a vickrey auction experiment* (ISER Discussion Paper). Institute of Social and Economic Research, Osaka University.

Thomas, C. (2021). Classification of priorities such that deferred acceptance is OSP implementable. *Proceedings of the 22nd ACM Conference on Economics and Computation*, 860.

A Additional Results

In this appendix, we present additional results and analyses of our data.

A.1 Sample Collection, Earnings, and Duration

We provide further details on our sample collection, including attrition rates by sample and treatment, and on earnings and completion times.

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. Specifically, they dropped out at or before an optional exit point, right after the common Null training, where they could quit for partial payment; we included this salient exit point to reduce attrition conditional on treatment. We had to drop another 2 observations that included corrupt DA round data due to technical issues, leaving a final sample of $N = 255$. At Cornell, 295 participants started the experiment and 293 completed it; of the two incompletes, one voluntarily dropped out, and one stopped due to a computer crash. We had to drop another 6 observations due to other technical data-saving errors unrelated to the experiment content, leaving a final sample of $N = 287$.

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, of the 2 participants who dropped out after allocation into treatment or whose computer crashed, 1 was allocated to Trad-DA and 1 to Textbook-SP. Treatment data was not saved for the additional 6 participants who had data-saving errors.

Recruitment and experiment texts were as similar as possible across platforms. Changes to experiment text were limited to differences in currency—British pounds (£) on Prolific and US dollars ($) at Cornell; 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. Prolific participants were required to participate in a camera-on video-conference meeting with the experimenter for the duration of the experiment. Cornell prescreening included only students. Cornell participants participated in physical lab sessions. For recruiting materials, a detailed description of the experiment protocol, and more details on the sessions, see Appendix C.4 in the Supplementary Materials. For demographics characteristics of our sample, see Appendix B.2 in the Supplementary Materials.

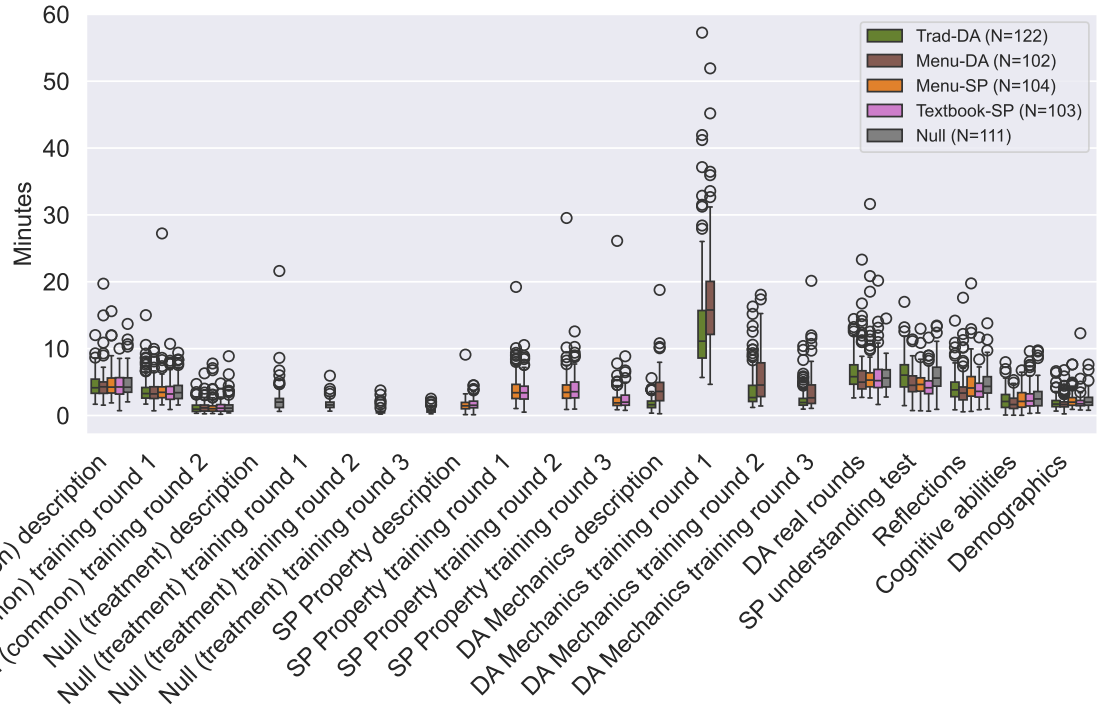
Prolific participants earned an average of £15.0 (a fixed £7 participation fee; an average

of £3.0 for TR and SP-U questions; and an average of £5.0 for incentivized rounds). Cornell participants earned 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; participants could choose to receive one credit for their studies at Cornell instead of the $17 participation fee; all monetary earnings were paid through Amazon gift cards).

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 on Prolific is 4 minutes higher than at Cornell.

Figure A.1 shows the distribution of experiment duration across treatments and the different main parts of the experiment.

Figure A.1: Experiment duration by treatment and part



A.2 Full Training Score Distribution

We present more details on participants' performance in our training modules, by showing the overall training-score distributions.

Figure A.2 shows the overall distribution of participants' performance in all training modules. From top to bottom, the figure rows focus on: (1) the two DA Mechanics treat-

ments, (2) the two SP Property treatments, (3) the Null treatment, and (4) participants' performance in the training round following the Null description, which is common to all participants. These are the only training questions in the experiment which are *not* a part of % TR measure.

For details on the mean question-level performance within each treatment, see Appendix B.3 in the Supplementary Materials.

A.3 Sub-Measures of % SP-U

We study the distributions of sub-measures of % SP-U, particularly the joint distribution of the Abstract and Practical sub-measures.

First, Figure A.3 shows the average success rates by treatment in *all* individual SP-U questions, along with summaries of the questions.

Next, in Figure A.4, we investigate the joint distribution between the main sub-measures of % SP-U, Abstract and Practical. The figure shows that high scores in Practical are predictive of high scores in Abstract; for example, among all participants with Practical $\geq$ 75%, the average Abstract score is 83%. In the other direction, the Abstract score is not as predictive regarding Practical; for example, even among participants with Abstract $\geq$ 90%, the average Practical score is 47%. This suggests that a high Abstract score is nearly a necessary condition—albeit an insufficient one—for a high Practical score.

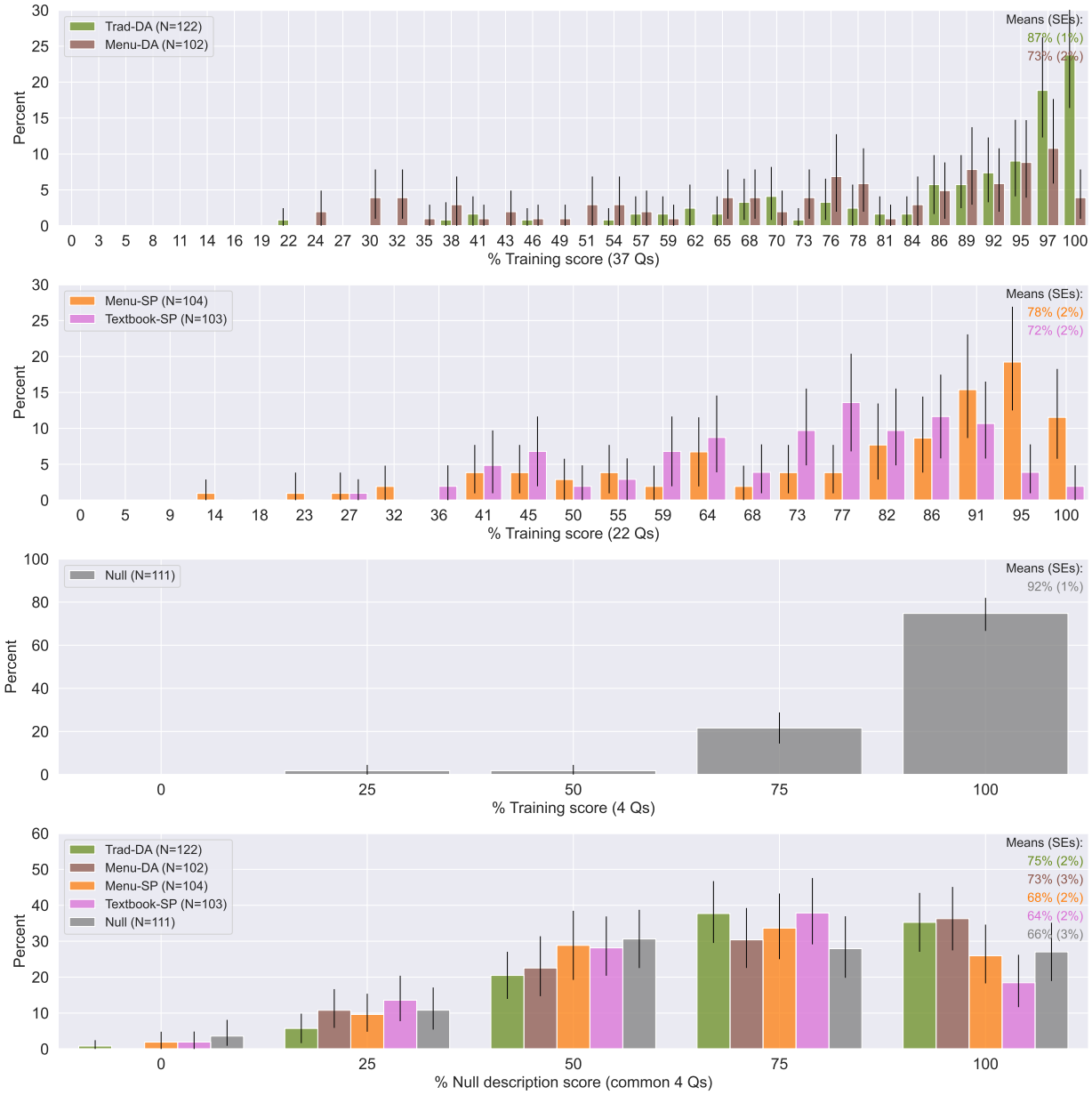
Next, Figure A.5 suggests that % SP-U can be understood as having two main ranges. The figure shows the means of Abstract and Practical conditional on different total % SP-U scores. The range of % SP-U $<$ 75% is governed mostly by the Abstract score, where Practical is approximately fixed in its low mode. The range of % SP-U $\geq$ 75% is governed mostly by the Practical score, where Abstract is approximately fixed near-perfect, and the weak bimodality observed in % SP-U in this range reflects Practical's bimodality.

A.4 Ranking Patterns Beyond % SF

We investigate the details of participants' ranking behavior.

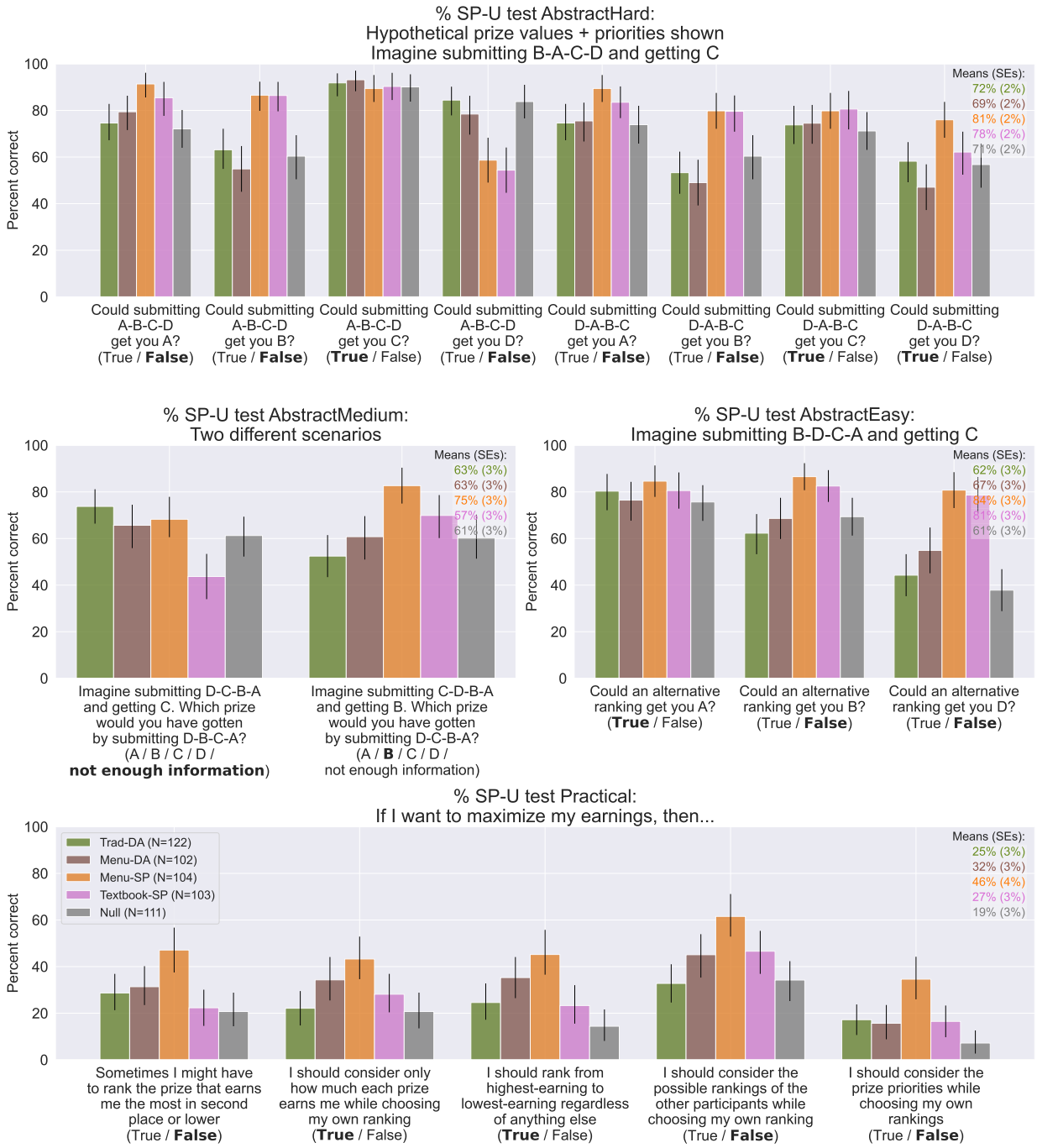
Figure A.6, top row, shows the distribution of the most common ranking patterns in the real rounds of playing DA. The SF behavior—ranking $[1, 2, 3, 4]$ where numbers reflect ordinal rank of prize value from highest- to lowest-earning—is the most common strategy across all treatments. Additionally, the 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. The figure's middle and bottom rows show this distribution among participants with different levels of SP understanding. Participants with low % SP-U

Figure A.2: Distribution of training scores and the Null-description training score



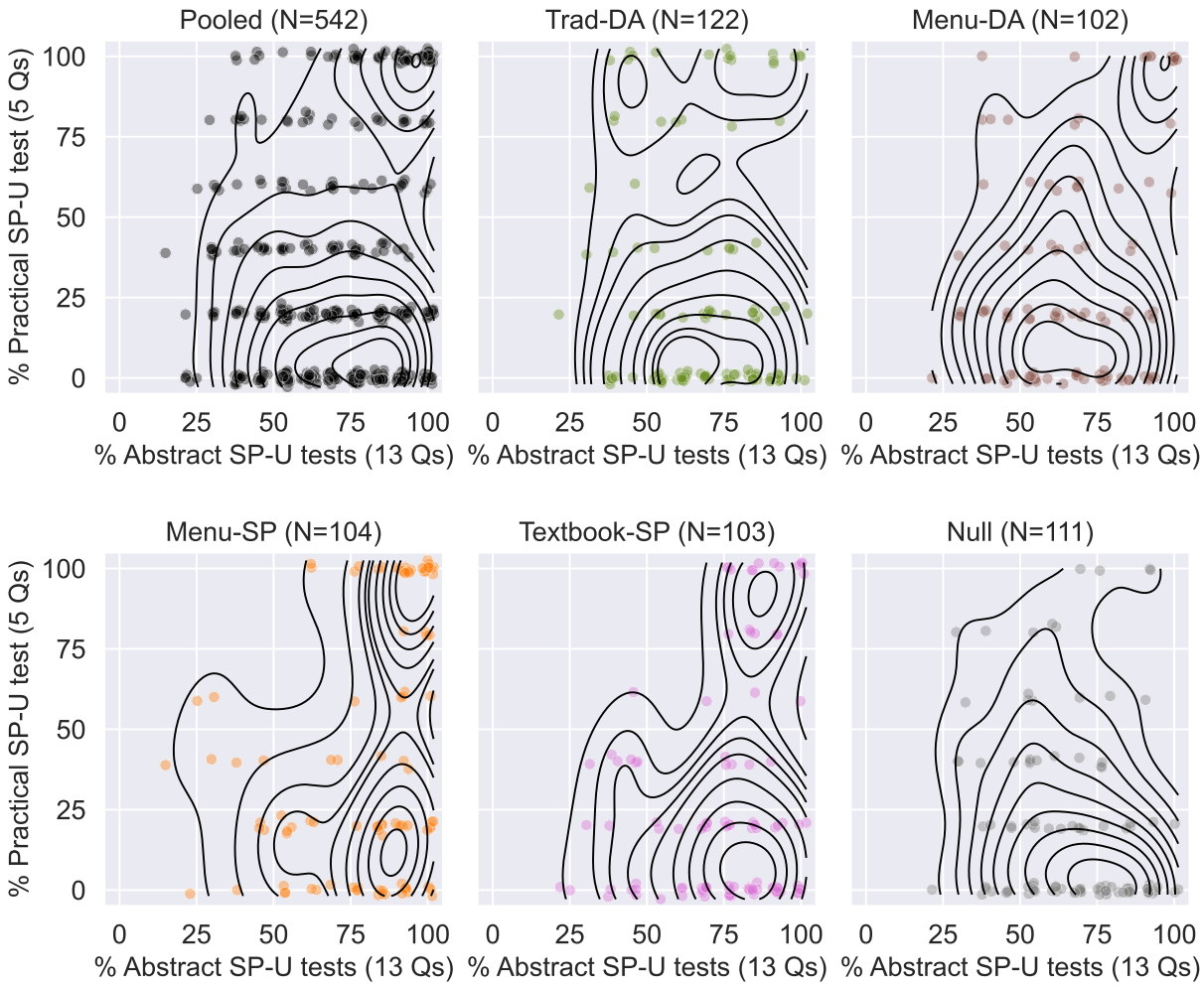
Notes: *First row*: The two DA Mechanics treatments. *Second row*: The two SP Property treatments. *Third row*: The Null treatment. *Fourth row*: The four Null training round questions, common to all treatments (and not included in their training scores). *All rows*: Mean values (SEs) are included in the upper right corner. "(# Qs)": the number of questions that the relevant score is based on. Error bars: 95% confidence intervals.

Figure A.3: Success rates in all SP-U questions



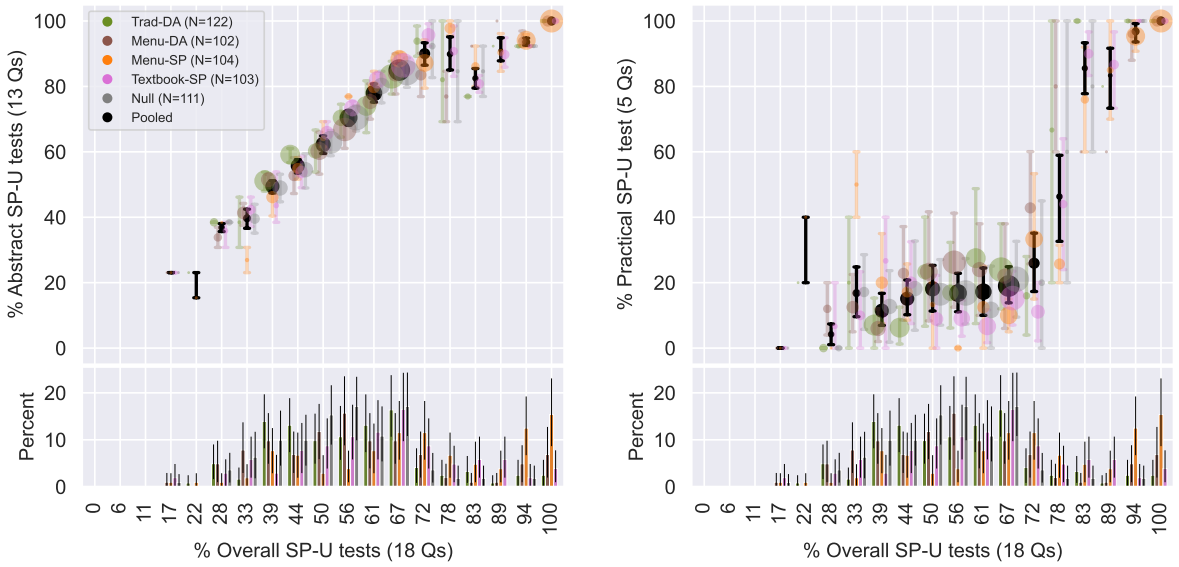
Notes: Each mini-figure shows the question-level success rates in one of the four SP-U sub-measures across treatments. Error bars: 95% confidence intervals.

Figure A.4: Joint distribution of the Abstract and Practical % SP-U sub-measures by treatment



Notes: Each mini-figure contains a jittered scatter plot and estimated contours smoothing the two-dimensional distribution. Smoothing is performed using a Gaussian kernel with a bandwidth chosen according to Scott's Rule (Scott, 1992).

Figure A.5: Relationship between % SP-U and its Abstract and Practical sub-measures



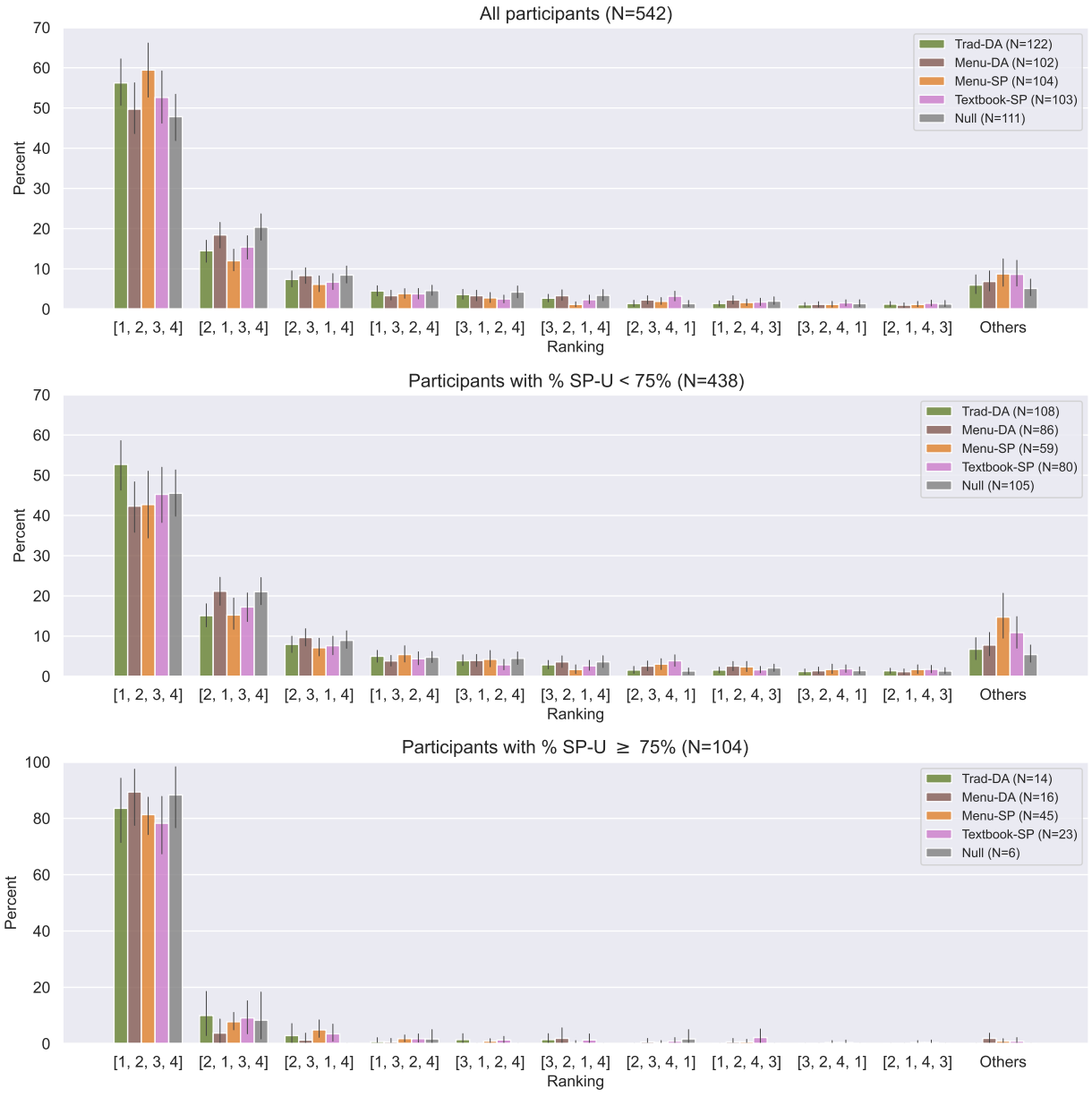
Notes: *Left*: Histogram of % SP-U at the bottom and the conditional means of the Abstract % SP-U sub-measure given all possible % SP-U scores at the top (see Figure 11, which is similar in structure, for more details). *Right*: a similar figure with the Practical sub-measure instead of Abstract.

(below 75%), who play NSF in higher rates, still mostly use the $[2, 1, 3, 4]$ strategy among NSF ones.

Next, Figure A.7 explores whether ranking behavior depends on key parameters of the DA rounds. The figure shows the frequency of three different categories of NSF ranking behavior, and compares these behaviors across different categories of DA rounds. The categories of ranking behavior 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).[22] The figure compares behavior across different binary conditions of the DA rounds. Each condition is such that, when the condition is true, we expect a higher rate of NSF play (under hypothesized common misconceptions, e.g., that prizes where the participant has low priority, or which do not offer a large increase in the money reward, are "not worth" ranking first). The conditions are: (1) the participant's priority of getting the highest-earning prize is not highest, (2) the participant's priority of getting that prize is strictly less than that of getting some other prize, (3) the

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

Figure A.6: Distribution of submitted rankings: overall and by SP understanding score



Notes: 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 prizes relative to the SF ranking ($[1, 2, 3, 4]$). Error bars: 95% confidence intervals.

difference between the monetary values of the highest- and second-highest-earning prizes is in the lower half of the distribution of such differences, and (4) a combination of the previous two categories.

Overall, we find that, in line with our expectations, NSF play generally increases when the conditions are true. However, that increase is fairly mild (generally around 10 percentage points), and we do not find any strong pattern across treatments beyond those conveyed by overall differences in mean % SF.

Figure A.8 shows that the weak patterns we find may be somewhat stronger for participants with % SP-U $< 75\%$, and are non-distinguishable from zero for participants with % SP-U $\geq 75\%$.

Next, we investigate whether ranking behavior changes over the 10 rounds that each participant play. Table A.1 shows the difference in participants ranking behavior in rounds 1–5 vs. rounds 6–10, and also shows the coefficient from an OLS regression of whether or not the participant played SF that round on the round number. Rates of SF play also seems stable across the 10 rounds, except perhaps 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.

Table A.1: SF play by round number and treatment

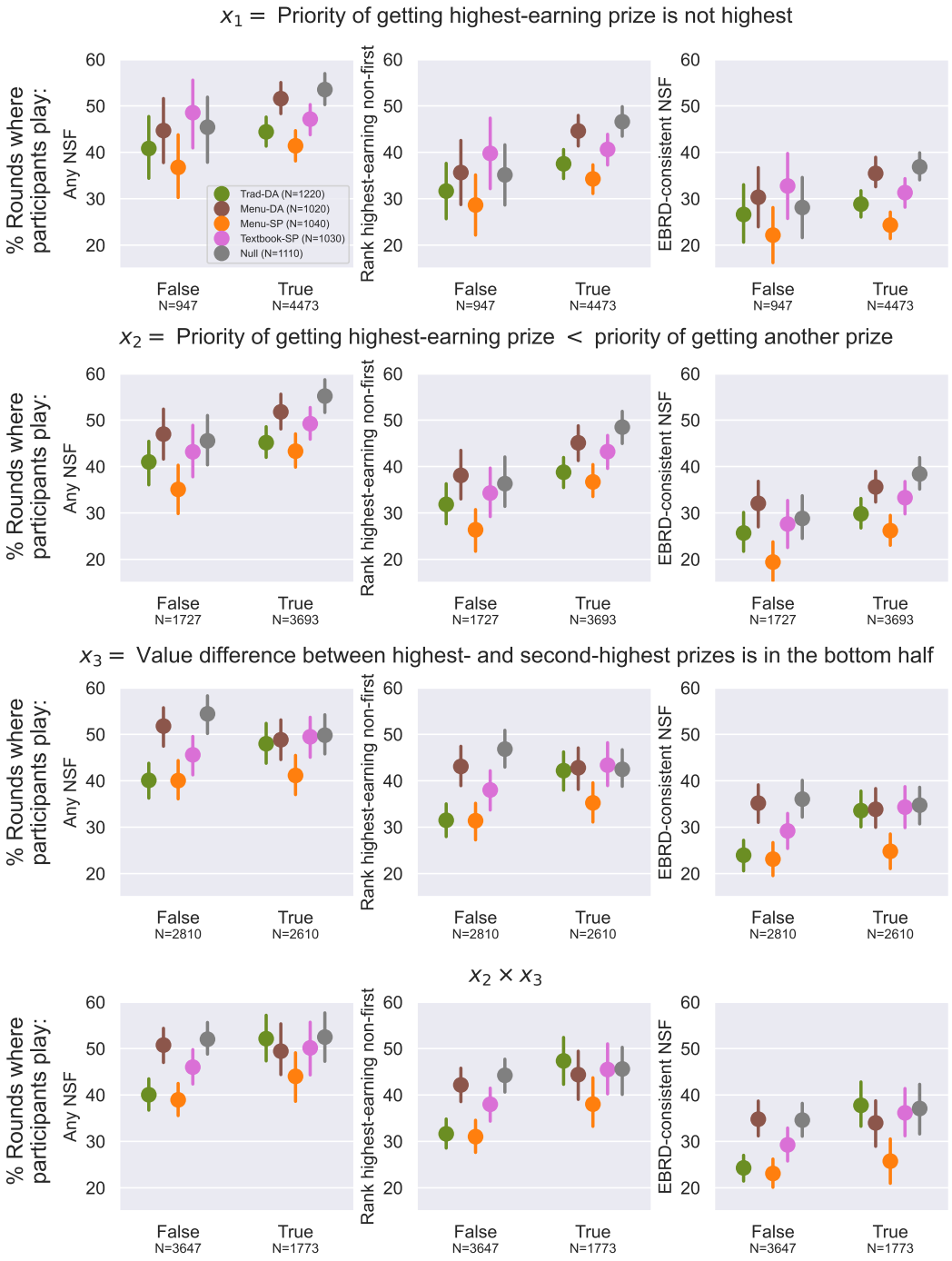
	Trad-DA		Menu-DA		Menu-SP		Textbook-SP		Null	
Rounds	1–5	6–10	1–5	6–10	1–5	6–10	1–5	6–10	1–5	6–10
% SF	57	56	49	50	56	63	54	52	47	48
	(3)	(3)	(3)	(4)	(4)	(4)	(3)	(4)	(3)	(3)
% All SF	26	26	22	28	28	36	23	25	20	21
	(4)	(4)	(4)	(4)	(4)	(5)	(4)	(4)	(4)	(4)
Slope	0.2 (0.5)		0.1 (0.5)		1.6 (0.5)		0.1 (0.5)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

Notes: "% All SF": fraction of participants with straightforward play in all relevant rounds (1–5 or 6–10). "Slope": coefficient from a OLS regression of an indicator for SF play (for an observation of participant × round) on round number. Robust standard errors in parentheses.

A.5 Relationship Between % SF and (Sub-Measures of) % SP-U

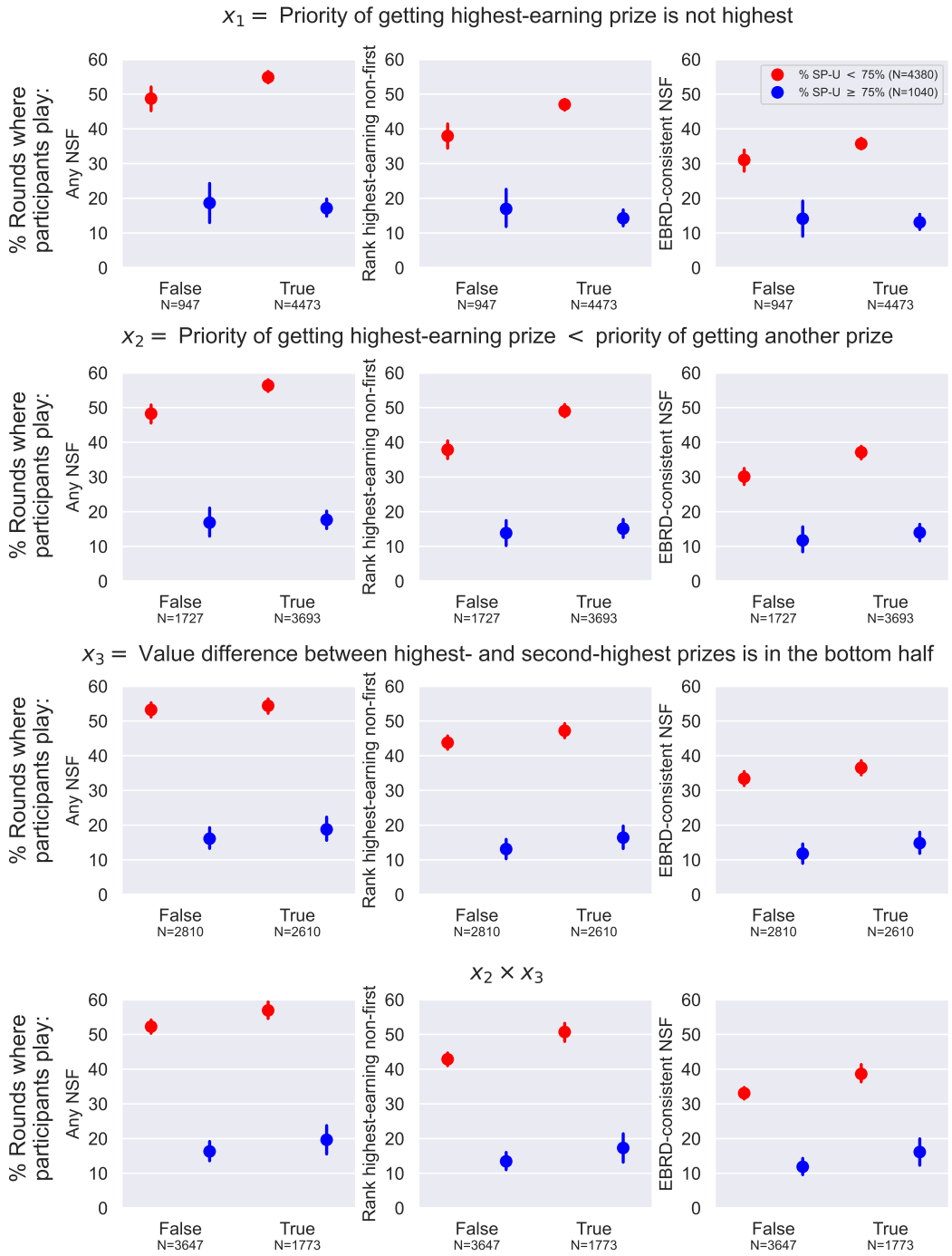
We first investigate the relationship between SP understanding and SF behavior shown in Figure 11 in more detail. Second, we investigate how the different sub-measures of % SP-U contribute to the overall relation.

Figure A.7: The frequency of different strategies conditional on round parameters



Notes: 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 for comparison. Error bars: 95% confidence intervals.

Figure A.8: The frequency of different strategies conditional on round parameters: participants with % SP-U $< 75\%$ vs. % SP-U $\geq 75\%$



Notes: See under Figure A.7, except that instead of comparing the five treatments, each panel compares two sub-samples of participants with % SP-U below and above 75%.

Figure A.9 shows the (non-conditional) joint distribution of % SP-U and % SF. The figure also displays the fractions of participants falling into four quadrants in the full joint distribution of % SP-U and % SF based on the threshold of 75% for both measures. The 75% cutoff of % SP-U and % SF is motivated by the location of the modes in Figure 8(b) and Figure 10(b), and of the jump in Figure 11.

We make two observations regarding Figure A.9, which add to the findings in Figure 11. First, 32% (SE = 5%) of participants in Menu-SP 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. Thus, Menu-SP is most effective in moving participants into the upper quadrant of the % SP-U, % SF distribution. Second, the figure suggests a *two*-dimensional bimodality of % SP-U and % SF in Menu-SP (and perhaps also in Menu-DA and Textbook-SP, albeit a weaker one).

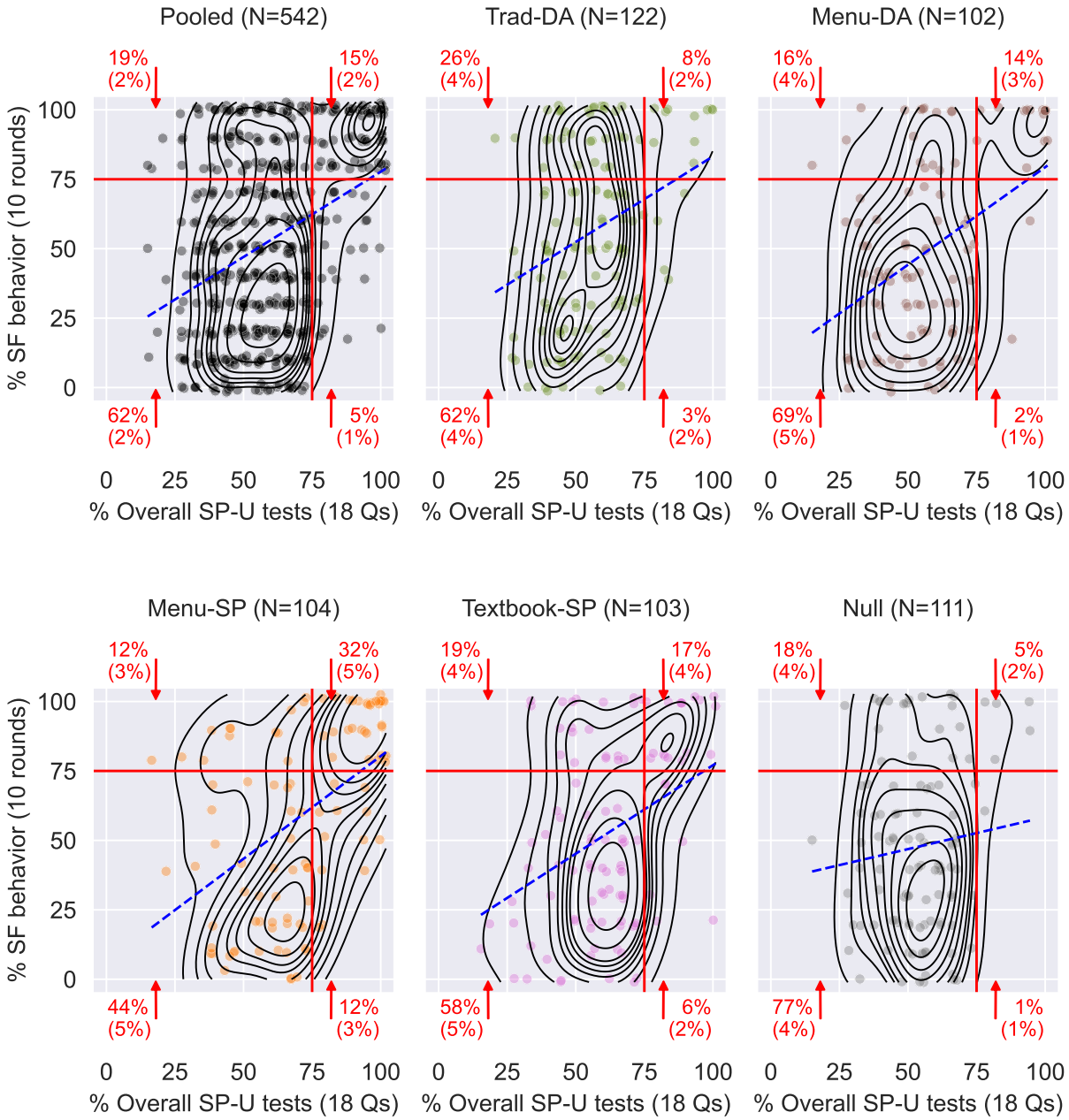
Next, we investigate how the different sub-measures of % SP-U contribute to its overall relation with % SF shown in Figure 11.

Figure A.10 shows two versions of Figure 11 focusing on Abstract and Practical *separately*. While Abstract is able to explain a small amount 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.

Finally, we investigate the extent to which Abstract and Practical are *jointly* required to explain the increase in % SF rates. Table A.2 summarizes OLS regressions showing the average relation between % SF, Abstract and Practical, and total % SP-U. (For a breakdown of Abstract into its three sub-measures AbstractHard, AbstractMedium, and AbstractEasy, see Appendix B.5.2 in the Supplementary Materials.) The regression results suggest that Practical is much more strongly related to % SF than Abstract or its sub-measures. They also suggest that the relation between Abstract alone and % SF is mostly due to the correlation between Abstract and Practical, since the coefficient on Abstract becomes close to zero when adding Practical as another independent variable. These findings are robust to the inclusion of control variables and treatment indicators.

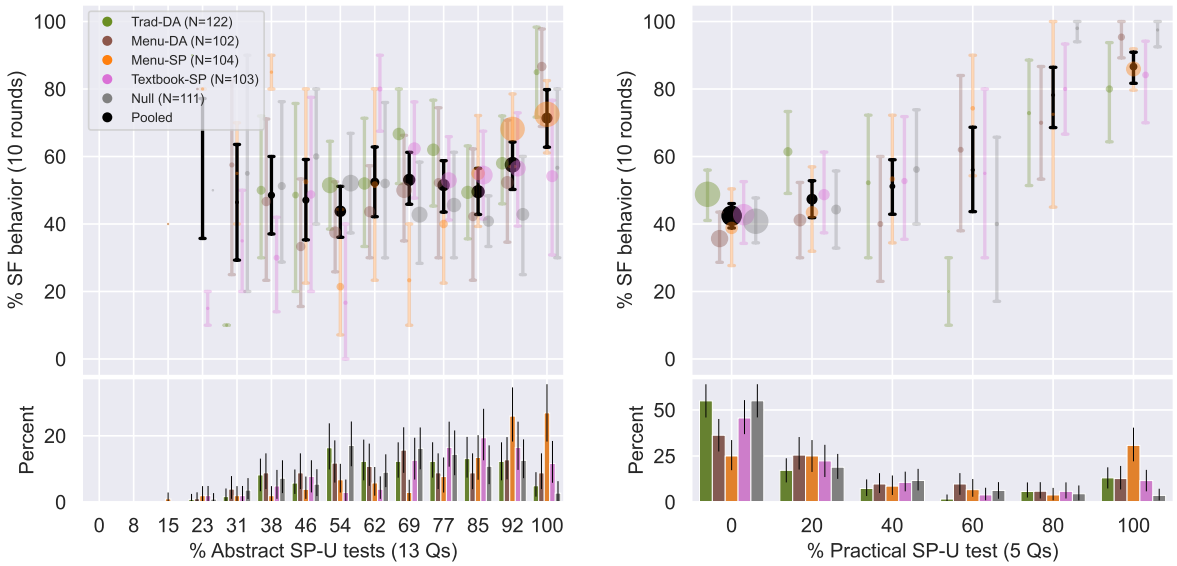
Despite that Abstract has an average insignificant relation to % SF when considered jointly with Practical, it may contribute to the overall relation in more specific cases. Figure A.11 non-parametrically investigates % SF rates at different combinations of Abstract and Practical. Pooling all treatments 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%,

Figure A.9: Joint distributions of % SP-U and % SF by treatment



Notes: Each mini-figure contains a jittered scatter plot and estimated contours smoothing the two-dimensional distribution. Smoothing is performed using a Gaussian kernel with a bandwidth chosen according to Scott's Rule (Scott, 1992). *Dashed blue lines*: Predicted % SF values according to OLS regressions on % SP-U (see Table B.1 in Appendix B.5.1 in the Supplementary Materials for full regression results and more details). *Solid Red lines*: % SP-U = 75% and % SF = 75% lines. *Red numbers*: fractions of participants in the four quadrants separated by the solid red lines (with SEs in parentheses).

Figure A.10: Relationship between % SF and the Abstract and Practical sub-measures of % SP-U



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

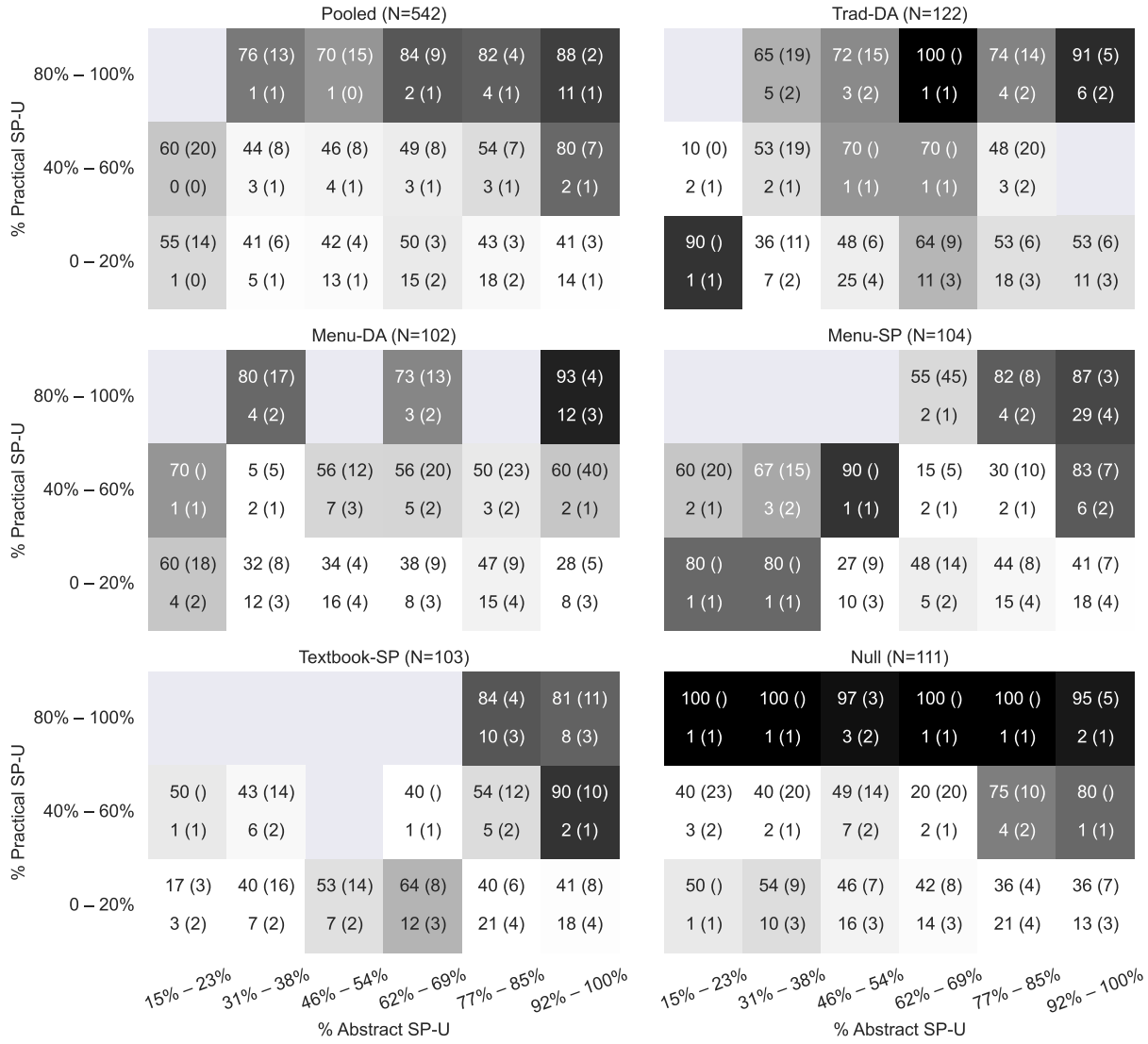
Table A.2: Relation of Abstract and Practical % SP-U sub-measures with SF behavior

Dependent variable: % SF

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
% Abstract	0.26	0.05			0.11	−0.00		
	(0.07)	(0.10)			(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	0.45	0.22	0.45	0.13	0.37
N	542	542	542	542	542	542	542	542

Notes: Coefficients from OLS regressions of % SF on % SP-U, including a separation of % SP-U into its sub-measures. Each regression is shown without and with including a full set of controls in the regression, described in Appendix B.2 in the Supplementary Materials. Treatment indicators are also included in the controlled regressions. Robust standard errors in parentheses.

Figure A.11: % SF rates and fractions of participants at different combinations of Abstract and Practical



Notes: 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 sub-measure. Standard errors are shown in parentheses, where no SE at the top row indicates a single-observation cell.

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 have these mid-level Practical scores, and there is insufficient statistical power to test whether it persists within treatments.

A.6 Prolific vs. Cornell Samples, Cognitive Score, Attention Score, and their Mediating Effects on Main Results

Among possible variables mediating our treatment effects, education, cognitive ability, attention, and ability to focus in a physical lab (vs. remote participation) 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.

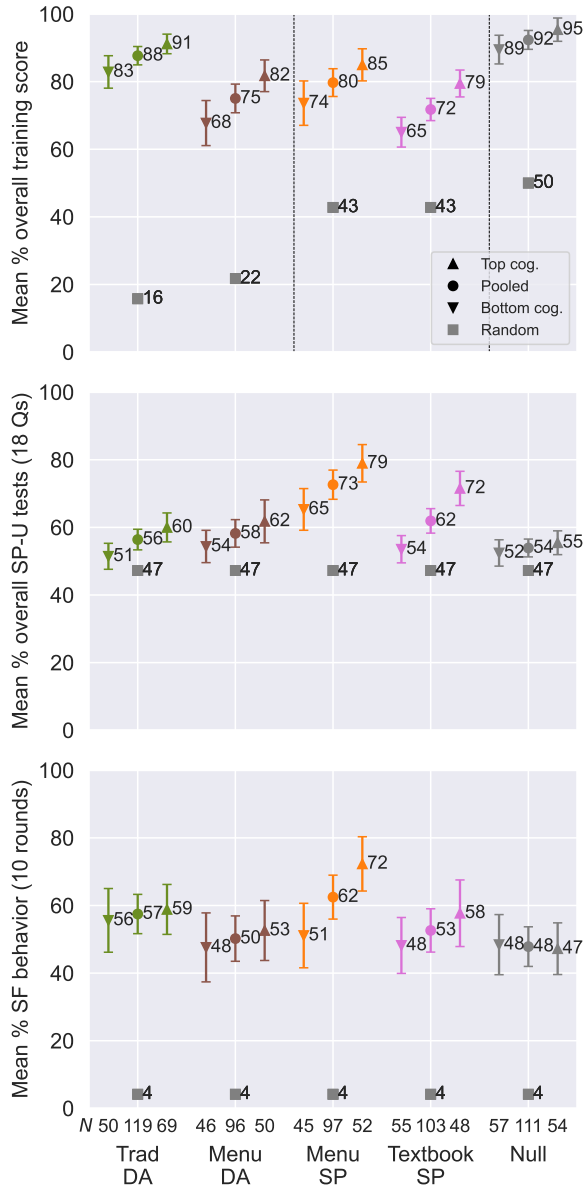
We use three key measures to study these effects: an indicator for whether a participant belongs to the Prolific or Cornell sub-sample, a cognitive score measure elicited at the end of the experiment, and an attention score elicited in two attention-check questions planted within the experiment. In addition, and according to our pre-registration, we test whether participants with extreme experiment duration—who take very little or very long time to complete the experiment compared to their treatment distribution—drive our main results. See Appendix B.4 in the Supplementary Materials for texts of our cognitive questions and attention-check questions, and for the distributions of cognitive and attention scores in our sample.

Figure 12 in Section 3.5 tests the effects of these measures on our main findings, by reporting differences between the Prolific and Cornell samples. Figure A.12 similarly adds a comparison of the bottom (roughly) half of cognitive scores (0–2) vs. the top half (3–4), and Figure A.13 adds a comparison of bottom attention scores (0–1) vs. a top (perfect) attention score (2). Both of these comparisons yield similar insights to the Prolific versus Cornell comparison from Section 3.5.

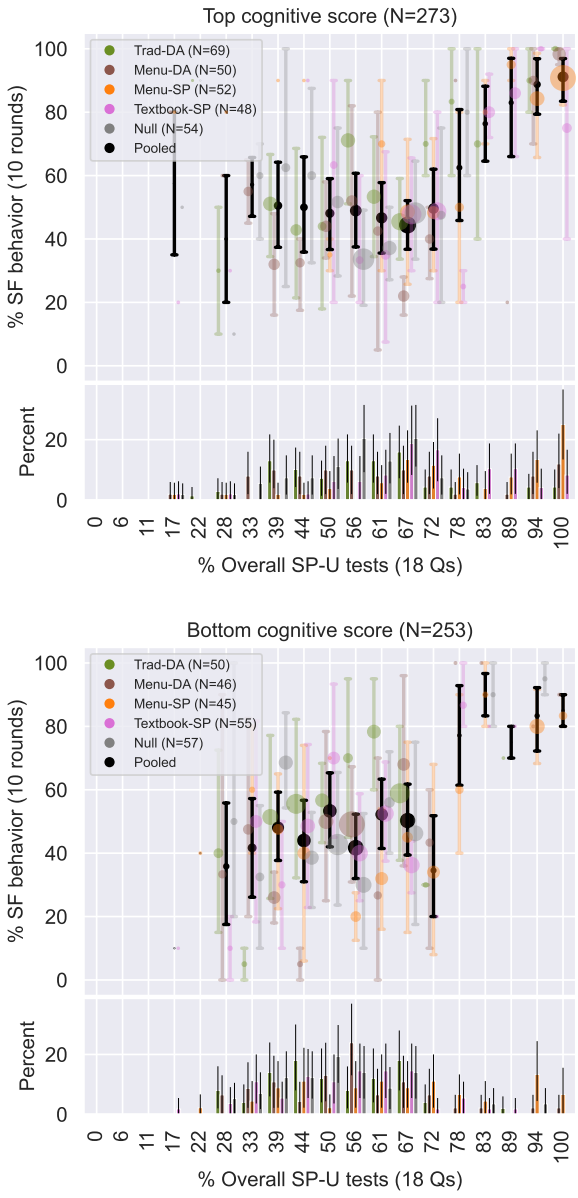
Finally, Figure A.14 reports a comparison of participants with non-extreme per-treatment completion time to participants with an extreme completion time. According to our pre-registration, extreme completion time is defined as belonging to the top or bottom 2.5 percentages of the per-treatment duration distribution. The per-treatment 2.5-percentile, median and 97.5-percentile duration are as follows: (31, 48, 95) minutes in Trad-DA, (30, 53, 101) minutes in Menu-DA, (19, 40, 75) minutes in Menu-SP, (22, 39, 78) minutes in Textbook-SP, and (21, 35, 68) minutes in Null. Our main results (among the full sample) remain largely unaffected when excluding extreme completion times.

Figure A.12: Main results by top vs. bottom cognitive scores

(a) Mean outcome variables by cognitive score



(b) Relation between % SF and % SP-U by cognitive score

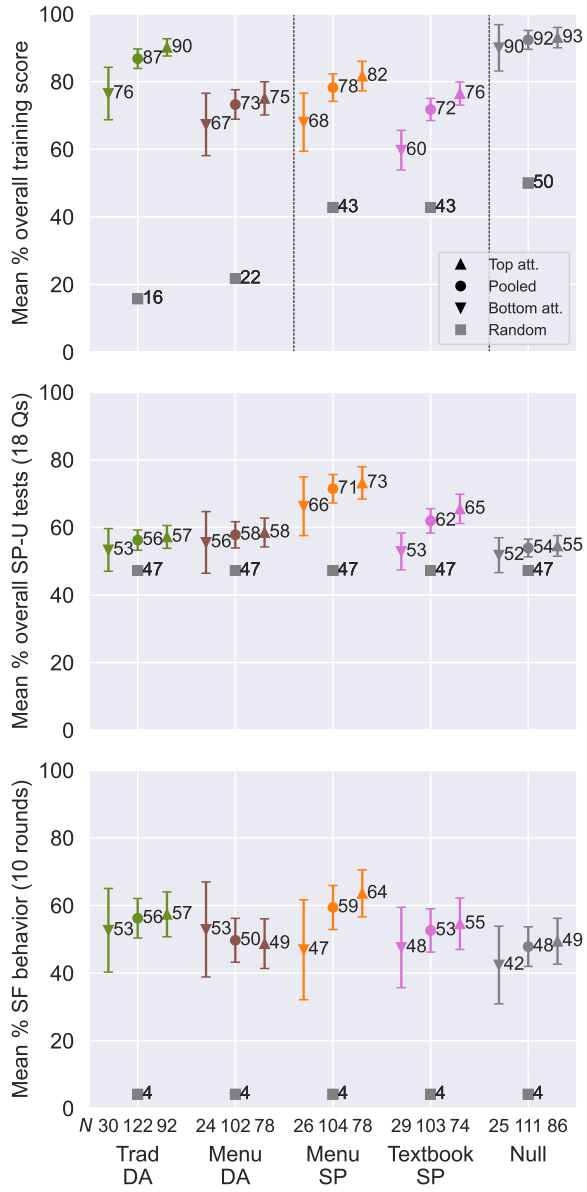


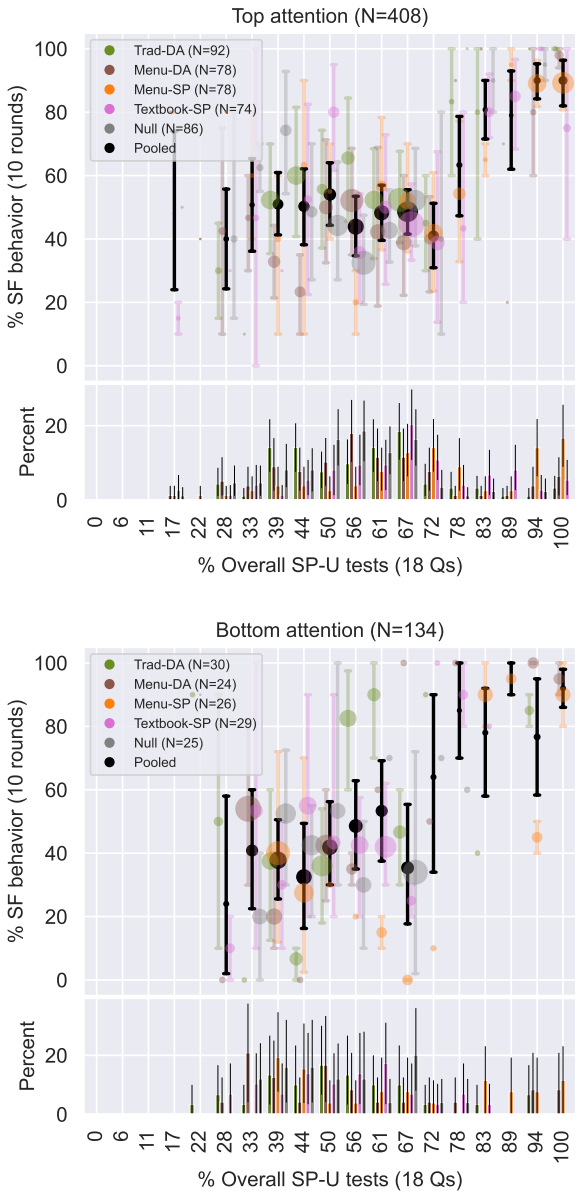
Notes: See under Figure 12; the sample split is according to top (3–4) vs. bottom (0–2) scores in the cognitive-ability test conducted at the end of the experiment.

Figure A.13: Main results by top vs. bottom attention scores

(a) Mean outcome variables by attention score

(b) Relation between % SF and % SP-U by attention score



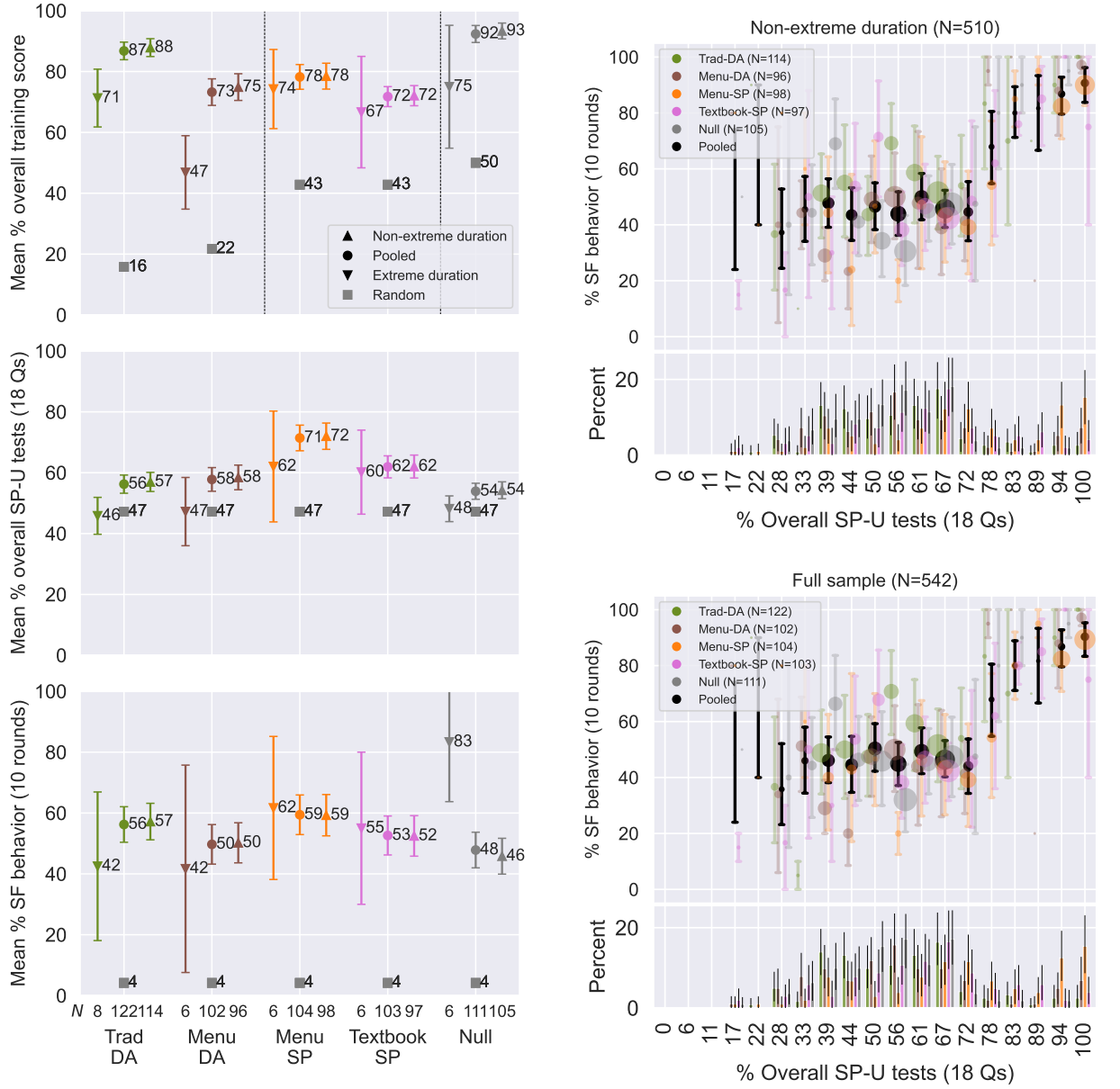


Notes: See under Figure 12; the sample split is according to top (2) vs. bottom (0–1) scores in the attention-check questions planted within the experiment.

Figure A.14: Main results by non-extreme vs. extreme completion time

(a) Mean outcome variables by completion time

(b) Relation between % SF and % SP-U by completion time



Notes: See under Figure 12; on the left, the sample split is according to non-extreme vs. extreme experiment duration (defined as being $\leq$ the 2.5 percentile or $\geq$ the 97.5 percentile of the per-treatment duration distribution); on the right, the sub-sample with non-extreme duration is compared to the full sample, since the extreme-duration sub-sample is too small for a meaningful comparison.

A.7 Robustness of Main Results to Adding Controls

We test whether our four main findings are robust to adding a rich set of controls as explanatory variables.

We first find that adding controls does not change the treatment-means of our three main outcome variables by much. Table A.3 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 Figures 7(a), 8(a), and 10(a) in terms of difference from the Null treatment mean. The control variables include (1) a set of demographic indicators, (2) indicators for every possible cognitive and attention scores (described in Appendix B.4 in the Supplementary Materials) and for every possible score in the Null training questions common to all treatments, and (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.

The demographic characteristics (described fully in Appendix B.2 in the Supplementary Materials) include state of residence (not used in regressions), age, number of people in household, number of people in household aged at least 18, gender, race, education, primary education focus, knowledge of the DA mechanism, participation in real-life DA, marital status, employment status, social views, economic views, identification with a political party, vote in the 2020 presidential elections, and combined household income.

Table A.3: Treatment effects on outcome variables without controls and with controls

	% TR	% TR	% SP-U	% SP-U	% SF	% SF
Trad-DA	−0.06	−0.08	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	0.13
	(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	0.63	0.48	0.11
	(0.01)	(0.14)	(0.01)	(0.11)	(0.03)	(0.24)
Controls		X		X		X
R^2	0.16	0.60	0.10	0.51	0.02	0.31
N	542	542	542	542	542	542

Note: Coefficients from OLS regressions of the 3 outcome variables on treatment indicators (with Null treatment the omitted category), both without and with the inclusion of a full set of controls in the regression, described in Appendix B.2 in the Supplementary Materials. Robust standard errors in parentheses.

Second, we find that adding controls does not change the relationship between % SP-U and % SF shown in Figure 11 by much. Table A.4 shows mean % SF when conditioning on SP-U $< 75\%$ vs. % SP-U $\geq 75\%$, within treatments. Comparing regression results without and with inclusion of controls suggests that the relationship between % SP-U and % SF remains consistent across treatments when adding controls.

Table A.4: Within-treatment relation between % SF and % SP-U without controls and with controls

Dependent variable: % SF

	(1)	(2)
Trad-DA × % SP-U $< 75\%$	0.53	0.53
	(0.03)	(0.03)
Trad-DA × % SP-U $\geq 75\%$	0.84	0.87
	(0.06)	(0.08)
Menu-DA × % SP-U $< 75\%$	0.42	0.41
	(0.03)	(0.04)
Menu-DA × % SP-U $\geq 75\%$	0.89	0.93
	(0.05)	(0.08)
Menu-SP × % SP-U $< 75\%$	0.43	0.47
	(0.04)	(0.04)
Menu-SP × % SP-U $\geq 75\%$	0.81	0.78
	(0.03)	(0.05)
Textbook-SP × % SP-U $< 75\%$	0.45	0.43
	(0.04)	(0.04)
Textbook-SP × % SP-U $\geq 75\%$	0.78	0.79
	(0.05)	(0.07)
Null × % SP-U $< 75\%$	0.46	0.47
	(0.03)	(0.03)
Null × % SP-U $\geq 75\%$	0.88	0.77
	(0.06)	(0.15)
Controls		X
R^2	0.20	0.43
N	542	542

Note: Coefficients from OLS regressions of % SF on indicators of % SP-U $< 75\%$ vs. % SP-U $\geq 75\%$, interacted with all five treatments, both without and with the inclusion of a full set of controls in the regression, described in Appendix B.2 in the Supplementary Materials. Robust standard errors in parentheses.

Supplementary Materials

B Supplemental Information and Analysis

In this appendix, we present further details, results, and analyses of our data.

B.1 Comparison of Analysis and Findings With Pre-Registration

Our pre-registration can be found at https://aspredicted.org/7eq7e.pdf. Our pre-registration specified our treatments and main outcome variables (% TR, % SP-U, and % SF), our sample size (see Appendix A.1), and our robustness check for outliers in completion time, (defined as bottom or top per-treatment 2.5%; test conducted in Appendix A.6). Our main pre-registered 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.[23] First, we hypothesized that Menu-DA would make it more complicated to understand DA mechanics than Trad-DA (which would be indicated by lower % TR in Menu-DA on comparative questions). 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).

While our results are organized differently than these hypotheses, they follow the analysis plan implied by our pre-registration by considering the mean levels of our main outcome variables. Our first three main results test and expand on the first two hypotheses, by comparing mean levels of % TR and % SP-U not only across Trad-DA and Menu-DA, but also across all other treatments. In a secondary analysis, reported in Section 3.5 and expanded in Appendix B.5.3 in the Supplementary Materials, we test the third hypothesis. Our fourth main result, on the correlation between % SP-U and % SF, is ex-post: We did not have clear hypotheses on how % SF would compare across treatments and on whether %

[23] 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.

SP-U and % SF behavior would correlate, since this requires assumptions on participants' preferences.

In contrast to our pre-registered 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 find little difference in the correlations between % TR and % SP-U across Trad-DA and Menu-DA, and a strong, non-treatment-specific relation between % SP-U and % SF. Our pre-registered robustness test verifies that completion-time outliers do not drive our main results.

B.2 Demographics

The last screen of the experiment elicits demographic characteristics of our participants. We now describe the categories used and their mean levels across treatments.

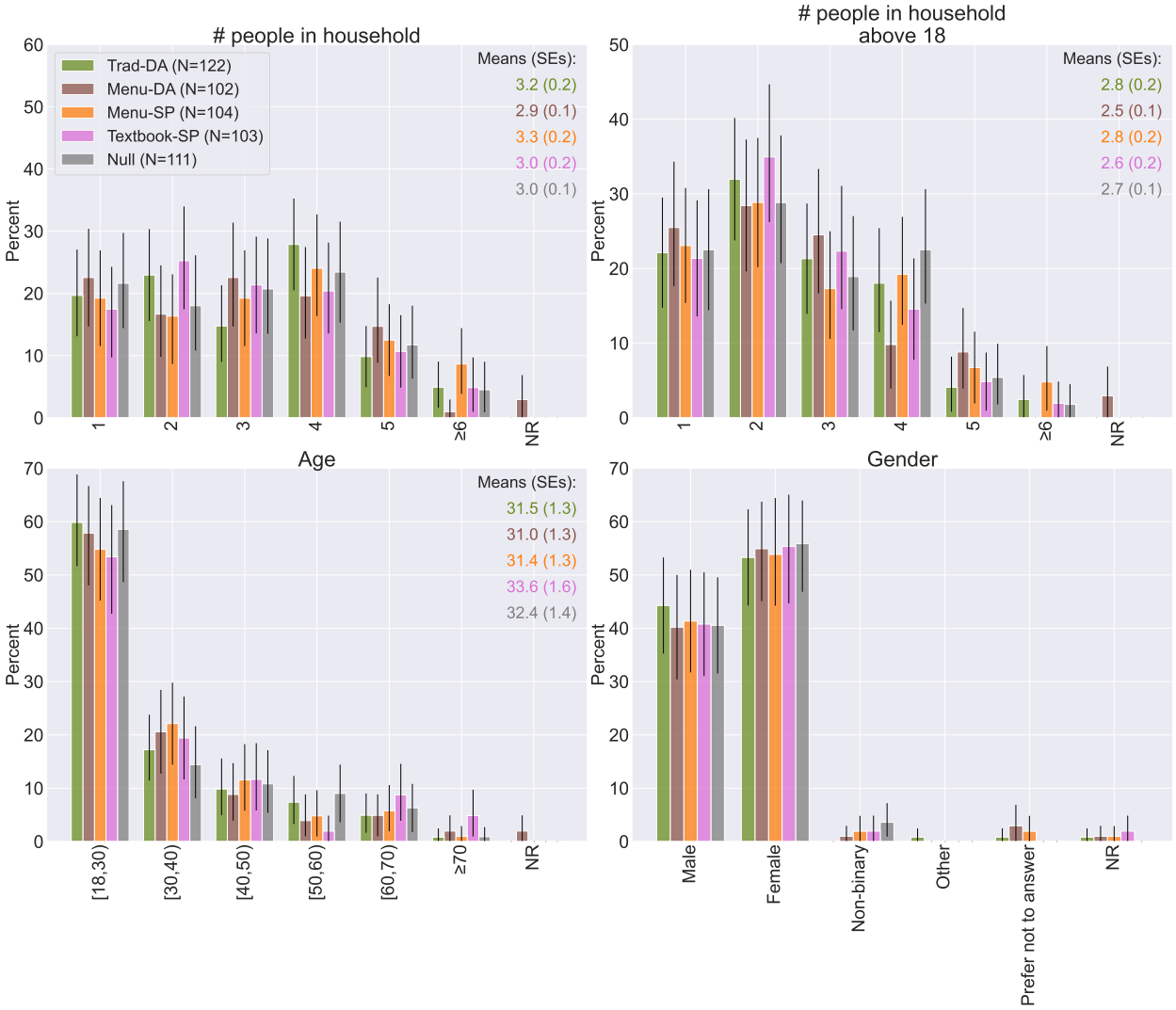
The set of demographic characteristics and their categories are as follows:

- State of residence (a choice of a US state)
- 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-year 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).

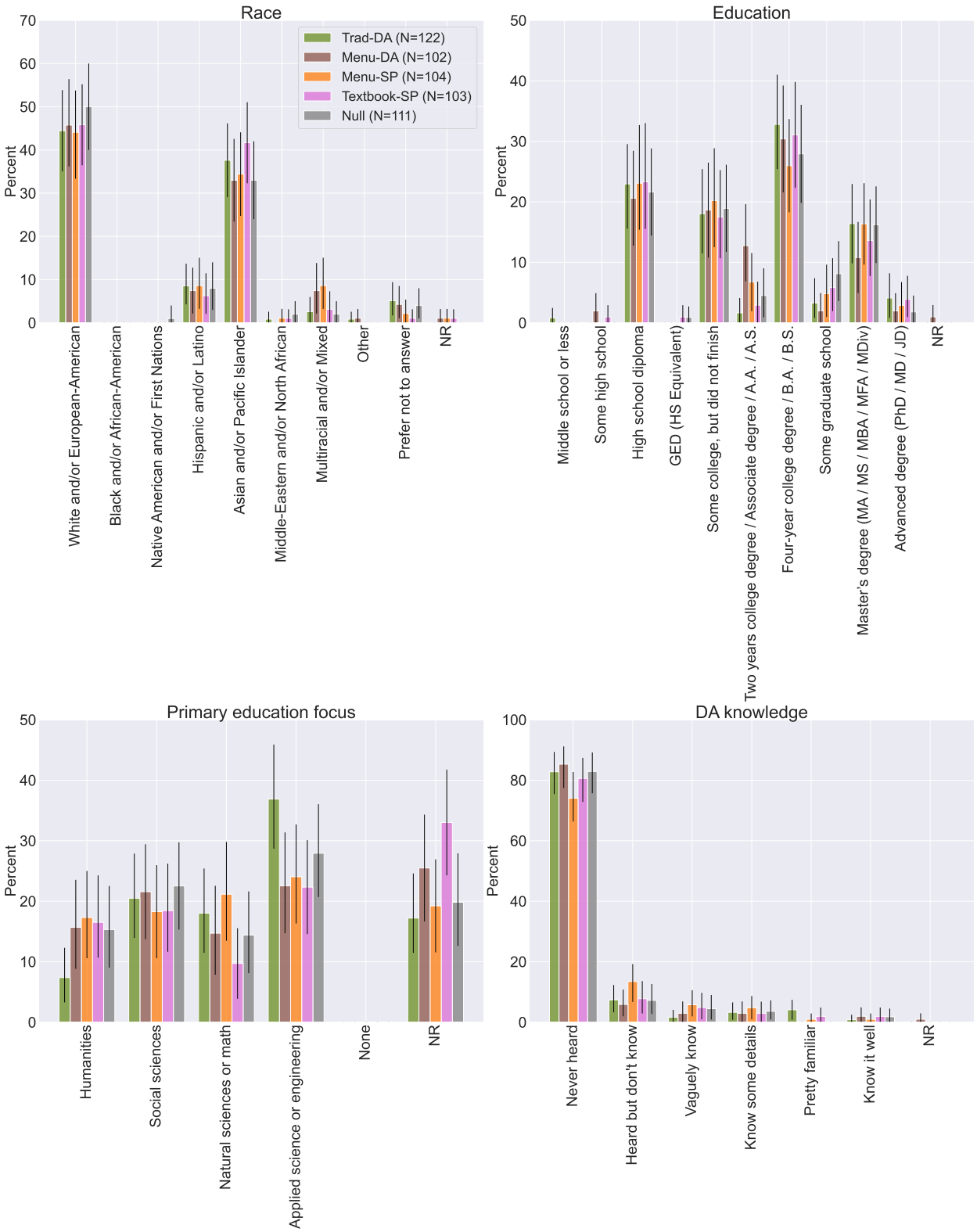
Figures B.1, B.2, B.3, and B.4 show the distribution of these characteristics in our sample across treatments (except for US state of residence).

Figure B.1: Distribution of demographic characteristics in our sample (1/4)



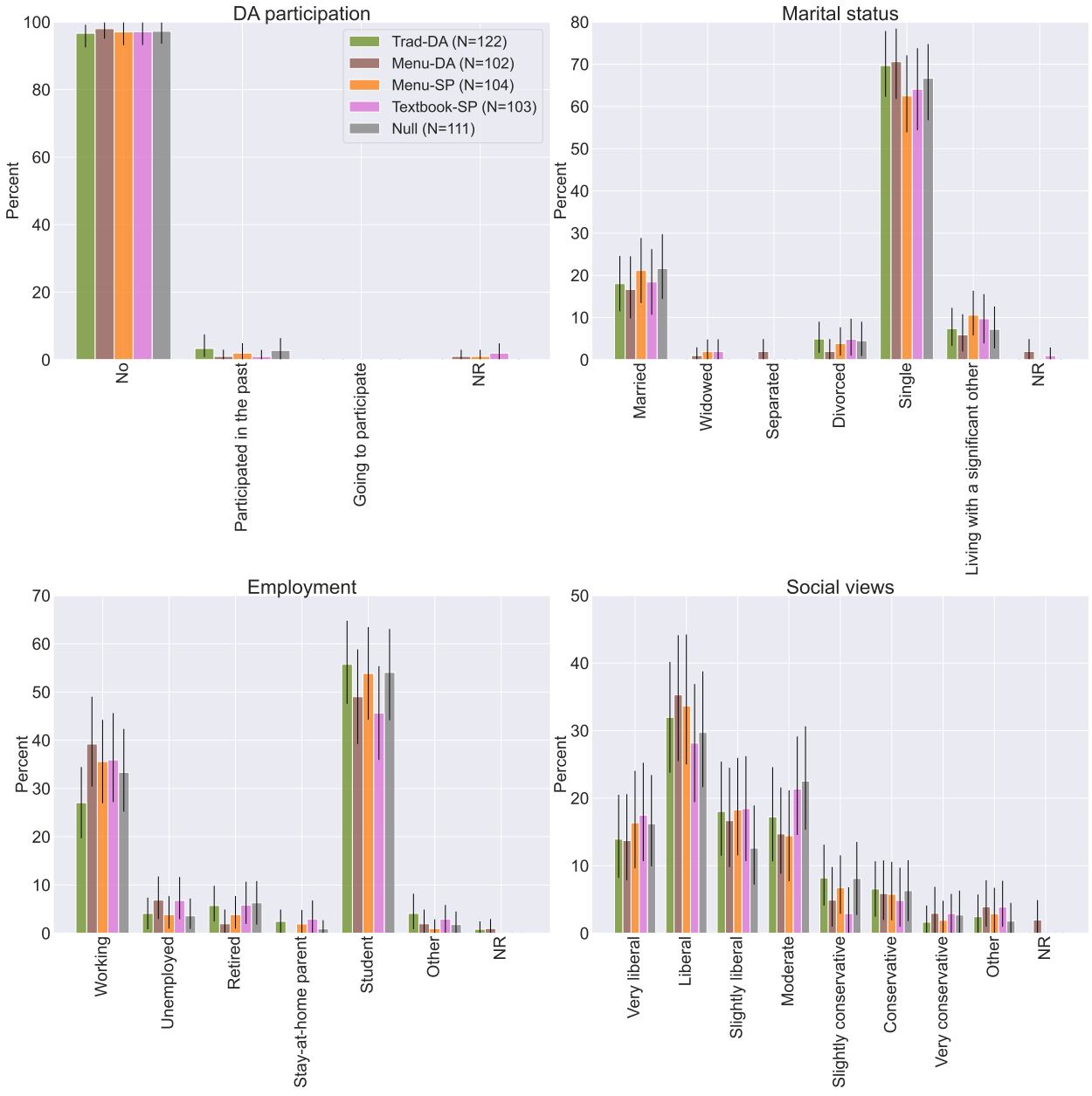
Notes: "NR": No response. Error bars: 95% confidence intervals.

Figure B.2: Distribution of demographic characteristics in our sample (2/4)



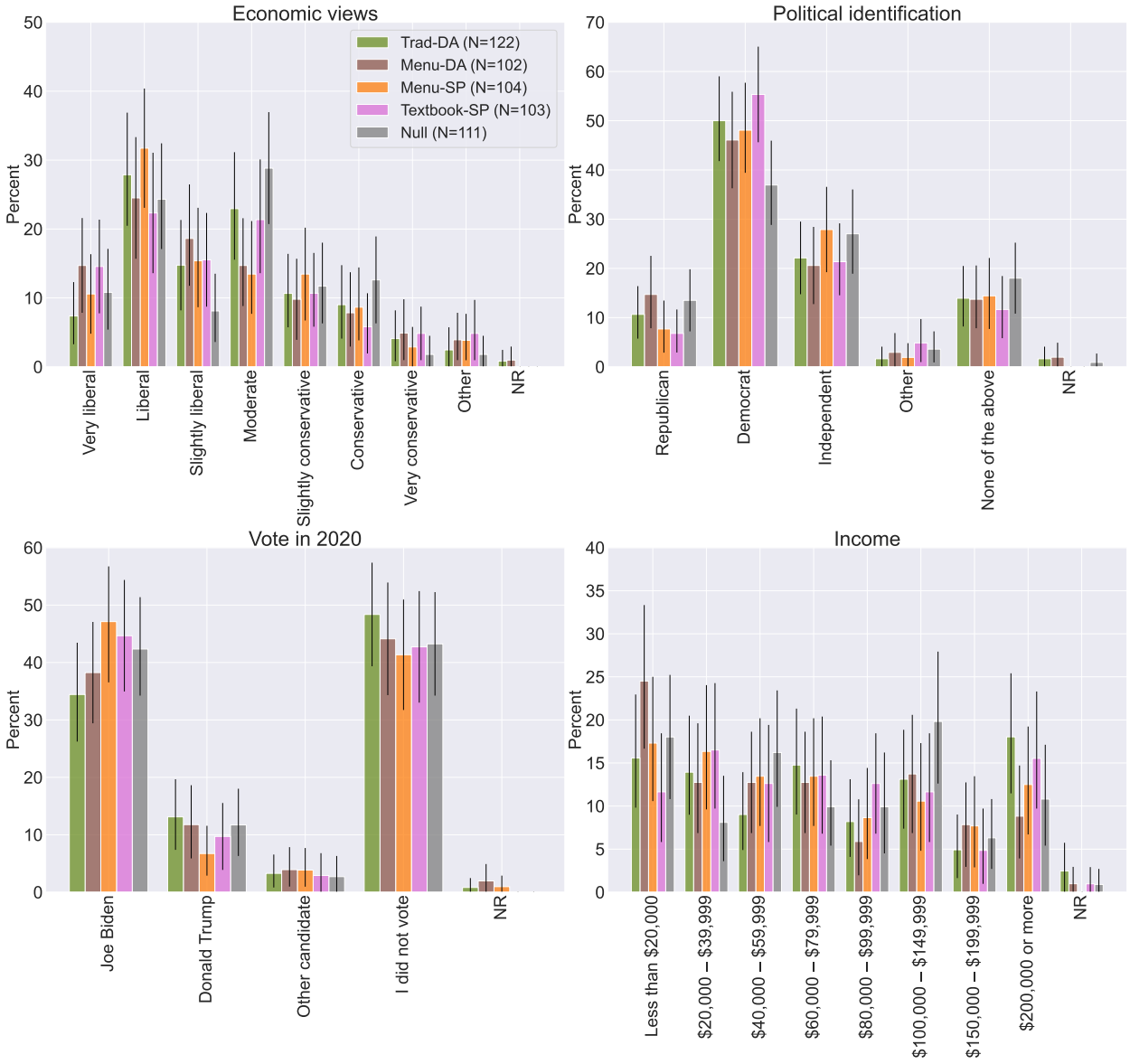
Notes: See under Figure B.1.

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



Notes: See under Figure B.1.

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



Notes: See under Figure B.1.

B.3 Detailed Performance in Training Questions

We present the details of participants' performance in training questions. We first show the overall training-score distributions, and then show mean performance at the question level within treatments.

B.3.1 Null Training Common to All Participants

Figure B.5 shows the distribution of participants' performance in the training round following the Null description, which is common to all participants and is not included in the treatment-specific training scores. There are four true/false questions about statements—shortly summarized above each top-row panel.

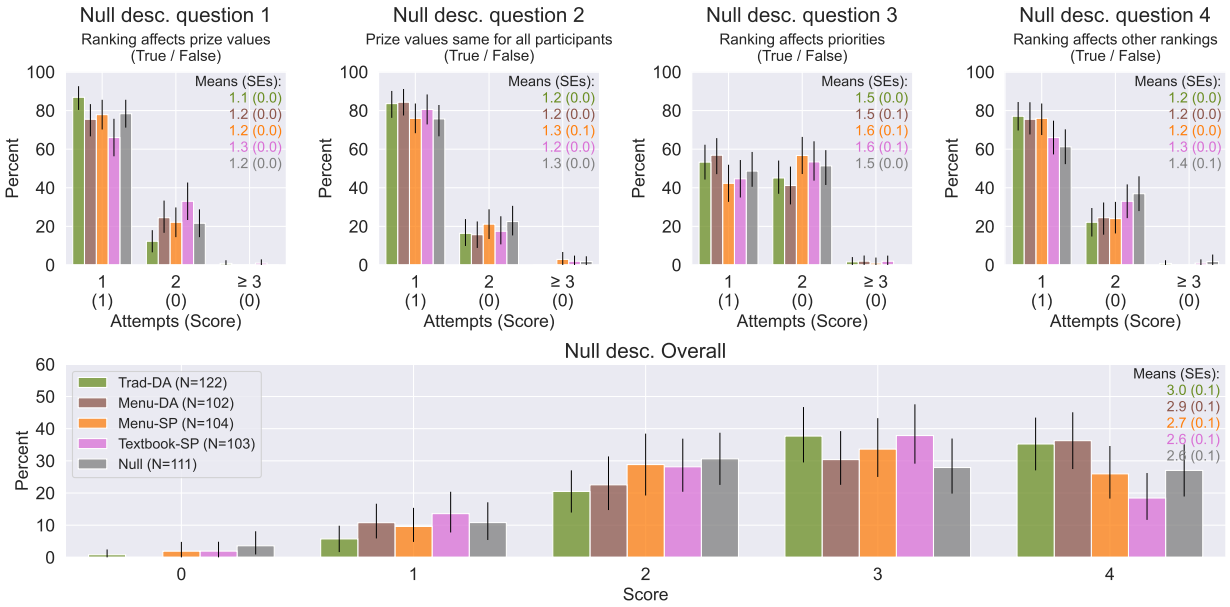
In this figure and others like it in the appendix, instead of only showing the binary score for each training question, we show more details on the number of attempts it took to answer the question correctly. Recall that participants cannot advance beyond a question until getting it correctly and receiving 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; we show the score awarded in parentheses below the number of attempts. The overall score in the round is displayed at the bottom row of the figure (which is also shown in the bottom row of Figure A.2 in terms of % success).

B.3.2 DA Mechanics Training

Figures B.6, B.7, B.8 and B.9 show participants' performance in the DA Mechanics training questions, including the DA-algorithm proposals-outcome calculations they needed to perform using the GUI, as well as their usage of an additional walkthrough video available in training round 2.

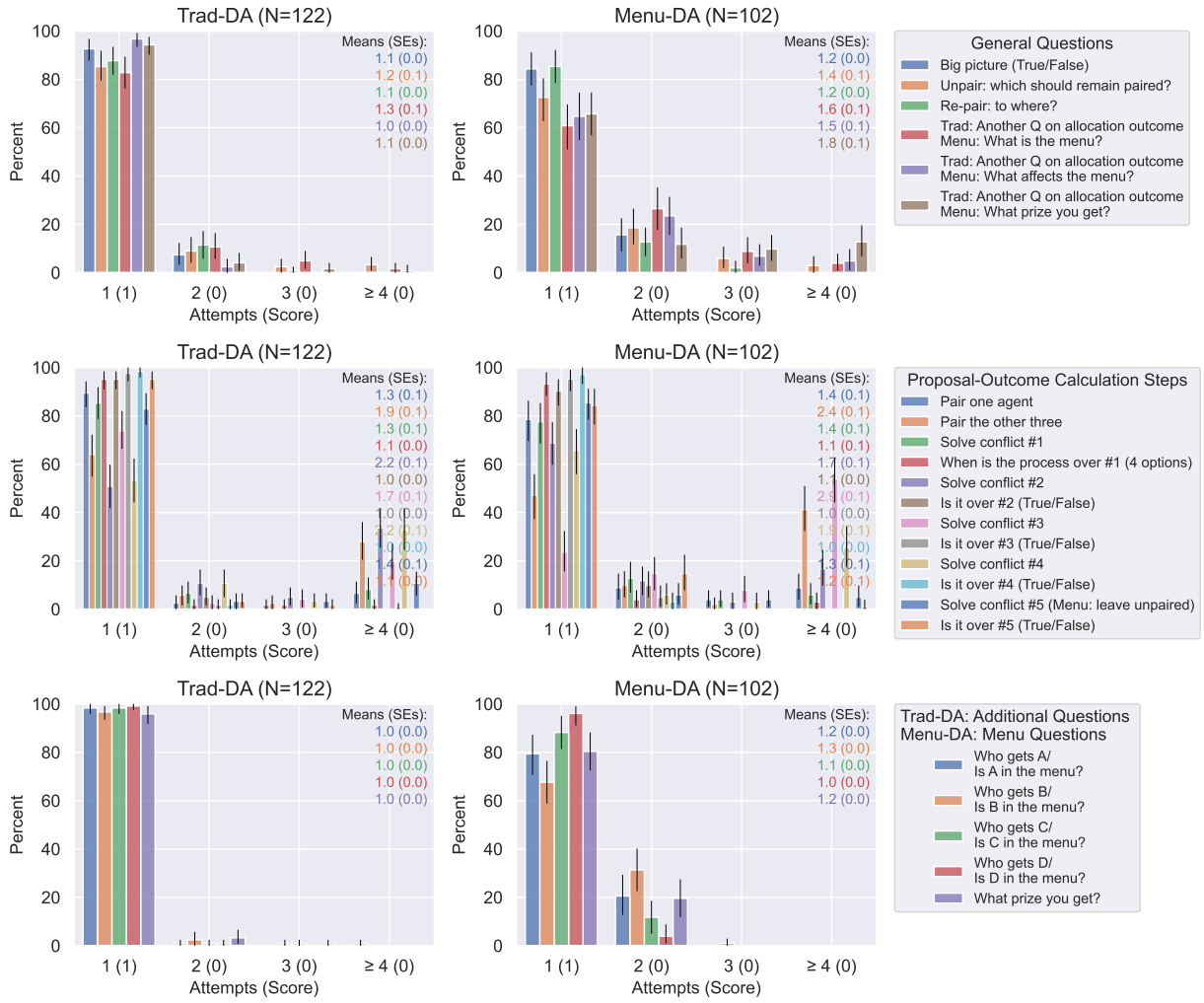
In DA Mechanics treatments, any error on an individual step of the DA-algorithm proposals-outcome calculation should, in theory, make the overall proposals outcome erroneous. However, despite a fairly high rate of errors on some calculation steps in training round 1, the overall success in rounds 2 and 3 is higher. This suggests that training round 1 was effective at teaching participants how to perform these calculations, perhaps through trial and error.

Figure B.5: Null training questions common to all participants



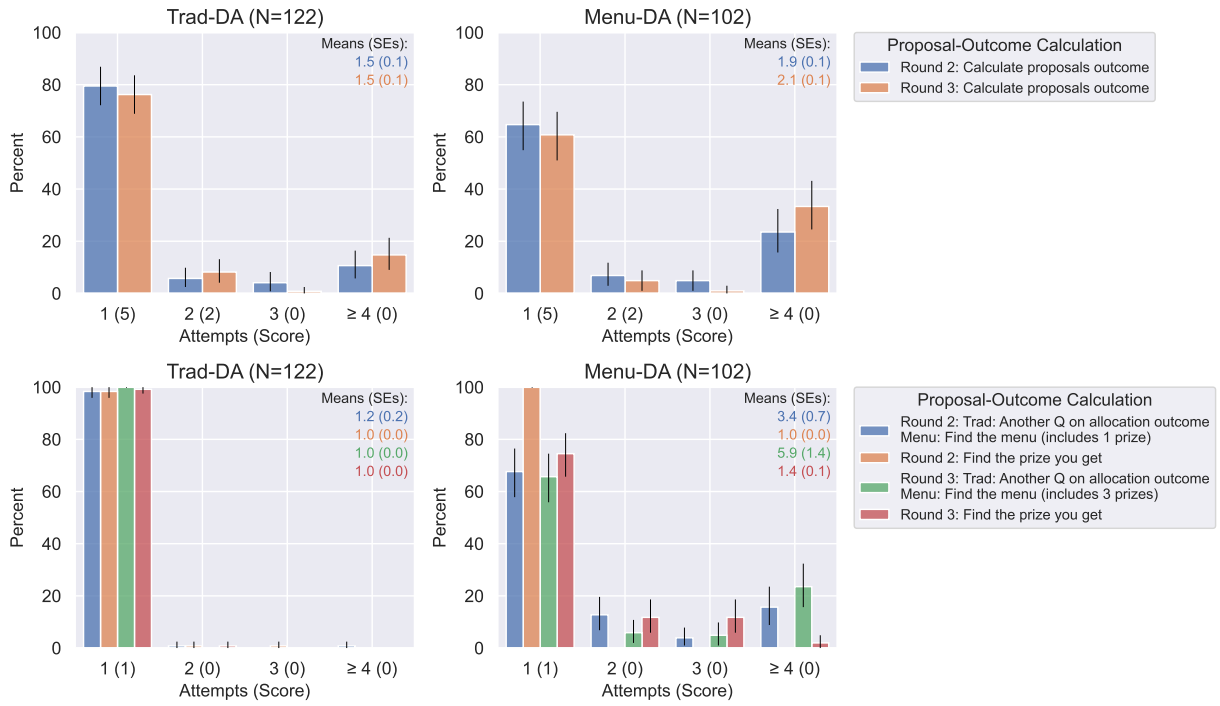
Notes: *Top row*: Performance at each of the four Null training round questions, by the number of attempts it took to answer the question correctly. Parentheses below amount of attempts show the score conditional on this amount. *Bottom row*: Overall score (replicating Figure A.2's bottom row in terms of % success). All panels include the mean values (SEs) in the upper right corner. *Both rows*: Error bars: 95% confidence intervals.

Figure B.6: DA Mechanics training questions: round 1



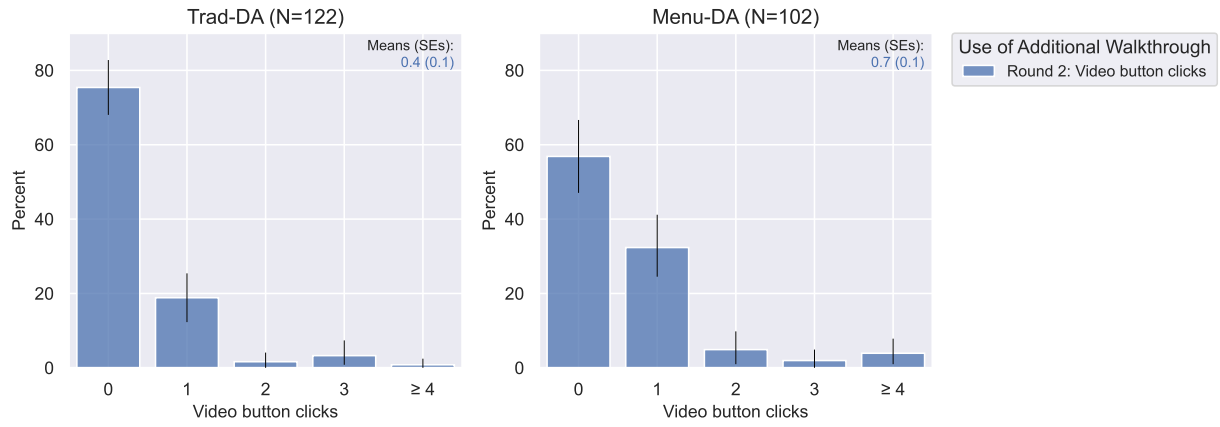
Notes: *Middle row*: ≥ 4 attempts in "Pair one agent," "Pair the other three," and "Solve conflict #i" (for $i = 1, 2, 3, 4, 5$) mean that the participant did not successfully calculate these proposals-outcome steps within three attempts, and was shown the correct answer by the computer in order to proceed. *All rows*: Error bars: 95% confidence intervals.

Figure B.7: DA Mechanics training questions: rounds 2–3



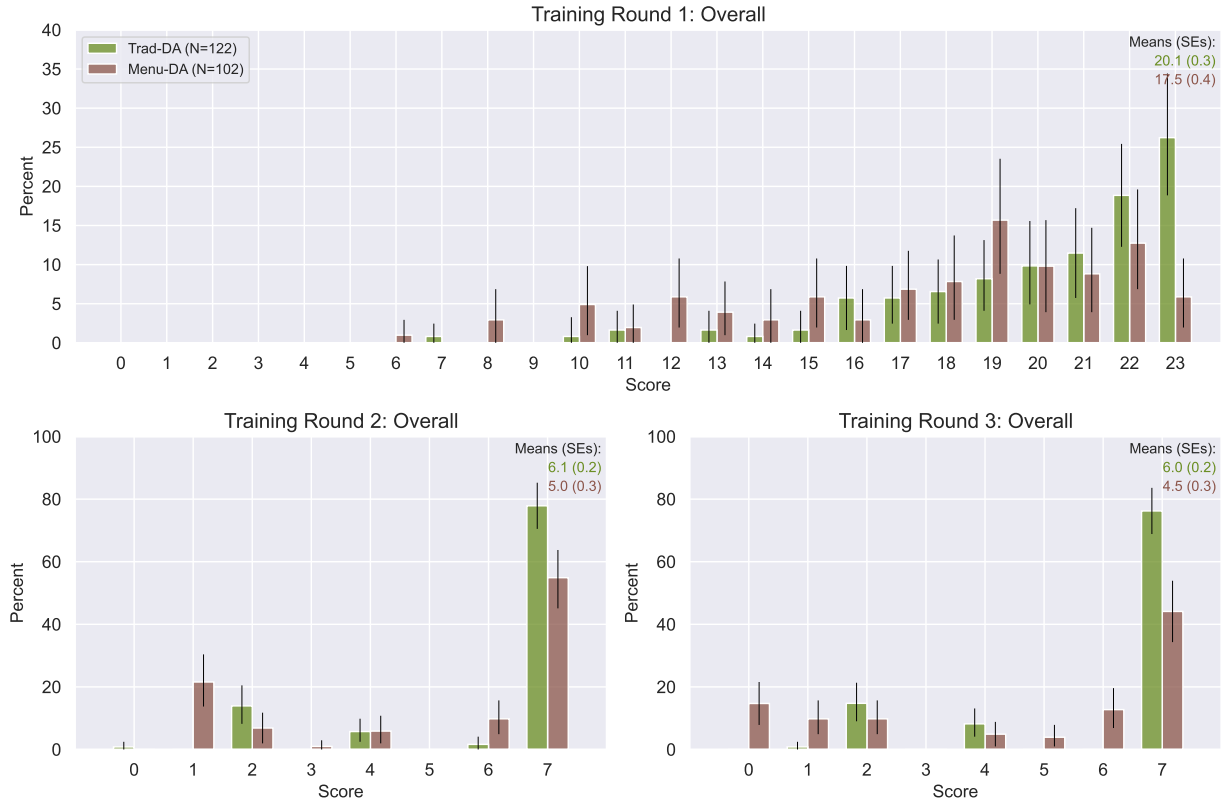
Notes: *Top row*: $\geq$ 4 attempts mean that the participant did not successfully calculate the proposals outcomes within three attempts, and was shown the correct answer by the computer in order to proceed. *Both rows*: Error bars: 95% confidence intervals.

Figure B.8: DA Mechanics walkthrough-video usage (round 2)



Notes: Error bars: 95% confidence intervals.

Figure B.9: DA Mechanics training scores

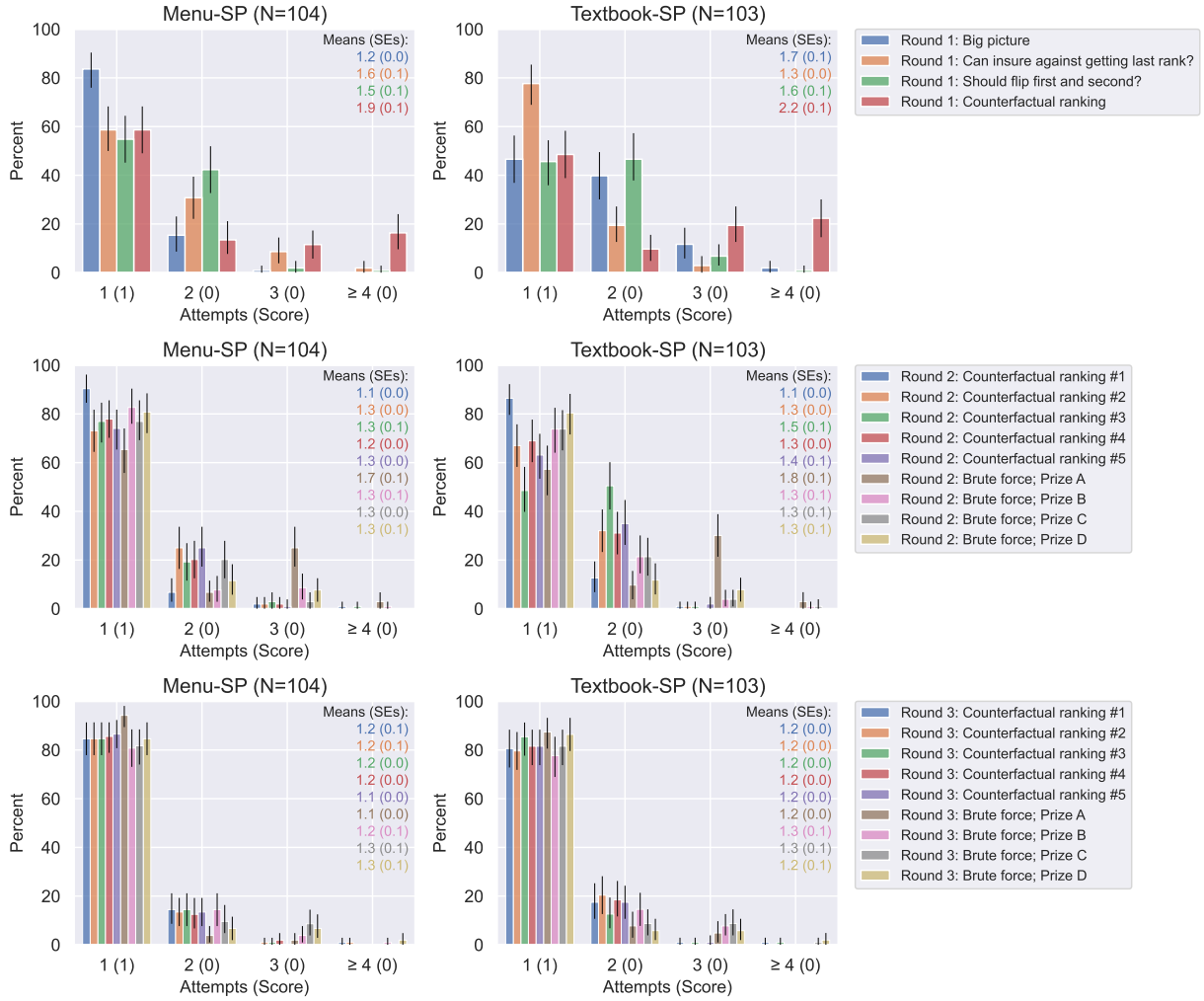


Notes: Error bars: 95% confidence intervals.

B.3.3 SP Property Training

Figures B.10 and B.11 show participants' performance in the the SP Property training questions.

Figure B.10: SP Property training questions

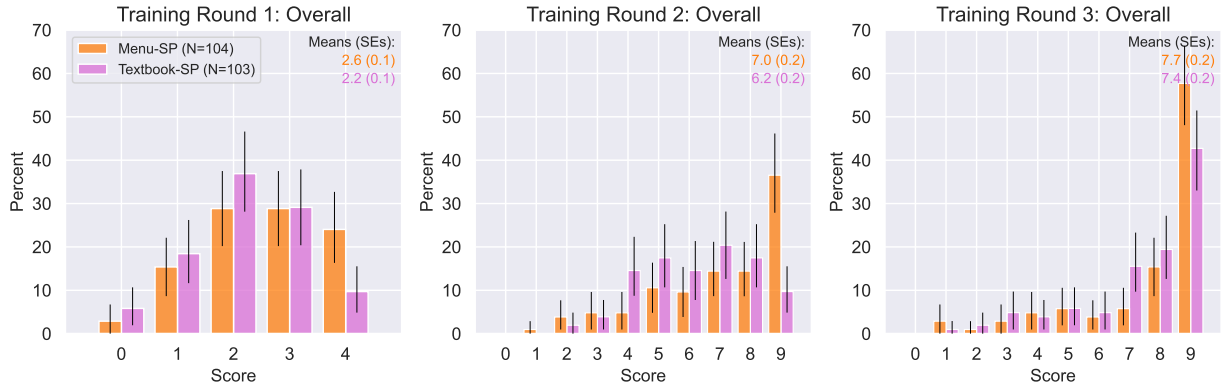


Notes: In "Counterfactual ranking" and "Brute force" questions, participants know which prize they received after submitting a particular ranking, and are asked whether they could have received other prizes by submitting other rankings (given that the other participants' rankings and prize priorities do not change). Counterfactual-ranking questions ask which prizes could be received by submitting a specific alternative ranking. Brute-force question ask whether there exists any alternative ranking whose submission could lead to receiving a specific prize. Error bars: 95% confidence intervals.

B.3.4 Null Treatment Training

Figure B.12 shows participants' performance in the Null treatment's training questions. These questions exactly repeated the Null training questions common to all participants

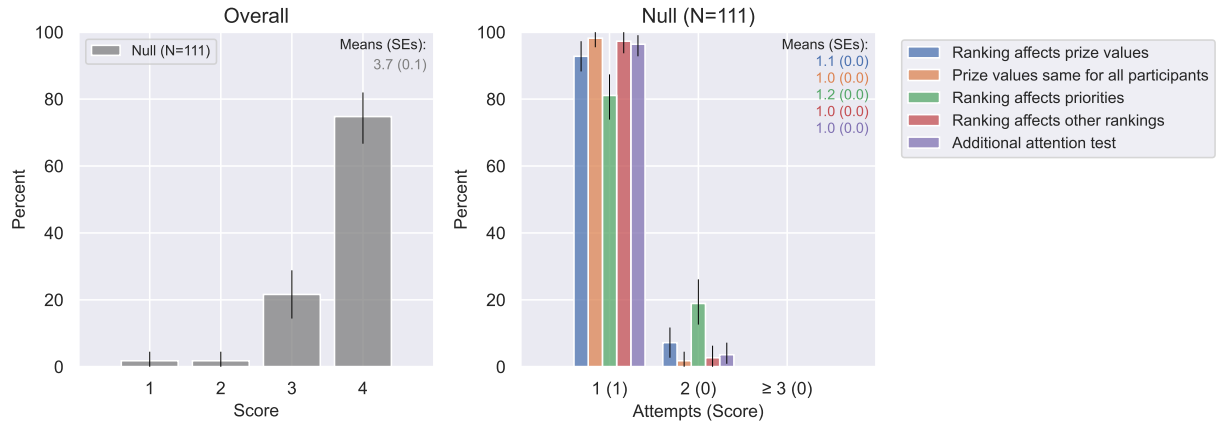
Figure B.11: SP Property training scores



Notes: Error bars: 95% confidence intervals.

(though, like the other treatments, the Null treatment's training score measure ignores the initial Null training, and instead measures only the repeated Null training).

Figure B.12: Null treatment training questions



Notes: Error bars: 95% confidence intervals.

B.4 Cognitive and Attention Scores

We present the details of our cognitive and attention scores.

First, as discussed in Section 2.6, the elicited cognitive score varies from 0 to 4. It uses the three questions of the Cognitive Reflection Task (Frederick, 2005), and measures numeracy using one question from the Berlin Numeracy Test (Cokely et al., 2012). The CRT questions are:

- (Bat-and-ball question) A bat and a ball cost $1.10 in total. The bat costs $1.00 more

than the ball. How much does the ball cost? Please enter an amount in cents. (Correct answer: 5).

- (Machines question) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? Please enter a number of minutes. (Correct answer: 5).

- (Lily-pads question) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? Please enter a number of days. (Correct answer: 47).

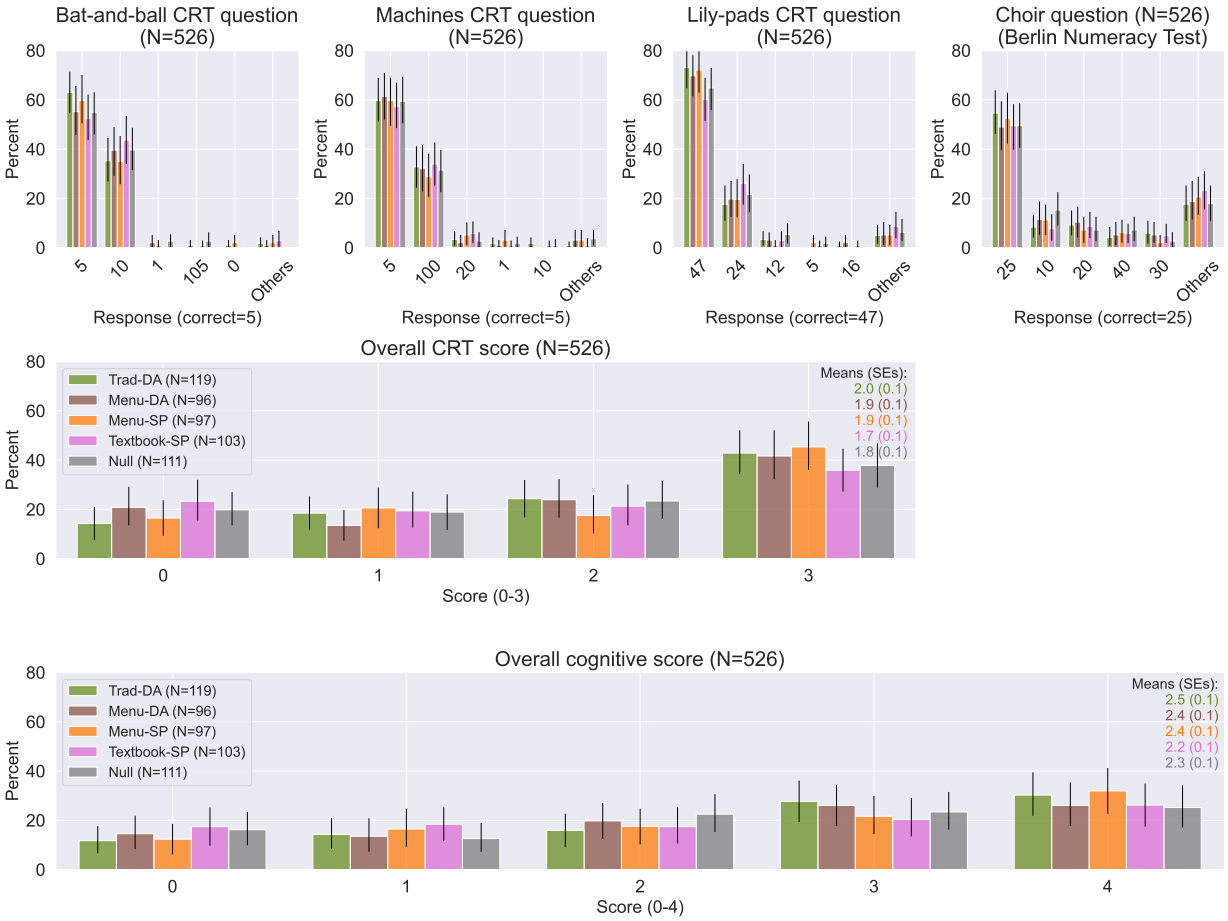
The Berlin Numeracy Test question is:

- (Choir question) Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please enter a percent chance between 0 and 100. (Correct answer: 25).

Figure B.13 shows the distribution of cognitive score, as well the distribution of responses to the individual underlying questions, across treatments. Due to a bug in the cognitive score elicitation interface in the first runs of the experiment, 16 observations from the Prolific sample lack this data and are omitted from this analysis, and the relevant sample size for this analysis is thus $N = 526$.

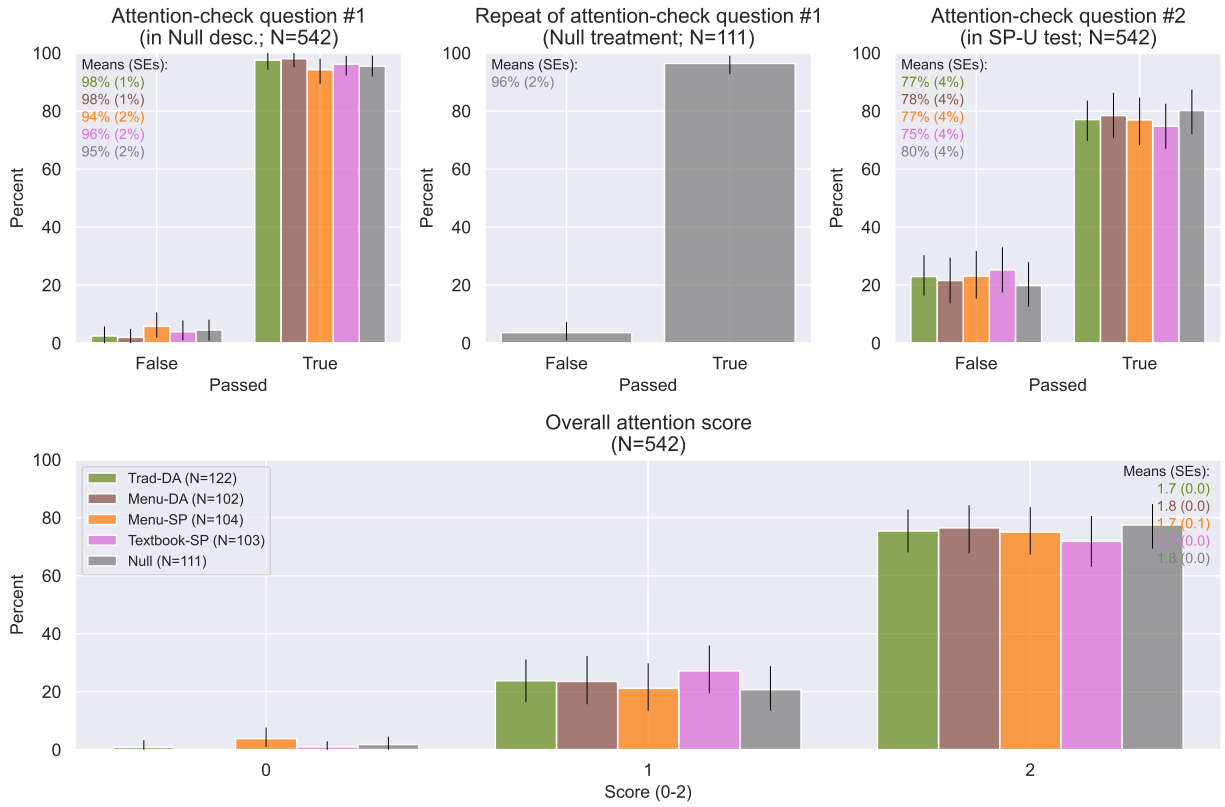
Second, the two attention-check questions were planted in the training questions following the Null description at the beginning of the experiment and in the middle of the SP understanding test, towards the end of the experiment. They include questions that explicitly instruct what response to submit, 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.14 shows the performance in the attention-check questions and the total attention scores. In the Null treatment, the first attention-check question is also repeated along the other questions from the Null training round common to all participants; this is shown in the figure but not included in the attention-score measure.

Figure B.13: Distribution of cognitive score and its underlying questions 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 D in the Supplementary Materials. Error bars: 95% confidence intervals.

Figure B.14: Distribution of attention score and its underlying questions across treatments



Notes: The panels describe the distribution of performance in the two attention-check questions which appeared during the experiment, and the success rate in an additional attention-check question given to Null treatment participants only, which is not included in the overall attention score. Error bars: 95% confidence intervals.

B.5 Additional Joint Distributions of Outcome Variables

B.5.1 Univariate Regressions and Correlations Between Main Outcome Variables

We investigate the relations between our main outcome variables, by reporting univariate regressions and correlations between them.

Table B.1 reports, separately for each treatment, three OLS regression coefficients 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 the three pairs. Focusing on the non-Null treatments, we make three observations which could serve as motivation for our other investigations.

Table B.1: Main outcome variables: regression and correlation coefficients

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)

Notes: "β": Coefficients from OLS regressions of dependent ("Dep.") on independent ("Indep.") variables, within each treatment separately. "r": Estimated Pearson regression coefficients. Robust standard errors in parentheses.

First, higher training scores are associated with higher SP understanding. This relationship is particularly strong for the SP Property treatments, which specifically aim to teach strategyproofness, but we also see a positive relation in the DA Mechanics treatments. We further investigate this relation in Appendix B.5.3 in the Supplementary Materials. 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. This gives additional motivation for our investigation into this relation in Section 3.4.2. Third, the correlations between % TR and % SF seem overall weaker than the other relations in the table. This finding is consistent with the idea that this relationship is essentially a noisy composition of those from first two rows of the table, i.e., that % TR

affects % SF mostly via % SP-U. We further investigate this relation in Appendix B.5.4 in the Supplementary Materials, and find additional support for this interpretation.

B.5.2 Relationship Between % SP-U and % SF: More Details

Following Table A.2 showing OLS regressions of % SF on Abstract and Practical, and total % SP-U, Table B.2 adds a breakdown of Abstract into its three sub-measures AbstractHard, AbstractMedium, and AbstractEasy.

Table B.2: Relation of all four % SP-U sub-measures with SF behavior

Dependent variable: % SF

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
% AbstractHard	0.15	−0.04							−0.01	−0.06		
	(0.07)	(0.08)							(0.07)	(0.08)		
% AbstractMedium			0.15	0.03					0.06	−0.02		
			(0.04)	(0.06)					(0.05)	(0.06)		
% AbstractEasy					0.17	0.11			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	0.32	0.22	0.45	0.22	0.45	0.13	0.37
N	542	542	542	542	542	542	542	542	542	542	542	542

Notes: See under Table A.2.

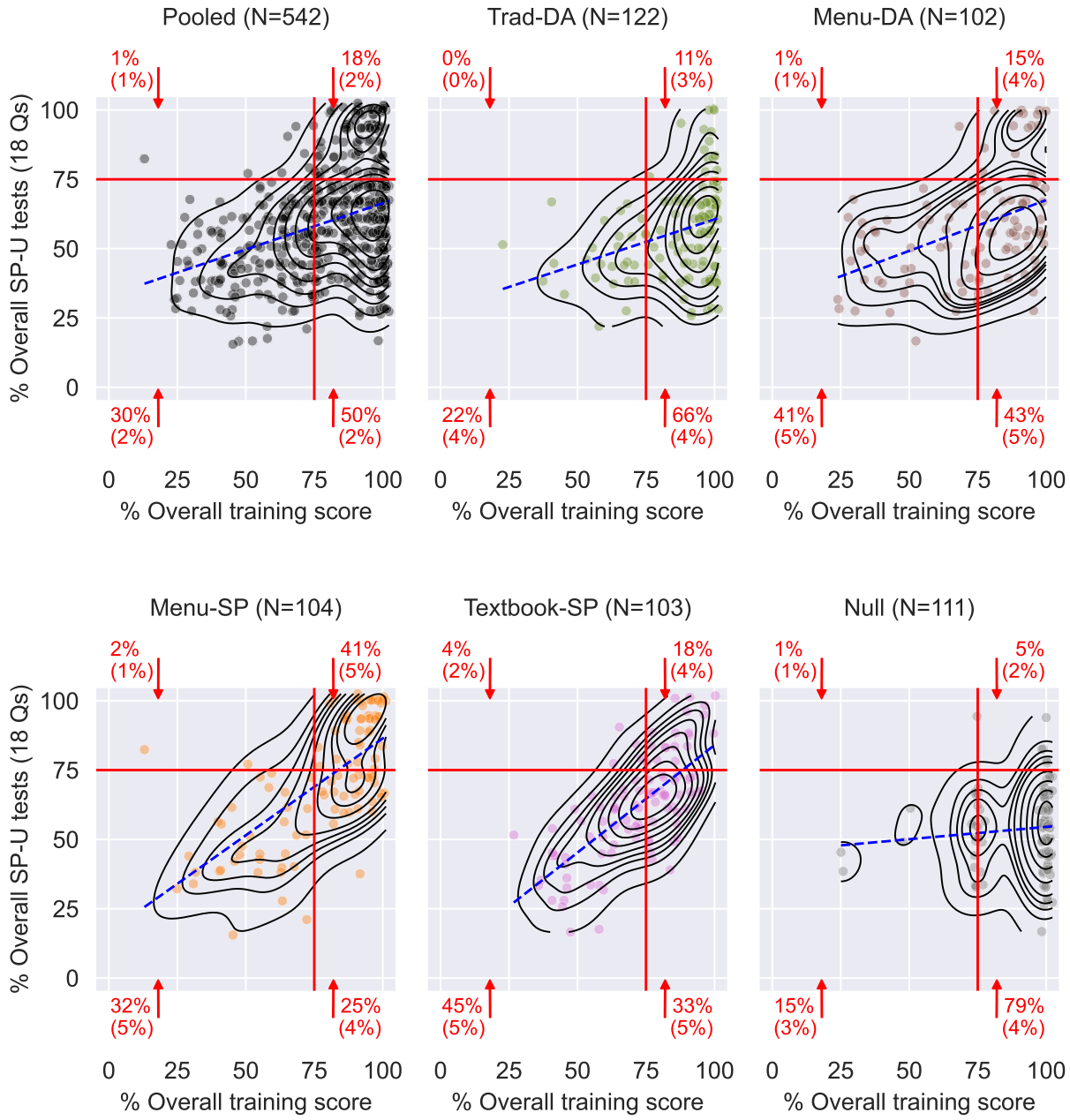
B.5.3 Relationship Between % TR and % SP-U

We investigate the relationship between training score and SP understanding in more detail.

Figure B.15 reports the joint distribution of % TR and % SP-U in all treatments. Note that, as discussed in Section 3.2, % TR is not comparable across treatments (though it is similar within DA Mechanics treatments and within SP Property treatments), i.e., the horizontal axis represents a different measure for each mini-figure in Figure B.15. However, the distribution among the pooled data is also added for completeness.

We first focus on our two DA Mechanics treatments. Despite the positive regression coefficient between % TR and % SP-U in Table B.1, most participants with high % TR do not reach particularly high levels of % SP-U. These findings support our second main result result, i.e., that understanding (the mechanics of) DA does not imply understanding strategyproofness. It provides within-treatment evidence (complementing our between-treatments

Figure B.15: Joint distribution of % TR and % SP-U by treatment



Note: See under Figure A.9.

evidence from Section 3.3) that even a high level of understanding of DA Mechanics does not imply understanding of strategyproofness.

To further illustrate the above—and to provide an additional comparison between Trad-DA and Menu-DA which accounts for the fact that the training questions are more challenging in Menu-DA—we consider the fraction of participants who have % SP-U above the threshold of 75% identified in Section 3.3.2 among certain upper percentiles of the per-treatment distribution of % TR. We find:

- Among participants in the top half of the % TR distribution (% TR $\geq$ 95% in Trad-DA, and $\geq$ 78% in Menu-DA; 52% of participants in each treatment), **19%** of participants in Trad-DA have % SP-U $\geq$ 75%; for Menu-DA, this fraction is **28%**.
- Among participants in the top quarter of the % TR distribution (% TR $\geq$ 97% in Trad-DA, and $\geq$ 92% in Menu-DA; 43% and 29% of participants, respectively), these fractions are **21%** for Trad-DA and **30%** for Menu-DA.
- Among participants with perfect training scores in Trad-DA and Menu-DA (24% and 4% of participants, respectively), these fractions are **24%** for Trad-DA and **25%** for Menu-DA.

The general similarity across Trad-DA and Menu-DA suggests that the two are fairly comparable in their ability to teach participants strategyproofness, even conditioned on (treatment-specific) high percentiles of understanding of the description.

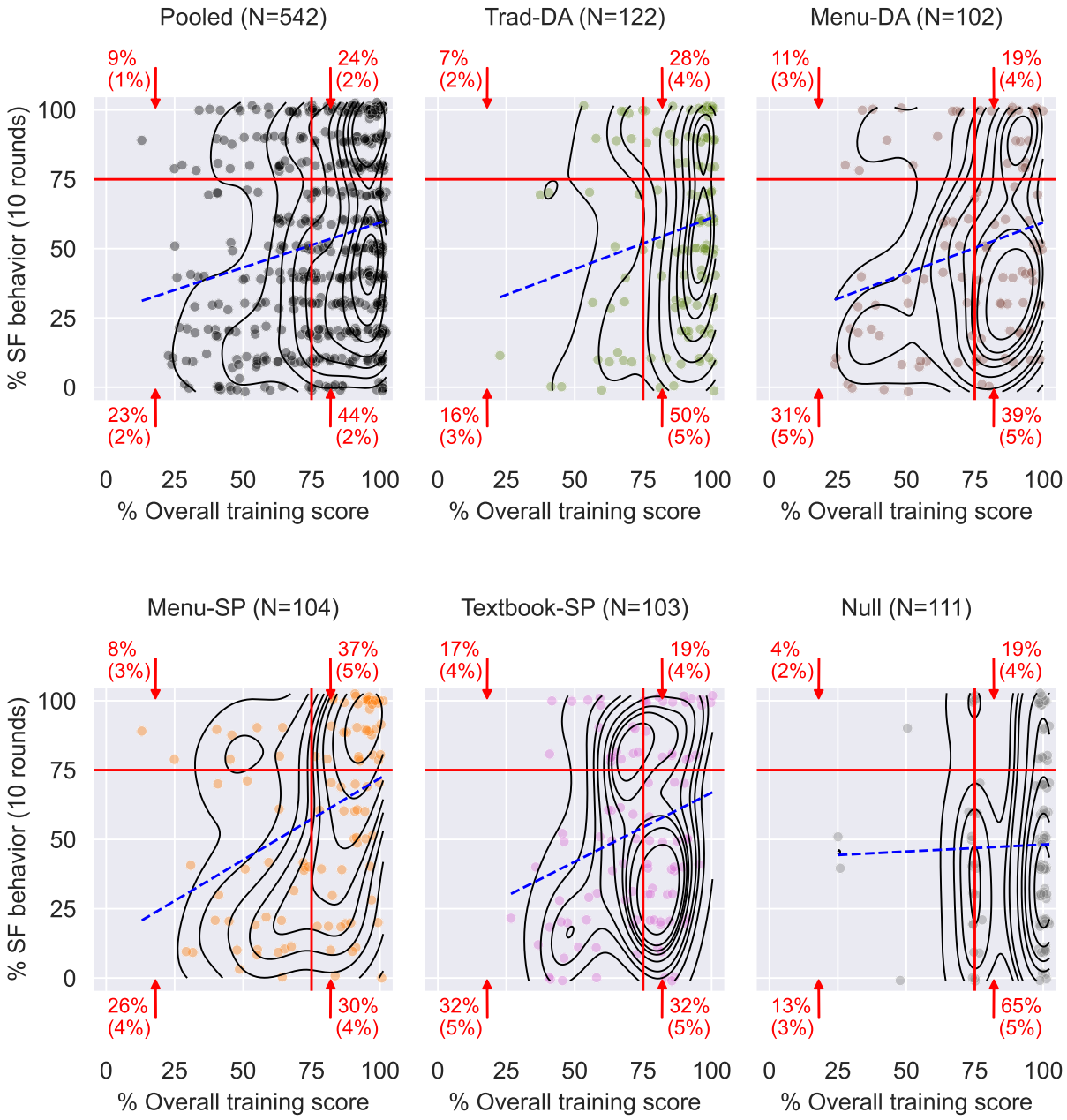
Next, we focus on our SP Property treatments, which specifically aim to teach participants strategyproofness. We find that the relationship between % TR and % SP-U is much stronger in these treatments than in the DA Mechanics treatments. This provides some empirical evidence that our SP Property treatments and our SP understanding test are aligned—they teach and test similar things.

B.5.4 Relationship Between % TR and % SF

We investigate the relationship between training score and SF behavior in more detail.

Figure B.16 shows the full joint distribution of % TR and % SF. The distribution among the pooled data is also added for completeness (however, recall again that training scores are not comparable outside the treatment groups of DA Mechanics, SP Property and Null). The figure, along with the small correlation coefficients in Table B.1, suggests that the relation between % TR and % SF is overall weaker than the relations between % TR and % SP-U and between % SP-U and % TR, and may mostly reflect a (noisy) composition of the two.

Figure B.16: Joint distributions of % TR and % SF by treatment



Note: See under Figure A.9.

C Experiment Materials

In this appendix, we provide experimental materials for easy reference, including screenshots of the main description texts, DA scenarios and parameter distributions used in incentivized DA rounds, and the full recruitment materials and procedures of running the experiment.

C.1 Description Screenshots

We provide screenshots of the per-treatment descriptions screens of our experiment. For full screenshots, see Appendix D in the Supplementary Materials. Note that in all the description screens, successive parts of the text did not appear at once, and instead appeared one-by-one as the participant advanced through reading.

Figures C.1 and C.2 show the description provided in the Traditional DA Mechanics treatment. Figures C.3 and C.4 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.

Figures C.5 and C.6 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.

Figures C.7 to C.9 show the full text of the Null treatment description. Note that this text is identical to the Null description, which is given in the initial screens of all treatments (excluding the small bit of text at the beginning noting that information will be repeated).

C.2 Scenarios in DA Mechanics Training Rounds

We describe the DA scenarios (i.e., the human and computerized participants' rankings and the prize priorities) used in the DA Mechanics training rounds. The same DA scenarios are presented in both Trad-DA and Menu-DA. See Figure C.10 for the scenarios used in the first and third training rounds (Figure 3 in Section 2.3 shows the second round).

These scenarios are chosen so that running each of these algorithms (whether participant- or prize-proposing) would be of comparable complexity. In the first round, correctly completing the training GUI requires 9 total proposals in both Trad-DA and Menu-DA; the menu consists of Prizes B and D. In the second round, the GUI requires 8 total proposals in both Trad-DA and Menu-DA; the menu consists of Prize A. In the third round, the GUI requires 10 total proposals in Trad-DA and 8 in Menu-DA; the menu consists of Prizes A, C, and D.

Figure C.1: Traditional DA Mechanics description main text (1/2)

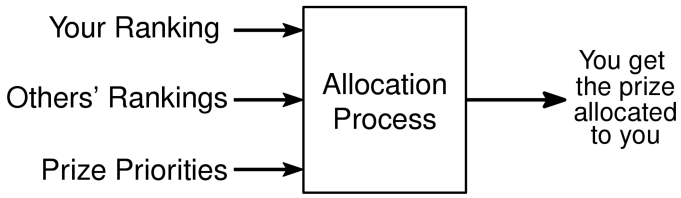
How the Allocation Process Works

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

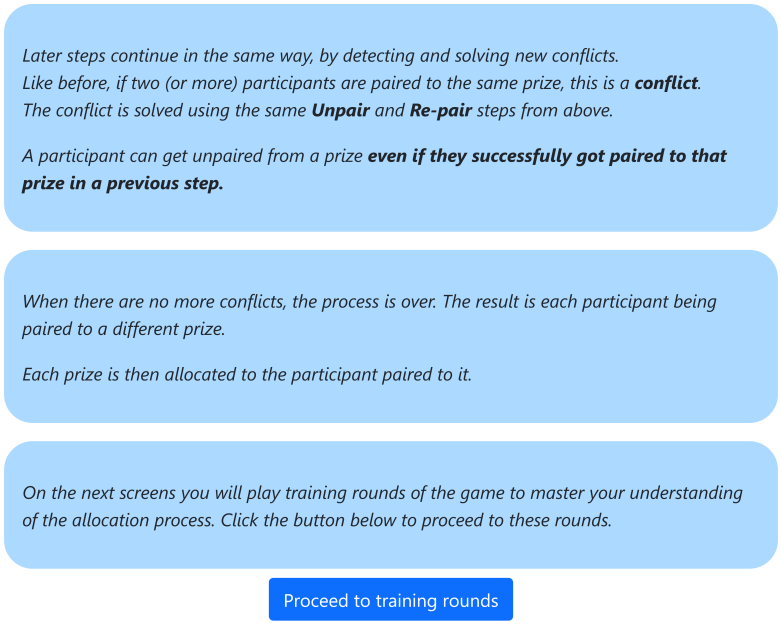
The allocation process is a multi-step process, as follows (it may look complicated, but don't worry, we will rehearse this in a moment):

*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.2: Traditional DA Mechanics description main text (2/2)



C.3 Randomization of Setting Components in Real DA Rounds

We describe the underlying randomization procedures which determine the three setting components of real DA rounds that participants play. In each round, these components include (1) the monetary values of the four prizes A, B, C and D for the human participant, (2) each participant's priorities for getting the different prizes (the "preferences of prizes" over participants), and (3) the computerized participants' submitted rankings (unobserved by the human participant).

These three components are randomly drawn from a joint distribution each round, where each (human) participant $\times$ round draw is independent from others. Each round randomization is performed 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.

Figure C.3: Menu DA Mechanics description main text (1/2)

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.

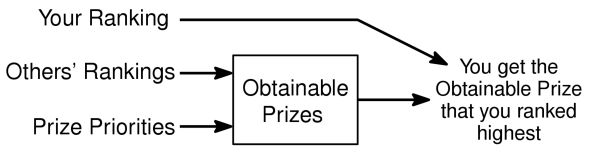
Overview of allocation process

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.4: Menu DA Mechanics description main text (2/2)

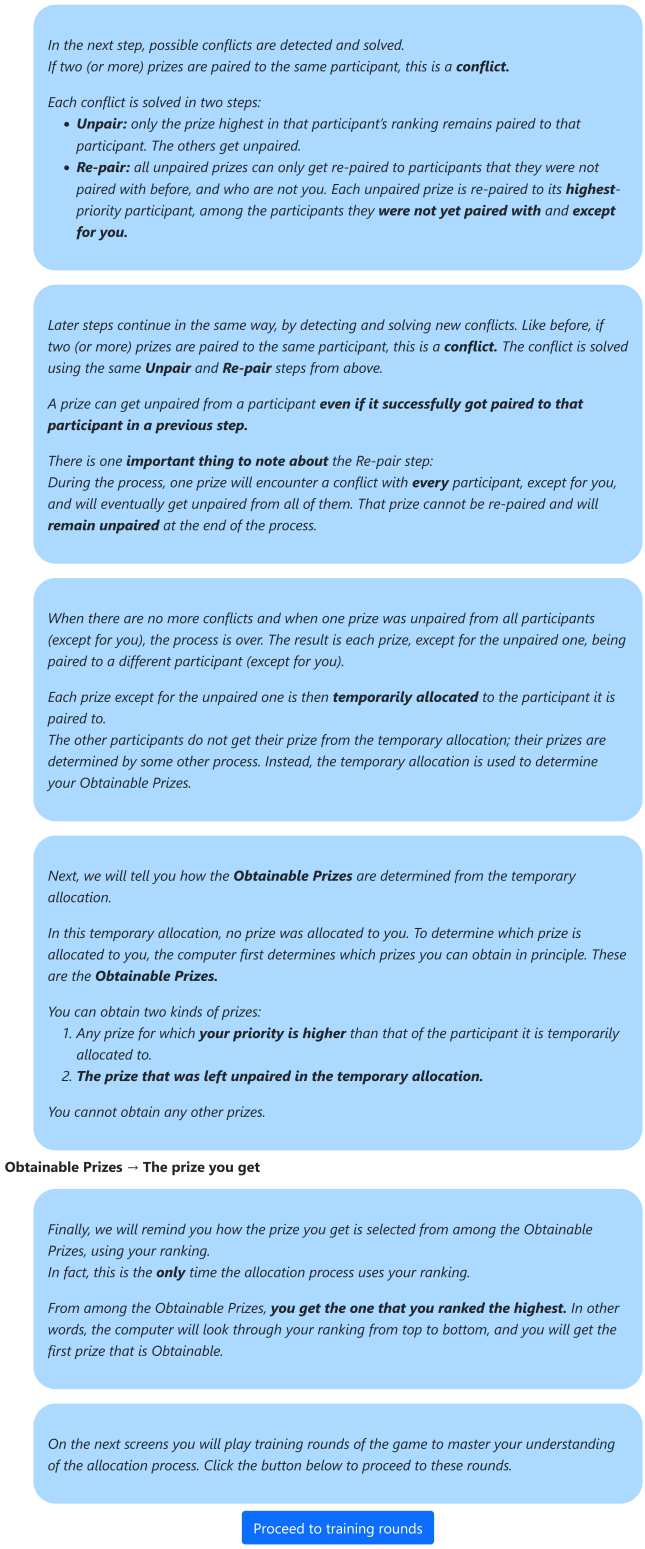


Figure C.5: Menu SP Property description main text

The Key Principle of Your Ranking

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

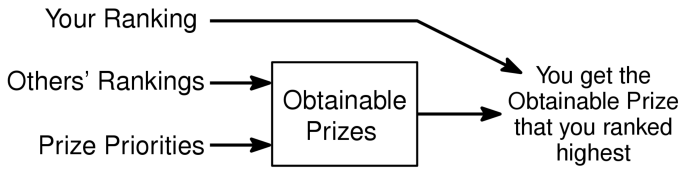
This principle is important to learn: *You may be able to apply your knowledge of it to choose your rankings in the upcoming real rounds of this study.*

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.6: Textbook SP Property description main text

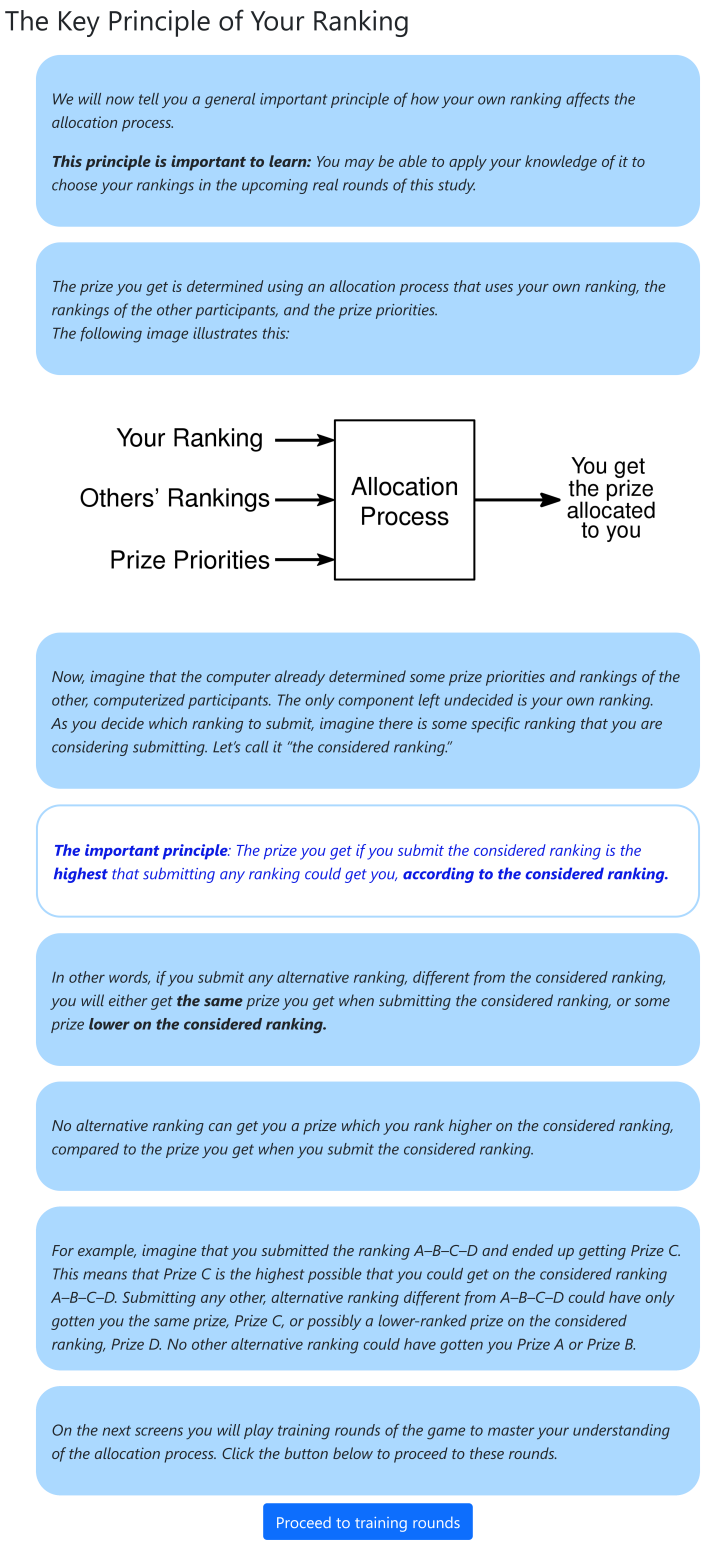


Figure C.7: Null treatment description main text (1/3)

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.8: Null treatment description main text (2/3)

*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.

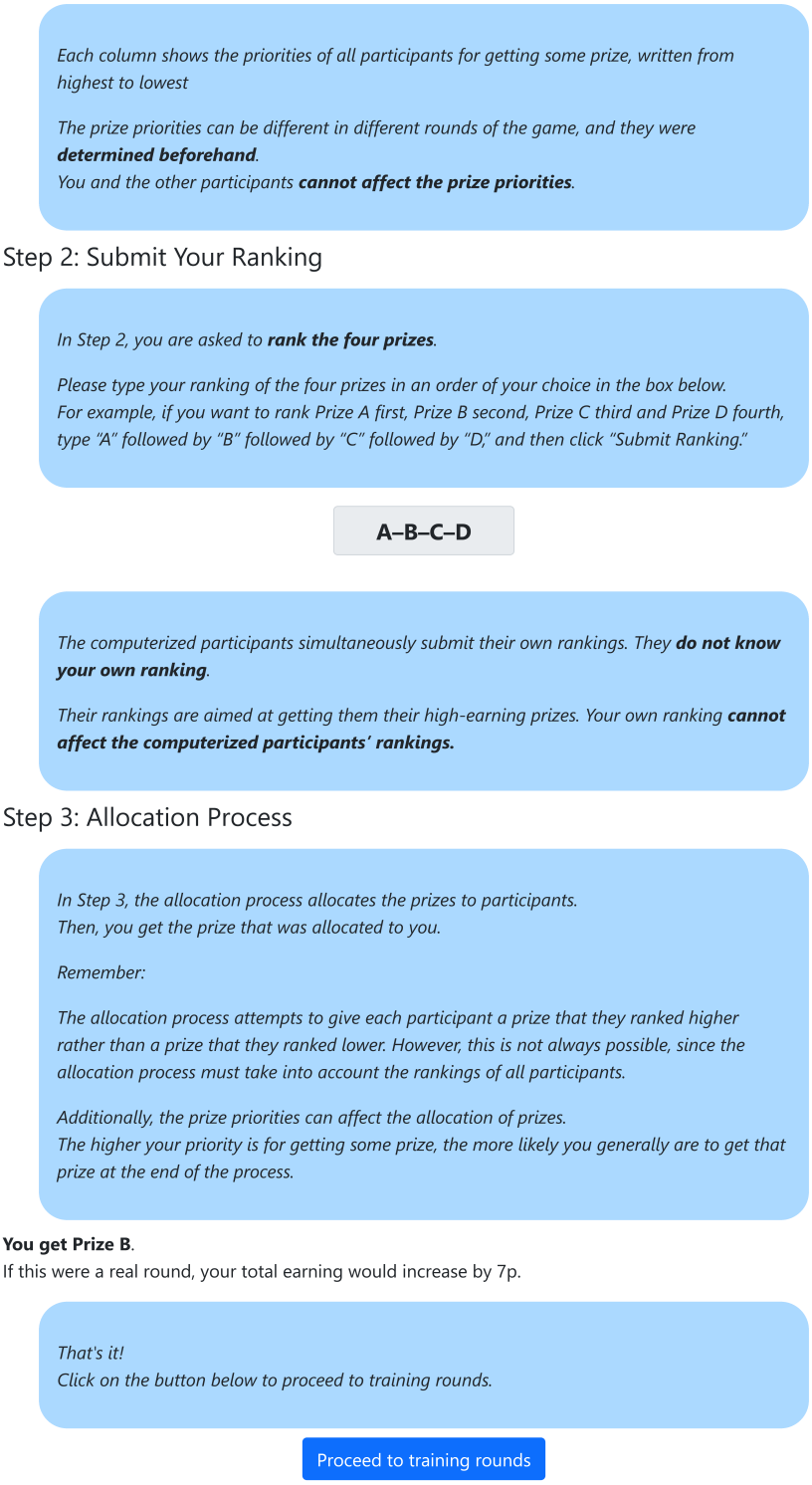
In this round, the **prize priorities** for you and for the other participants are:

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.9: Null treatment description main text (3/3)



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**.*

Step 2: Submit Your Ranking

*In Step 2, you are asked to **rank the four prizes**.*

Please type your ranking of the four prizes in an order of your choice in the box below.
For example, if you want to rank Prize A first, Prize B second, Prize C third and Prize D fourth, type "A" followed by "B" followed by "C" followed by "D," and then click "Submit Ranking."

A–B–C–D

*The computerized participants simultaneously submit their own rankings. They **do not know your own ranking**.*

*Their rankings are aimed at getting them their high-earning prizes. Your own ranking **cannot affect the computerized participants' rankings.***

Step 3: Allocation Process

In Step 3, the allocation process allocates the prizes to participants.
Then, you get the prize that was allocated to you.

Remember:

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.

Additionally, the prize 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.

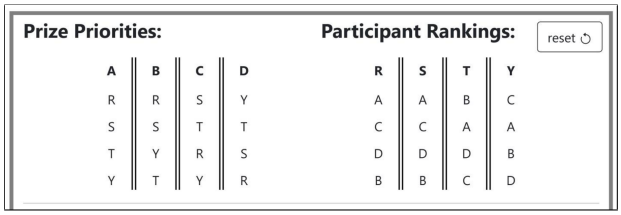
You get Prize B.
If this were a real round, your total earning would increase by 7p.

That's it!
Click on the button below to proceed to training rounds.

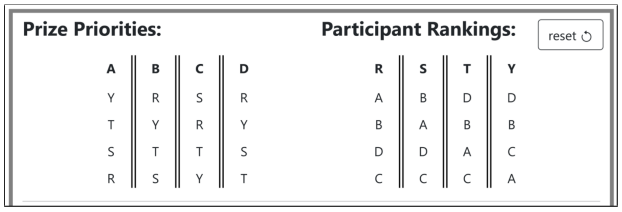
Proceed to training rounds

Figure C.10: DA Mechanics training scenarios

(a) First training round



(b) Third training round



Notes: Screenshots are taken from the Trad-DA treatment. See Figure 3 for the scenarios used in the second training round.

- All the above numeric values are determined the same in Prolific and Cornell, but the currency attached to them is British pounds (£) in Prolific and US dollars ($) at Cornell.

2. For each prize, the four participants' priorities for getting it are sequentially determined from highest to lowest, by iteratively selecting a participant without replacement in four steps. In more detail:

   - For all prizes except the highest-valued one from (1), each participant has an equal chance of having each of the four priorities, i.e., of being chosen at each step.
   - For the highest-valued prize, the human participant's chances are different from the computerized participants': at each step, each computerized participant has an r_1-times higher chance of being picked to that step's 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-valued 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 submitted ranking is sequentially determined from top to bottom rank, by iteratively selecting a prize without replacement, where the chances to select among the different prizes at each step is a function of the order of prize values for the human participant from (1). In more detail:

   - First, the participant's first (highest) rank is determined. The chance to select each prize at this rank is proportional to r_2^{P-1}, where r_2 is a fixed parameter (see

below) and P is the ordinal position of that prize when prizes are sorted from highest to lowest values for the human participant (i.e., $P = 1$ for the highest-valued prize for the human participant and $P = 4$ for the lowest-valued one for the human participant).

- Then, that participant's second rank is similarly determined, after sorting the remaining three prizes from highest to lowest, and so on until the fourth rank.
- Due to the dependence of the ranking on the prize values for the human participant, the r_2 parameter effectively sets a type of correlation between the computerized participants' ranking and the human's participant straightforward (SF) ranking. We use a parameter value of $r_2 = 0.5$. (One can show that $r_2 \to 0$ would generate a perfect correlation of all rankings with the human participant's SF ranking, and $r_2 = 1$ would generate zero correlation and uniformly random rankings for the computrized participants.)

The joint distribution of the three setting components following the above procedure is designed to achieve two main 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 salient non-straightforward ranking patterns observed in previous studies (e.g., Li 2017, Dreyfuss, Heffetz, and Rabin 2022, Dreyfuss, Glicksohn, et al. 2022). Intuitively, it does so by making the setting feel "hard" and competitive, in order to encourage participants to think more carefully on their strategy. This is done by (1) giving participants a lower probability of having first priority for getting the highest-earning prize than to the computerized participants (recall that all these priorities are shown in a table each round), and (2) correlating the computerized participants' rankings with the participant's straightforward ranking.[24]

Second, the distribution is designed to often induce large differences between prize values, such that conditional on NSF behavior being costly ex-post, i.e., earning-decreasing, the cost is 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 ex-post. However, participants are neither informed on this nor can easily calculate this, as they do not know the computerized participants' rankings or their distribution. This common NSF ranking pattern is only costly when ranking the highest-earning prize first would win it,

[24]Using a simpler, uniform distribution over the computerized participants' rankings instead results in an "easy" setting on average, with at least a 0.25 probability of getting the highest-earning prize whenever the participant ranks it highest, regardless of how the priorities for getting the prizes are distributed (since one can show that one uniformly random prize will go unmatched when considering DA run on only the computerized participants).

and making such conditions frequent would interfere with our goal of creating a competitive setting, where it is unlikely to win the highest-earning prize. Evidence from Dreyfuss, Glicksohn, et al. (2022)—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 of ranking patterns, we opt for inducing high rates of non-costly NSF behavior over inducing lower rates of costly NSF behavior.

C.4 Experiment Procedures

We describe the full protocol used to run the experiment on Prolific and at the Cornell Business Simulation Lab (BSL), and provide more details on the sessions in each of the platforms.

C.4.1 Prolific

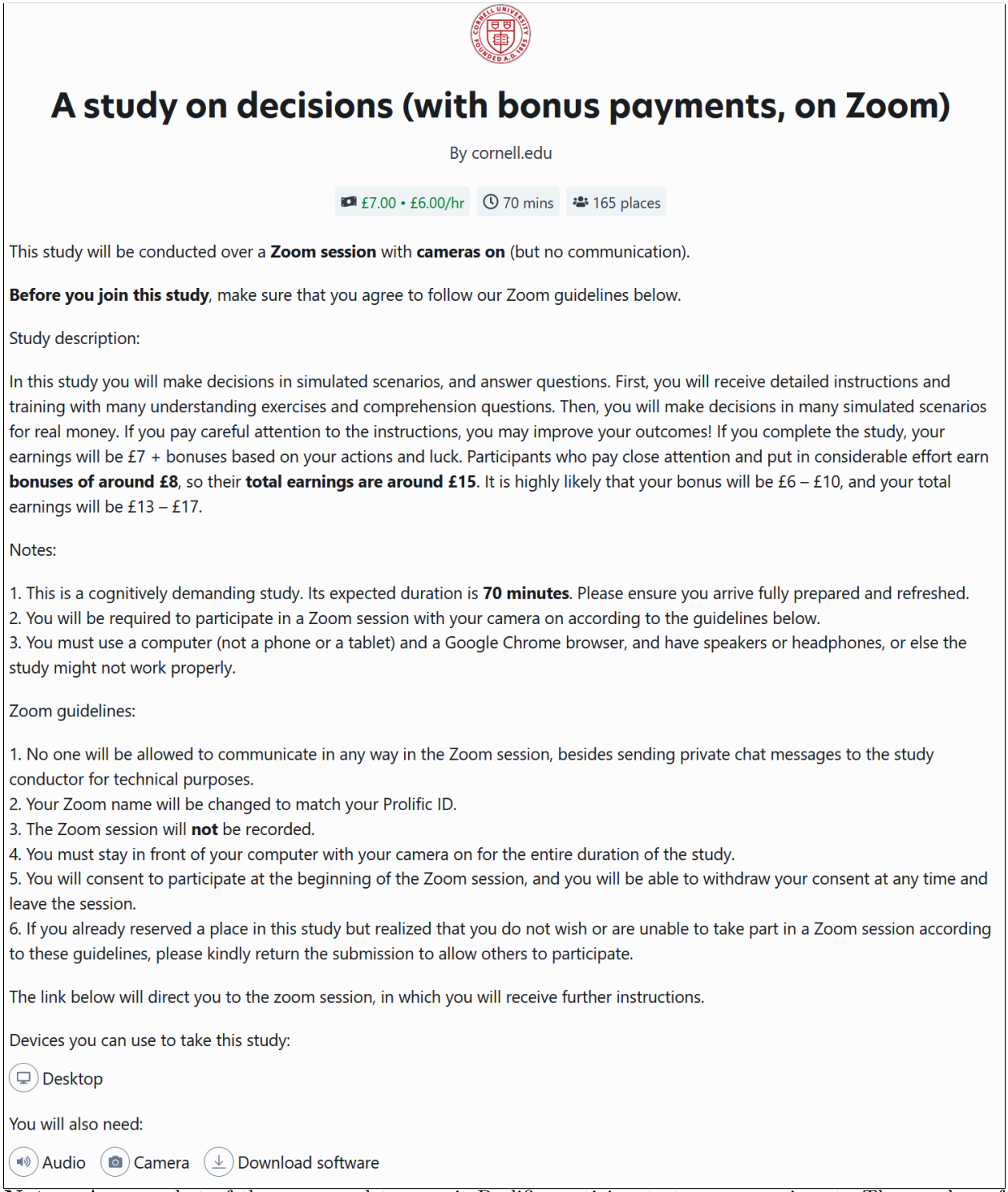
Protocol. Participants were recruited using the page shown in Figure C.11. Only Prolific participants who passed the pre-screening criteria of living in the US and having at least 99% past approval rate and at least 50 past completed tasks were able to see the page. Participants who wished to participate 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 computer, 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, and (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).

Figure C.11: 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 slots mentioned at the top ("165 places") is session-specific.

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 re-joining the room using their link.

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.

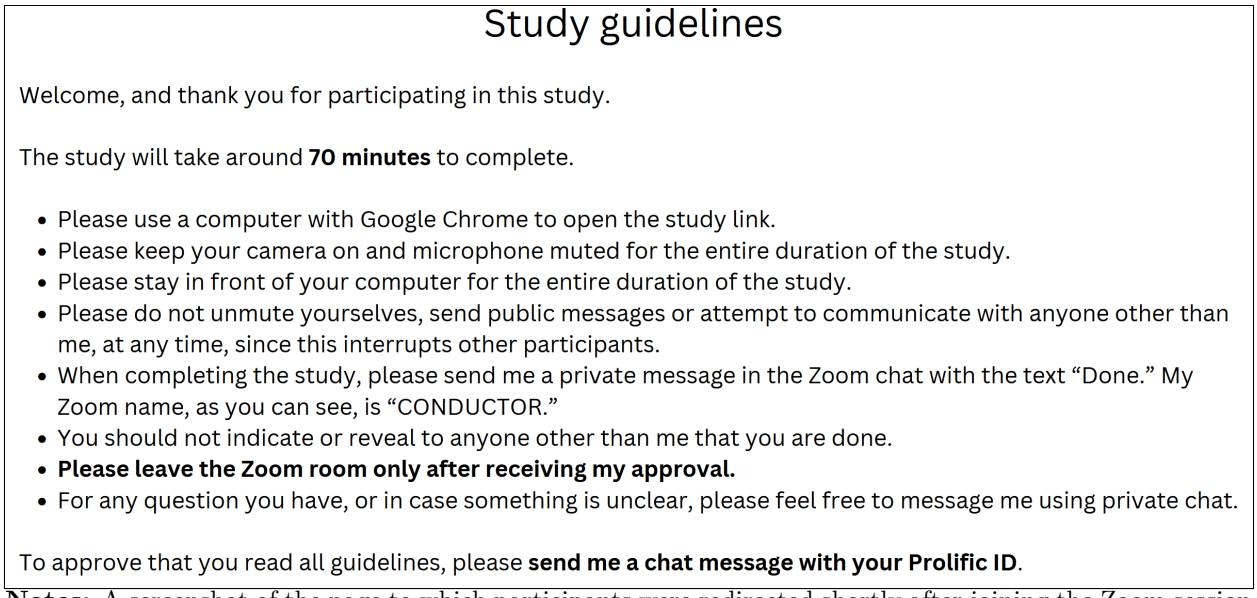
The Zoom session was opened shortly prior to posting the recruitment message on Prolific. The session conductor was named "CONDUCTOR," and was required to join the session using another device named "BACKUP," in case the main device disconnects.

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 timing 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 the conductor noted that the participant's camera is already on, the message read "Dear participant, please enter...").

The participant then saw the slide shown in Figure C.12. 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 then received the link to the experiment using 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 started the experiment, they were renamed 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.

Figure C.12: 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.

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' completion was approved in the Prolific platform and they were paid according to their responses in the experiment, typically within a day.

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

The August 3 session was a final pilot including 43 participants. We opted (and preregistered) to include it in our data, since at that point the experiment no longer changed except for two small bugs which were fixed toward the final two sessions.

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

C.4.2 Cornell Business Simulation Lab (BSL)

Protocol. Participants were recruited using the page shown in Figure C.13. Only Cornell 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.

Figure C.13: 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.

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 during some time window, BSL sessions required all participants to start at the same time. Sessions at the 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 complete the experiment before another session began, they could stay at the lab and complete it 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 at 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 at the lab for at least 50 minutes, to avoid potential effects of participants randomized into shorter treatments leaving before participants who were randomized into 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 by the lab within a few days.

Sessions. 295 Prolific participants participated over 32 days, each including between 1 and 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.

On October 11, 2023, one session including 16 participants was accidentally 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 Trad-DA participants which were collected in the other sessions. In our robustness analysis we control for session-date fixed effects and do not find them to change our results.

D Full Screenshots

For Appendix D, which contains screenshots of every screen of every treatment, see the supplementary material on the authors' websites and in the ancillary files on arXiv. For screenshots of the per-treatment description screens, see Appendix C.1 in the Supplementary Materials.