Perceived Fairness in Envy-Free Allocations of Indivisible Goods

HADI HOSSEINI, Pennsylvania State University, USA
JOSHUA KAVNER, Rensselaer Polytechnic Institute, USA
SUJOY SIKDAR, Binhamton University, USA
ROHIT VAISH, Indian Institute of Technology Delhi, India
LIRONG XIA, Rensselaer Polytechnic Institute, USA

The problem of fairly allocating indivisible resources among agents with differing preferences has been extensively studied in recent years, giving rise to several fairness concepts. While the theoretical aspects of such concepts have been thoroughly investigated, their perceived fairness by stakeholders is not well understood. In this work, we examine the perceived fairness of two relaxations of envy-freeness: *envy-freeness up to one good* (EF1), a counterfactual notion, and *envy-freeness up to k hidden goods* (HEF-k), an epistemic notion. Our main contribution is a framework for conducting human experiments to evaluate human perceptions of different notions of fairness. Through extensive crowdsourced experiments, we demonstrate that HEF-k allocations are perceived to be fairer than two popular variants of EF1.

1 INTRODUCTION

Dividing resources fairly among agents with subjective preferences is a significant aim in economics and computer science research. Despite significant interest in the theoretical and algorithmic aspects of fairness (see e.g., [Amanatidis et al., 2022, Aziz et al., 2022, Moulin, 2019]), little has been done to investigate their practical suitability among human subjects. The perception of fairness has been recently studied in the context of loan decisions [Saxena et al., 2019] and within machine learning [Srivastava et al., 2019]. Yet, in the context of fair allocation of resources, it is unclear which fairness concepts are perceived to be fairer.

Our task is to fairly partition a set of indivisible goods and assign bundles to the agents. One of the gold standards of fairness, *envy-freeness* (EF), requires that each agent weakly prefers her own bundle to those of all others according to her own subjective valuation [Foley, 1967]. EF is a natural and compelling solution concept because it does not rely on interpersonal utility comparisons and eliminates the need for identifying which agent derives the most benefit from a bundle of resources.[1] However, EF allocations of indivisible goods do not always exist and determining their existence is computationally intractable [Lipton et al., 2004]. In light of the impossibility of guaranteeing EF exactly, several relaxation of EF have been proposed. This motivates the following research question:

> *Which relaxation of envy-freeness for allocating indivisible goods is perceived to be fairer by individuals?*

In this paper, we compare two prominent variants of EF. First, *envy-freeness up to one good* (EF1) is based on the *counterfactual* thinking that any pairwise envy can be eliminated by the hypothetical removal of a single good from the envied agent's bundle [Budish, 2011]. An EF1 allocation always exists and can be computed in polynomial time [Lipton et al., 2004]. Second, *envy-freeness up to k hidden goods* (HEF-k) assumes agents have common information about an allocation except for a small subset of k goods which are hidden [Hosseini et al., 2020]. HEF-k strikes a balance between the counterfactual removal of goods, as in EF1, and *epistemic envy-freeness*, which assumes agents have different information about how goods are distributed [Aziz et al., 2018].

We seek to investigate the perceived fairness of these relaxations of EF through an experiment with human subjects. We concentrate on perceptions of envy, the specific aspect of fairness that

[1] See e.g., [Harsanyi, 1990, Sen, 2018] for a comprehensive discussion on interpersonal utility comparisons.

Paper presented at the New Directions in Social Choice workshop held at EC'24, July 8, 2024, New Haven, CT, USA.

underlies envy-based fairness notions, and estimate how likely it is for human subjects to experience envy given allocations that satisfy different relaxations of EF. At a high level, our work is aligned with a large body of work in distributive justice concerning socially fair outcomes (in contrast to procedural justice, which concerns fair processes for determining outcomes). [2]

1.1 Our Contributions

We develop a novel framework, data set, and measure of perceived fairness of allocations of indivisible goods through an empirical study. In our empirical study, each participant is presented with a series of scenarios in which they take on the perspective of an agent in a fair division instance with an initial allocation that satisfies one of three fairness criteria: sEF1, EF1, or HEF-k (defined in Section 2). Participants may either keep their given bundle or *swap* it with the bundle of an agent of their choice. Our approach measures perceptions of fairness by evaluating whether an agent is envious of another's bundle. If an agent is envious, then it is likely the agent would be willing to swap her bundle should she get the opportunity to do so. This indicates the individual's perceived envy (but not the degree of envy). Participants' responses are aggregated into a single *swap rate*, the percentage of scenarios where a swap was chosen, measuring aggregate perceived fairness under each treatment.

Our results show that HEF-k allocations are perceived to be fairer than in the sEF1 and EF1 treatments. In particular, we show that there is a statistically significant difference between swap rates of HEF-k and both sEF1 and EF1 treatments ($p < 0.001$). Participants under the HEF-k treatment displayed the lowest swap rate, followed by sEF1 and then EF1 (Section 4.1). We subsequently control for the effect of variables such as instance size, allocation balance (defined in Section 3.1), and scenarios for which it is optimal to swap, and find that the qualitative results still hold.

Additionally, we study *cognitive effort*, as measured by response time and self-reports of scenario difficulty, to understand how treatment affects participants' reasoning and how it correlates with perceived fairness. We find that there is a significant difference in the cognitive effort exercised by participants, measured by response time and self-reports of difficulty, between the HEF-k and both sEF1 and EF1 treatments ($p < 0.001$) (Section 4.2). Hence, perceived fairness appears correlated with the cost of increased task complexity.

1.2 Related Work

Our work is in line with the broader research agenda on empirically testing the relevance of different fairness notions and theories of distributional preferences. While it is evident people trade off self-interest for fairness [Kahneman et al., 1986], it is still not clear to what extent and which theories of fairness are the most valid. Prior experiments have employed several methods to evaluate perceived fairness of allocations, often asking participants which they prefer. For instance, Herreiner and Puppe [2009] empirically investigated EF in a free-form bargaining experiment. In their setting, participants had subjective preferences over goods and collaborated with another to choose the allocation (see also [Herreiner and Puppe, 2010]). The authors subsequently analyzed the fairness criterion satisfied by and efficiency of the chosen allocations. This work is most similar to ours, except that we measure the envy experienced by participants and focus on the *relative* fairness of relaxations of EF.

Herreiner and Puppe [2009]'s work follows a tradition of questionnaire methodology for evaluating distributive justice, popularized by Yaari and Bar-Hillel [1984] and Konow [2003], who asked whether participants perceive given allocations as just or not (see also [Gaertner, 2009, Chapter 9]). Similarly, Herreiner and Puppe [2007] asked participants to choose which of a set of allocations was the most fair. While these studies provide some evidence in favor of certain fairness notions, payoffs were

[2] We refer the reader to the extensive literature in social justice theory, e.g., [Adams, 1963, Rawls, 2004]. See [Lee et al., 2019, Rawls, 2004, Tyler and Allan Lind, 2002] on procedural justice.

identical, so intrapersonal theories like EF could not be tested. In this vein, Engelmann and Strobel [2004] ran an experiment were participants would, with some probability, received the allocation of money they chose. Their aim was to compare the explanatory power of distributional preferences models by Fehr and Schmidt [1999], Bolton and Ockenfels [2000], and Charness and Rabin [2002]. [3]

A separate line of work by Lee and colleagues focused on perceived fairness of algorithmic decision-making. Participants in Lee and Baykal [2017]'s study perceived allocations prescribed by Spliddit[4] to be less fair than those chosen in group discussions one third of the time. The authors explain this distinction as the algorithms excluding the effects of individual participation, interpersonal power, and altruism on fairness. Lee [2018] suggested that perceived fairness depends on task characteristic. Their participants recognized that algorithms produce less fair decisions on tasks requiring human skills, such as those requiring subjective judgement, but equally fair on mechanical tasks, such as processing data. Lee et al. [2019] measured the effect of transparency and outcome control (i.e., the ability to manually adjust prescribed outcomes) on perceived fairness of EF1 allocations prescribed by Spliddit. They showed that perceived fairness increased after participants were given an opportunity to modify the allocation, either individually or through group discussions. These studies substantially differ from ours in that (1) there is an impact of 'personal image' and 'social pressure' in bargaining and collective decision-making, which may provide a justification for inequality aversion, and (2) there is a sense of 'agency' within discussions or ability to modify the outcome, which may result in higher satisfaction via the *IKEA Effect*.[5]

Other empirical research includes Kyropoulou et al. [2022], who tested the effect of participants' strategic behavior in choosing allocations of divisible resources on total envy. Separately, König et al. [2019] measured the suitability of two well-adopted matching mechanisms, the Boston mechanism and assortative matching, under the *veil of ignorance* [Rawls, 2004] assumption. They concluded that which procedure participants prefer depends on how much autonomy they have to report their preferences. The empirical validity of fairness axioms in cooperative games [De Clippel and Rozen, 2022, d'Eon and Larson, 2020] and machine learning [Chakraborti et al., 2020] has also been studied.

2 MODEL AND SOLUTION CONCEPTS

Model. For any $k \in \mathbb{N}$, we define $[k] := \{1, \dots, k\}$. An instance of the fair division problem is a tuple $\mathcal{I} = \langle N, M, V \rangle$, where $N := [n]$ is a set of n *agents*, $M := [m]$ is a set of m *goods*, and $V := \{v_1, \dots, v_n\}$ is a *valuation profile* that specifies for each agent $i \in N$ her preferences over the set of all possible *bundles* 2^M. This *valuation function* $v_i : 2^M \to \mathbb{N} \cup \{0\}$ maps each bundle to a non-negative integer. We write $v_{i,j}$ instead of $v_i(\{j\})$ for a single good $j \in M$. We assume that the valuation functions are *additive* so that for any $i \in N$ and $S \subseteq M$, $v_i(S) := \sum_{j \in S} v_{i,j}$, where $v_i(\varnothing) = 0$.

Allocation. An allocation $A := (A_1, \dots, A_n)$ is a (complete) n-partition of the set of goods M, where $A_i \subseteq M$ is the bundle allocated to agent $i \in N$.

Definition 1 (Envy-freeness)**.** An allocation A is: (i) *envy-free* (EF) if for every pair of agents $h, i \in N$, $v_i(A_i) \geq v_i(A_h)$ [Foley, 1967], (ii) *strongly envy-free up to one good* (sEF1) if for each agent $h \in N$ such that $A_h \neq \varnothing$, there exists a good $g_h \in A_h$ such that for every $i \in N$, $v_i(A_i) \geq v_i(A_h \setminus \{g_h\})$ [Conitzer et al., 2019], and (iii) *envy-free up to one good* (EF1) if for each pair of agents $h, i \in N$, there exists a good $g_h \in A_h$ such that $v_i(A_i) \geq v_i(A_h \setminus \{g_h\})$ [Budish, 2011, Lipton et al., 2004].

[3]See also the subsequent experiments by [Bereby-Meyer and Niederle, 2005, Kritikos and Bolle, 2001] and the back-and-forth discussion of [Bolton and Ockenfels, 2006, Engelmann and Strobel, 2006, Fehr et al., 2006].

[4]www.spliddit.org

[5]The IKEA effect is a cognitive bias in which people tend to value on products they helped to create highly [Norton et al., 2012].

	g_1	g_2	g_3	g_4	g_5	g_6
v_1	2^*	$\boxed{2^*}$	4	1	1	4
v_2	1	4	1^*	1^*	4	1
v_3	4	1	3	3	$\boxed{2^*}$	2^*

(a) sEF1

	g_1	g_2	g_3	g_4	g_5	g_6
v_1	2^*	2	4	1	$\boxed{1^*}$	4
v_2	1	4	1^*	1^*	4	1
v_3	4	$\boxed{1^*}$	3	3	2	2^*

(b) EF1

	g_1	g_2	g_3	g_4	g_5	g_6
v_1	2	2	4	1^*	1	4^*
v_2	1	4^*	1	1	4^*	1
v_3	4^*	1	3^*	3	2	2

(c) HEF-k

Fig. 1. Allocations satisfying (a) sEF1, (b) EF1 and (c) HEF-k for a fair division problem instance. Elements marked by a circle, rectangle, and diamond must be hidden or counterfactually removed to eliminate the envy from agents 1, 2, and 3 respectively.

Definition 2 (Envy-freeness with hidden goods)**.** An allocation A is *envy-free up to k hidden goods* (HEF-k) if $\exists S \subseteq M$, $|S| \leq k$, such that for every pair of agents $h, i \in N$, we have that $v_i(A_i) \geq v_i(A_h \backslash S)$ [Hosseini et al., 2020].

By the above definitions, EF implies sEF1, which implies EF1 and subsequently HEF-k for some $k \leq m$. Moreover, an allocation is EF if and only if it is HEF-0 and $\forall k \geq 0$ HEF-k implies HEF-$(k+1)$ [Hosseini et al., 2020]. We distinguish these classes throughout this paper with the following qualifications. First, we recognize two variants of envy-freeness up to one good by discerning allocations that are EF1 but not sEF1. Through an abuse of notation, we henceforth label this *weak* variant "EF1." Both variants (weak and strong) correspond to the *counterfactual* removal of goods when agents have full information about the entire allocation. Second, for any HEF-k allocation with hidden set S, each agent i knows their own bundle A_i but only has partial information about the goods in the bundle of any other agent h. Then, i has no envy among the observable (partial) allocation (i.e., $v_i(A_i) \geq v_i(A_h \backslash S)$). Furthermore, we assert that $|S| = k$ and that A is not HEF-k' with respect to any strict subset $S' \subset S$, where $|S'| = k' < k$.

Example 1. Figure 1 demonstrates three allocations for the same instance with three agents $1, 2, 3$ and six goods $g_1, \ldots, g_6$. They are demonstrated by the shaded elements in subfigures (a), (b), and (c), satisfying sEF1, EF1, and HEF-1 respectively. Elements outlined by a circle, rectangle, and diamond must be counterfactually removed (in the sEF1 and EF1 allocations) or hidden (in the HEF-1 allocation) to eliminate the envy of agents 1, 2, and 3 respectively.

Consider the EF1 allocation A where $A_1 = \{g_1, g_5\}$, $A_2 = \{g_3, g_4\}$, and $A_3 = \{g_2, g_6\}$. Although agent 1 is envious of agents 2 and 3, we have $v_1(A_1) \geq v_1(A_2 \backslash \{g_3\})$ and $v_1(A_1) \geq v_1(A_3 \backslash \{g_6\})$. For the HEF-1 allocation, rather, agent 1 is not envious of agent 3 because they only observe a partial allocation: $v_1(A_1) \geq v_1(A_3 \backslash S)$ where $S = \{g_3\}$. Agent 3 is not envious of agent 1 because they observe the entire allocation and $v_3(A_3) \geq v_3(A_1)$.

Notice that at most a single good is outlined in each agent's bundle in the sEF1 allocation, whereas multiple goods may be outlined in each bundle in the EF1 allocation.

3 EXPERIMENTAL DESIGN

We conducted an empirical study to compare the perceived fairness of multiple relaxations of envy-freeness—sEF1, EF1, and HEF-k—using a gamified pirate scenario (see Figure 2). Participants

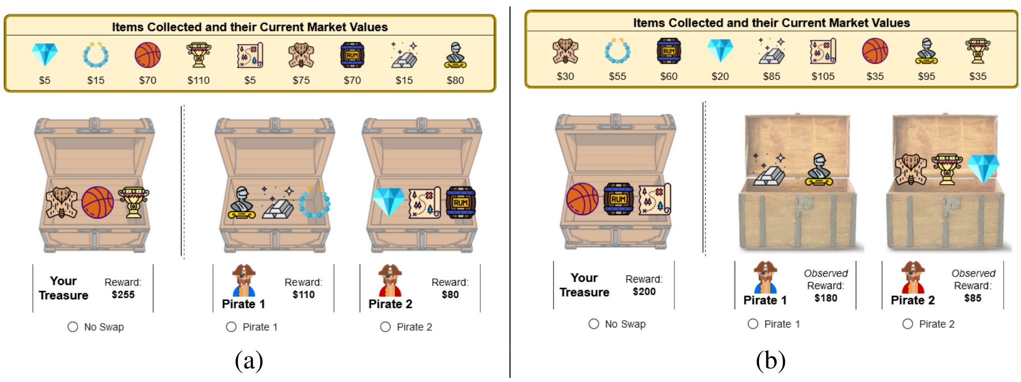

Fig. 2. Sample scenarios from the (a) EF1 and sEF1 treatments and (b) HEF-k treatment. In the EF1 and sEF1 treatments, participants have a 'birds-eye view' of all goods in all bundles. In the HEF-k treatment, participants observe only the revealed goods from other pirates' 'upright' boxes. All HEF-k treatment participants complete a training tutorial emphasizing this point.

were split into three treatments and given twelve scenarios. In each scenario, the participant was assigned the role of one member of a crew of pirates (agents) whose captain (a central authority) wished to divide goods, the spoils of a recent adventure, among the crew. Each scenario consisted of a number of goods, presented in a *marketplace*, and the bundles of (revealed) goods for each pirate in an allocation determined by the captain. Participants' *subjective* values for each bundle were determined by the given instance and the perspective of the participant. For instance, a participant could be offered the instance and allocation demonstrated by Figure 1(a) from the perspective of agent 1 and would value their bundle at $v_1(A_1) = 1 + 1 = 2$. Alternatively, their value for A_1 from the perspective of agent 3 would be $v_3(A_1) = 4 + 1 = 5$.

Given this information, participants were asked whether they wanted to *s*wap their bundle with that of another pirate of their choice, in its entirety, or keep their initial bundle. Participants had a stake in the outcome of their choices: they received a bonus payment if the total value of goods they collected surpassed a threshold. We measured participants' *swap rate*, the percentage of scenarios where a swap was chosen, and compared treatments using the Chi-square (χ^2) [McHugh, 2013] and Fisher's exact tests [Kim, 2017]. We then compared treatments upon segmenting our data by (i) the number of agents and goods (instance size), (ii) the distribution of goods across agents (allocation balance), and (iii) whether it is optimal for participants to swap or not, including the value of hidden goods (optimal choice).

Treatment details. Each participant was subjected to exactly one of three treatments – sEF1, EF1, and HEF-k – such that each allocation presented to the participant satisfied the same relaxation of EF. Across treatments, participants were shown their subjective values of the visible portions of the bundles of each agent. Participants in the sEF1 and EF1 treatments had full information about the allocations (see Figure 2(a)). Participants assigned the HEF-k treatment were shown their own bundles but only the visible portions of other agents' bundles (recall Definition 2; see Figure 2(b)). We explained through a tutorial that the visible allocation was incomplete by detailing the possibilities of the missing information: some goods may be allocated to and hidden by other pirates or discarded altogether. Participants could therefore enumerate the possible values of the other agents' bundles.

Our study employed 120 mutually exclusive participants for each of three Human Intelligence Tasks (HITs), corresponding to the three treatments, in Amazon's Mechanical Turk platform, totalling 360 participants. Our study was single-blind; participants were not aware of their treatment.

Perceived fairness. We measured perceptions of one aspect of fairness, envy, via swaps. Specifically, given an allocation A, we say that agent i *swaps* her bundle A_i with agent h if the agents exchange all goods within their bundles (including hidden goods). An agent choosing to swap bundles indicates that they are envious of another agent and thus does not perceive their bundle A_i as fair. We call the proportion of participants that swap under A its empirical *swap rate*, representing the aggregate perceived fairness of the scenario.

3.1 Data Set

The scenarios were sampled from a novel data set of 166 scenarios, each consisting of a fair division instance, an allocation partitioning the goods, and an assignment of the participant to one of the agent's perspectives.

Instances. We generated twenty-eight instances involving nine or ten goods: twenty-one `small` instances with three agents and seven `large` instances with five agents. Each valuation $v_{i,j}$ for $i \in N$ and $j \in M$ was sampled uniformly at random from $\{5, 10, \ldots, 120\}$ for `small` instances and $\{0, 10, \ldots, 150\}$ for `large` instances.

Allocations. For each instance, we computed three allocations satisfying sEF1, EF1, and HEF-k for a pre-specified $k \in \{0, 1, 2\}$ for the three corresponding treatments. Allocations were computed by randomly shuffling goods across agents until the desired properties were achieved. As we observe in Example 1, EF1 allocations can sometimes require the counterfactual removal of a larger number of goods than sEF1 allocations. To reflect this, and emphasize the distinction between EF1 and sEF1 allocations in our experiments, we pick EF1 allocations that require at least $n + 2$ goods to be counterfactually removed to eliminate envy among agents.

There were two levels of *balance* for allocations. A `balanced` allocation gives every agent a bundle of equal size, three (respectively, two) goods to each agent in a small (respectively, large) instance. In an `unbalanced` allocation, agents may have bundles consisting of different number of goods, with bundle sizes $(2, 4, 4)$ for `small` instances and either $(4, 2, 2, 1, 1)$ or $(3, 2, 2, 2, 1)$ for `large` instances.

Scenario properties. Our 166 scenarios were made with the following combinations: 63 allocations were affiliated with `small` instances, of which 45 were `balanced` and 18 were `unbalanced`, while 20 allocations were affiliated with `large` instances, of which 14 were `balanced` and 6 were `unbalanced`. Table 2 in Appendix A presents the number of allocations in each treatment succinctly. Each of the these 83 allocations provided two scenarios to the data set, corresponding to two perspectives we offered participants, yielding the 166 total scenarios.

3.2 Survey Outline

Participants undertook the following workflow (see Figure 6 in Appendix A). First, participants gave their consent to partake in our IRB-approved study after being informed of the study description, benefits, risks, rights, and project manager contact information. After being assigned a treatment and a randomly determined perspective, they subsequently answered twelve scenario questions, two questions soliciting scenario difficulty, and two attentiveness check questions. The scenarios were organized into four sections, each consisting of scenarios of different instance size and allocation balance that were selected uniformly at random from the appropriate data set, and then randomly permuted within the section. A complete survey therefore consisted of:

Variable	value	Pairs of Treatments					
		HEF-k, sEF1	HEF-k, EF1	sEF1, EF1	HEF-0, sEF1	HEF-1, sEF1	HEF-2, sEF1
All scenarios		0.286	0.173	0.604	0.150	0.306	0.371
Optimal Choice	`stay-is-opt`	4.533	0.512 (ns)	0.113 (†, #)	4.415	5.384	2.753 (ns)
	`swap-is-opt`	0.346	0.353	1.021 (#)	`N/A`	0.334	0.357
Repeated scenario (S7)		0.286	0.211	0.737	0.091	0.490 (#)	0.338

Table 1. Ratio of swap rates for pairs of treatments and adjusting for different variables. The χ^2 test is used for all comparisons except where denoted by a "†", in which the Fisher's exact test is used. All tests are statistically significant with $p < 0.001$, except where denoted by "#," which is significant with $p < 0.05$, and "ns," which is not statistically significant.

- Section 1 (S1–3): 3 `small-balanced` scenarios. If the treatment is HEF-k, then $k \in \{0, 1, 2\}$ respectively.
- Section 2 (S4–7): 3 `small-unbalanced` scenarios followed by S7 which is a repeat of S4. If the treatment is HEF-k, then $k \in \{0, 1, 2\}$ respectively for (S4–S6).
- Difficulty: self-reported rating for `small` scenarios.
- Section 3 (S8–10): 3 `large-balanced` scenarios. If the treatment is HEF-k, then $k \in \{0, 1, 2\}$ respectively.
- Section 4 (S11-12): 2 `large-unbalanced` scenarios. If the treatment is HEF-k, then $k = 1$.
- Difficulty: self-reported rating for `large` scenarios.

Tutorials. All participants were required to correctly answer a few tutorial questions prior to the scenarios.

The first tutorial taught participants that the value of a bundle was equal to the sum of values of the goods inside that bundle. Participants were presented with a bundle consisting of three goods, which were highlighted in the marketplace, and were asked to compute the bundle's value.

The second tutorial taught participants that whether they received a monetary bonus upon completing the survey is dependent on the total value they collect throughout its course. The participants were presented with three bundles, similar to Figure 2(a), and were asked if they wanted to keep their bundle (left) or swap it with either Pirate 1's bundle (middle) or Pirate 2's bundle (right). The bundle with the highest value was enforced as the correct choice.

HEF-k treatment participants were provided a third tutorial designed to teach them about goods in the marketplace that were not visibly allocated. Participants were presented with three bundles, similar to Figure 2(b), and were told that the missing goods may be either allocated to and hidden by the other pirates or discarded altogether. Participants were asked about the maximum number of goods that could be found in any one pirate's bundle, thus requiring them to reason about the location of missing goods.

Self-reported difficulty. The groups of seven `small` and five `large` scenarios were each succeeded by a question asking participants to rate the difficulty of the scenarios on a 5-point Likert scale from Very Easy (1) to Very Hard (5).

Attentiveness check questions. We incorporated many checks to ensure high quality responses from attentive human participants and dissuade fraud, which is a known problem for Mechanical Turk [Kennedy et al., 2020]. Additional details about participant qualifications can be found in Appendix A.

4 EXPERIMENTAL RESULTS

We test the empirical swap rate of each treatment as a measure for perceived fairness across all scenarios and while controlling for several variables. We further partition the HEF-k treatment into three separate sub-treatments—HEF-0, HEF-1, and HEF-2—and compare their swap rates with sEF1, with a focus on whether increasing the number of hidden goods affects perceived fairness. Separately, we compare the effect of treatment and size of instance on participants' cognitive effort, as measured by response time and self-reports of difficulty, for answering the scenarios.

Of particular interest is whether swap rates differ between treatments when a participant's *optimal* (i.e., *value-maximizing*) choice is to either *stay* or *swap* bundles, due to the following two observations. First, participants may be biased to accept their default bundle and maintain the status quo rather than make adjustments [Samuelson and Zeckhauser, 1988]. Second, HEF-k differs from the other treatments in that participants may not have enough information to distinguish which bundle is optimal. This raises the question of whether swap rates differ depending on optimal choice.

4.1 Perceived Fairness

We formalize our research questions as follows:

Research Questions: For any two treatments $X, Y \in \{\text{sEF1}, \text{EF1}, \text{HEF-}k\}$ or $\{\text{sEF1}, \text{HEF-0}, \text{HEF-1}, \text{HEF-2}\}$, do swap rates differ between X and Y overall and when adjusted independently for the variables: (i) *instance size*: `small` or `large`, (ii) *allocation balance*: `balanced` or `unbalanced`, and (iii) *optimal choice*: whether the value-maximizing choice is to keep the initial bundle (`stay-is-opt`) or to swap (`swap-is-opt`)?

Null Hypothesis: Swap rate is independent of treatment.

Alternate Hypothesis: Swap rate depends on treatment.

Our experiments provide statistically significant evidence for rejecting the null hypothesis that swap rate is independent of treatment. We draw this conclusion using the Chi-square (χ^2) test with $p < 0.001$ for nearly all combinations of pairs of treatments and values for the different confounding variables in our study and $p < 0.05$ for the remaining combinations.

Table 1 summarizes our findings. For each pair of treatments X and Y, identified by the column named 'X, Y', and for certain confounding variables, we present the ratio of swap rates between X to Y. For example, the ratio of swap rates from all HEF-k scenarios to all sEF1 scenarios is 0.286. In Appendix B we present Tables 4 and 9 which include more specific information about the p-values of the χ^2 and Fisher's exact test statistics and *effect size*, as measured by Cramer's V [Kim, 2017], about the tests. We include further tests conditioning on instance size and allocation balance.

Our main finding is that across all scenarios, (1) the perceived envy of HEF-k is significantly lower than that of either sEF1 and EF1, and (2) sEF1 allocations are less likely to be perceived as unfair than EF1 allocations, as we show in Figure 3. This holds true upon adjusting for instance size (`small` or `large`) and the allocation balance (`balanced` or `unbalanced`), and among scenarios where `swap-is-opt`. Thus, our main takeaway message is:

Allocations that are visibly envy-free through hiding goods are perceived to be fairer than allocations that are counterfactually envy-free via removing goods.

Segmented Data. Upon realizing this conclusion, we segment our data to draw additional insights. In particular, among HEF-k allocations, swap rate increases as the number of hidden goods increases (Table 6 in Appendix B): HEF-0 allocations induce less envy among the participants than either HEF-1 or HEF-2. This is perhaps because as more goods are hidden, participants are more cautious, more uncertain about the allocation's fairness, and spend more time on average to choose bundles (see Figure 8 in Appendix B). Further studies may be necessary to explain these results.

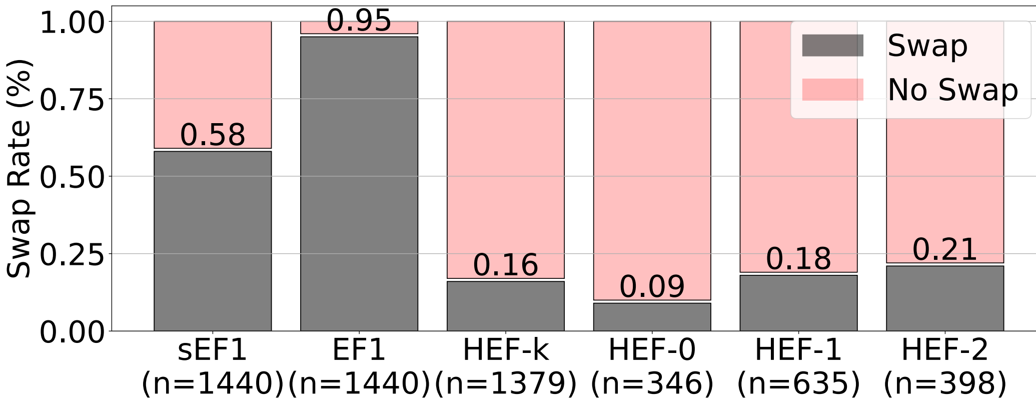

Fig. 3. Swap rates per treatment, all scenarios. Here, n is the number of scenarios per treatment.

Additionally, to control for any preferential bias toward our choices of pirate-related goods, we repeated a scenario and replaced the goods with identically-shaped gems of different colors (see "Repeated scenario (S7)" in Table 1). This test is described further in Appendix B.

Optimal Choice. We find that participants' perceived fairness is indeed affected by their optimal choice. Specifically, for each treatment (except HEF-0), participants' swap rates are statistically different between `swap-is-opt` and `stay-is-opt` scenarios (Table 5 in Appendix B).

Among `stay-is-opt` scenarios (see Figure 4), where participants' bundles have the highest value, we observe that sEF1 allocations are perceived with significantly lower envy than HEF-k allocations, and in turn EF1 allocations. Participants subjected to the sEF1 treatment could verify with certainty that their bundles have the highest value since all goods were visible. It may not be possible to make such determinations under the HEF-1 and HEF-2 treatments, where goods may be hidden to eliminate envy between other pairs of pirates. Indeed, the hidden goods may all be allocated to another pirate, hypothetically raising the value of that pirate's bundle to be the highest, justifying a swap. Surprisingly, HEF-0 and EF1 induce higher envy than sEF1 allocations, despite it being equally possible to verify that the participant's bundle has the highest value. This may be due to *framing effect* biases by which participants may not have incorrectly assumed that goods were missing [Tversky and Kahneman, 1985]. However, further tests are needed to confirm this conjecture. Swap rates between either HEF-k and EF1, and HEF-2 and sEF1, are not statistically significant in this case.

When `swap-is-opt` (Figure 7 in Appendix B), participants swap their bundles significantly less under the HEF-k treatment than the sEF1 and EF1 treatments. This supports our overall conclusion that participants desire allocations that are not visibly unfair. Since all goods are visible under the sEF1 and EF1 treatments, the participant has clear evidence that her allocated bundle has a lower value than that of another pirate. Under the HEF-k treatment, rather, where the allocation of some goods is hidden, participants perceive significantly lower envy even when they are allocated a lower-valued bundle. Recall that HEF-0 is equivalent to EF, so there are no such scenarios when `swap-is-opt`.

4.2 Cognitive Effort

In addition to our tests of perceived fairness, we investigate the extent to which cognitive effort varies by treatment. Specifically, we measure:

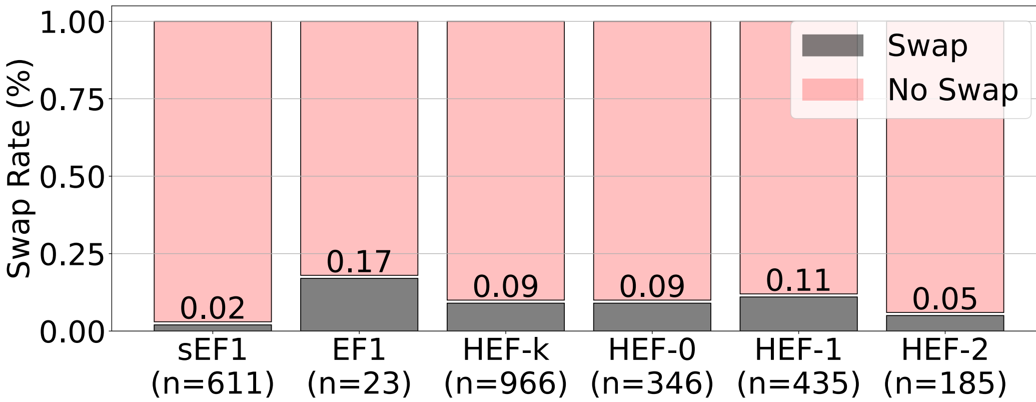

Fig. 4. Swap rates per treatment, `stay-is-opt`. Here, n is the number of questions per treatment.

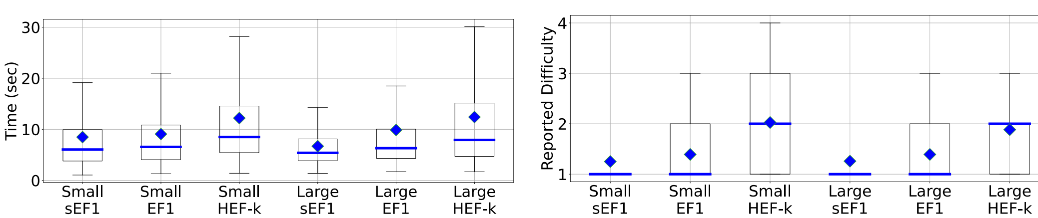

Fig. 5. Box-plots of (left) time spent per scenario and (right) reported difficulty (higher scores indicate higher difficulty) by treatment. Outliers excluded.

- *response time*, the time elapsed between each scenario page being made available to the participant and the participant submitting her choice, and
- *scenario difficulty*, using the self-reports of scenario difficulty solicited immediately after the `small` and then the `large` scenarios.

We check whether the mean response time or reported difficulty on a five-point Likert scale is different between pairs of treatments, while adjusting for different variables such as (i) *optimal choice* – whether the value-maximizing choice is to stay with the allocated bundle (`stay-is-opt`) – and (ii) *instance size* – `small` or `large`.

Our experiments provide sufficient evidence to reject the null hypothesis that cognitive effort for HEF-k is the same as either sEF1 or sEF1, using a two-sided Welch t-test ($p < 0.001$). Tables 7 and 8 in Appendix B summarize our findings for response times per scenario and reported feedback, respectively. Figure 5 (left) illustrates that the average sEF1 response time is lowest and HEF-k is highest, while EF1 splits the two. Similarly, in Figure 5 (right), participants report that sEF1 scenarios are easiest while HEF-k is the most difficult and EF1 lies in between. These observations hold for either instance size and demonstrate that HEF-k instances cause higher cognitive burden on participants. Effect size for these statistical tests, as measured by Cohen's D [Cohen, 1992], is reported in Tables 12 and 13 in Appendix B.

4.3 Descriptive Comments from Participants

We identify the participants anonymously using a letter S, E, or H corresponding to their treatment (sEF1, EF1, or HEF-k).

Participants in the EF1 treatment consistently noted that other pirates' bundles *"were usually more valuable"* ($E22$), so they should *"swap with the highest yielding chest"* ($E17$, $E8$, $E15$). On the other hand, HEF-k participants noted *"it seemed a no brainer to just never swap"* ($H8$, $H28$), either because it was the *"safest bet"* ($H49$) or the *"greatest statistical chance of getting higher reward"* ($H59$). These comments are consistent with our data that swap rates were significantly lower for HEF-k than the other treatments.

A few participants explicitly addressed concerns about fairness. Participant $S57$ suggested *"it didn't seem like a fair split"* while $S63$ declared they wouldn't swap in real life *"because it would be unfair to the other person."* Despite this hesitation, participant $E96$ reasoned that because *"there was no defining reason why anyone would get more than others"* due to differing effort, they should still select the most valuable treasure. These comments resemble Herreiner and Puppe [2009]'s findings that participants care more about inequality aversion than EF to ensure fairness. Still, it is unclear to what extent participants' opinions and choices are affected by strategic interaction with other humans, as in [Herreiner and Puppe, 2009], as opposed to confederates, as in our work. We leave this intriguing question for future work.

5 LIMITATIONS AND FUTURE WORK

Our experiment was limited, in part, by the scenario size, uncontrolled bias, and the type of fairness notions we tested. First, our experiment tested scenarios for a cross-section of the numbers of goods m, agents n, and goods hidden k. We sought to provide meaningful information to participants without causing cognitive overload. Future work may determine how sensitive our results are to scaling these values.

Second, we controlled for effects of our pirate-related goods on participants' decision-making by randomly permuting good images and repeating a scenario with identical multi-colored gems. However, we may not have accounted for all confounding variables, such as framing effects. For example, while participants' values were *subjective*, they could have reasoned that values were *objective* based on the scenario presentations. Furthermore, our setup of the HEF-k treatment conveyed to participants the possibility that hidden goods may not be allocated at all. This subsumes the reality that all goods were indeed allocated, yet is par to the definition of HEF-k [Hosseini et al., 2020]. Still, this is not be the only way to implement an information scheme, as exemplified by Herreiner and Puppe [2009], who explicated all subjective values for all agents. Further work may be necessary to determine the sensitivity of our results to framing effects and what information is provided.

Finally, our experiments compared the relative perceived fairness of two intrapersonal envy-based concepts. Both EF1 and HEF-k presume that people find allocations fair if they are not envious of others' bundles; we measured perceived envy via swap rate according to this standard. Our results confirm that people experience less envy among allocations for which they theoretically and epistemically should not experience envy (HEF-k) than those requiring counterfactual reasoning (EF1). Whether envy-based notions of fairness are more appropriate than comparative forms like inequality aversion is a topic of ongoing debate [Herreiner and Puppe, 2009].

Our work presents an important first step to provide an empirical comparison about perceived fairness of relaxations of EF. Future empirical research may investigate perceived fairness of other notions, such as maximin-share [Budish, 2011] and proportionality, attitudes towards procedural versus distributive fairness, and whether moral judgements are affected by participants either having a stake in the resource division task or making decisions as outside observers.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for helpful comments. HH acknowledges support from NSF grants #2144413, #2107173, and #2052488. RV acknowledges support from DST INSPIRE grant no. DST/INSPIRE/04/2020/000107 and SERB grant no. CRG/2022/002621. LX cknowledges support from NSF grants #2007994 and #2106983, and a Google Research Award.

REFERENCES

J Stacy Adams. 1963. Towards an Understanding of Inequity. *Journal of Psychopathology and Clinical Science* 67, 5 (1963).

Georgios Amanatidis, Georgios Birmpas, Aris Filos-Ratsikas, and Alexandros A Voudouris. 2022. Fair Division of Indivisible Goods: A Survey. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*. 5385–5393.

Haris Aziz, Sylvain Bouveret, Ioannis Caragiannis, Ira Giagkousi, and Jérôme Lang. 2018. Knowledge, Fairness, and Social Constraints. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 4638–4645.

Haris Aziz, Bo Li, Herve Moulin, and Xiaowei Wu. 2022. Algorithmic Fair Allocation of Indivisible Items: A Survey and New Questions. In *ACM SIGecom Exchanges*, Vol. 20. 24–40.

Yoella Bereby-Meyer and Muriel Niederle. 2005. Fairness in bargaining. *Journal of Economic Behavior & Organization* 56, 2 (2005), 173–186.

EB Bolton and Axel Ockenfels. 2006. Measuring efficiency and equity motives. A comment on "inequality aversion, efficiency, and maximin preferences in simple distribution experiments". *American Economic Review* 96, 5 (2006), 1906–1911.

Gary E Bolton and Axel Ockenfels. 2000. ERC: A theory of equity, reciprocity, and competition. *American economic review* 91, 1 (2000), 166–193.

Steven J. Brams and Alan D. Taylor. 1996. *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge University Press.

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

Eric Budish, Gérard P Cachon, Judd B Kessler, and Abraham Othman. 2017. Course Match: A Large-Scale Implementation of Approximate Competitive Equilibrium from Equal Incomes for Combinatorial Allocation. *Operations Research* 65, 2 (2017), 314–336.

Tapabrata Chakraborti, Arijit Patra, and J Alison Noble. 2020. Contrastive Fairness in Machine Learning. *IEEE Letters of the Computer Society* 3, 02 (2020), 38–41.

Gary Charness and Matthew Rabin. 2002. Understanding social preferences with simple tests. *The quarterly journal of economics* 117, 3 (2002), 817–869.

Jacob Cohen. 1992. Statistical Power Analysis. *Current Directions in Psychological Science* 1, 3 (1992), 98–101. http://www.jstor.org/stable/20182143

Vincent Conitzer, Rupert Freeman, Nisarg Shah, and Jennifer Wortman Vaughan. 2019. Group Fairness for the Allocation of Indivisible Goods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1853–1860.

Harald Cramér. 1946. *Mathematical Methods of Statistics*. Princeton University Press Press.

Geoffroy De Clippel and Kareen Rozen. 2022. Fairness through the Lens of Cooperative Game Theory: An Experimental Approach. *American Economic Journal: Microeconomics* 14, 3 (2022), 810–36.

Greg d'Eon and Kate Larson. 2020. Testing Axioms against Human Reward Divisions in Cooperative Games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 312–320.

Dirk Engelmann and Martin Strobel. 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review* 94, 4 (2004), 857–869.

Dirk Engelmann and Martin Strobel. 2006. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Reply. *American Economic Review* 96, 5 (2006), 1918–1923.

Ernst Fehr, Michael Naef, and Klaus M Schmidt. 2006. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review* 96, 5 (2006), 1912–1917.

Ernst Fehr and Klaus M Schmidt. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* 114, 3 (1999), 817–868.

Duncan Karl Foley. 1967. *Resource Allocation and The Public Sector*. Yale University. 45–98 pages.

Wulf Gaertner. 2009. *A primer in social choice theory: Revised edition*. Oxford University Press, USA.

John C Harsanyi. 1990. Interpersonal Utility Comparisons. In *Utility and Probability*. Springer, 128–133.

Dorothea K Herreiner and Clemens Puppe. 2007. Distributing indivisible goods fairly: Evidence from a questionnaire study. *Analyse & Kritik* 29, 2 (2007), 235–258.

Dorothea K Herreiner and Clemens Puppe. 2010. Inequality aversion and efficiency with ordinal and cardinal social preferences—An experimental study. *Journal of Economic Behavior & Organization* 76, 2 (2010), 238–253.

Dorothea K Herreiner and Clemens D Puppe. 2009. Envy Freeness in Experimental Fair Division Problems. *Theory and Decision* 67, 1 (2009), 65–100.

Hadi Hosseini, Sujoy Sikdar, Rohit Vaish, Jun Wang, and Lirong Xia. 2020. Fair Division through Information Withholding. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, Vol. 34. 2014–2021.

Daniel Kahneman, Jack L Knetsch, and Richard Thaler. 1986. Fairness as a constraint on profit seeking: Entitlements in the market. *The American economic review* (1986), 728–741.

Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D Waggoner, Ryan Jewell, and Nicholas JG Winter. 2020. The Shape of and Solutions to the MTurk Quality Crisis. *Political Science Research and Methods* 8, 4 (2020), 614–629.

Hae-Young Kim. 2017. Statistical Notes for Clinical Researchers: Chi-Squared Test and Fisher's Exact Test. *Restorative Dentistry & Endodontics* 42, 2 (2017), 152–155.

Tobias König, Dorothea Kübler, Lydia Mechtenberg, and Renke Schmacker. 2019. *Fair Procedures with Naive Agents: Who Wants the Boston Mechanism?* Rationality and Competition Discussion Paper Series 222. Discussion Paper. https://ideas.repec.org/p/rco/dpaper/222.html

James Konow. 2003. Which is the fairest one of all? A positive analysis of justice theories. *Journal of economic literature* 41, 4 (2003), 1188–1239.

Alexander Kritikos and Friedel Bolle. 2001. Distributional concerns: equity-or efficiency-oriented? *Economics Letters* 73, 3 (2001), 333–338.

Maria Kyropoulou, Josué Ortega, and Erel Segal-Halevi. 2022. Fair Cake-Cutting in Practice. *Games and Economic Behavior* 133 (2022), 28–49.

Min Kyung Lee. 2018. Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society* 5, 1 (2018), 1–16.

Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1035–1048.

Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 182 (nov 2019), 26 pages. https://doi.org/10.1145/3359284

Richard J Lipton, Evangelos Markakis, Elchanan Mossel, and Amin Saberi. 2004. On Approximately Fair Allocations of Indivisible Goods. In *Proceedings of the 5th ACM Conference on Electronic Commerce*. 125–131.

Mary L McHugh. 2013. The Chi-Square Test of Independence. *Biochemia Medica* 23, 2 (2013), 143–149.

Hervé Moulin. 2019. Fair Division in the Internet Age. *Annual Review of Economics* 11 (2019), 407–441.

Michael I Norton, Daniel Mochon, and Dan Ariely. 2012. The IKEA Effect: When Labor Leads to Love. *Journal of Consumer Psychology* 22, 3 (2012), 453–460.

Parag A Pathak, Tayfun Sönmez, M Utku Ünver, and M Bumin Yenmez. 2021. Fair Allocation of Vaccines, Ventilators and Antiviral Treatments: Leaving No Ethical Value Behind in Health Care Rationing. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 785–786.

John Rawls. 2004. A Theory of Justice. In *Ethics*. Routledge, 229–234.

William Samuelson and Richard Zeckhauser. 1988. Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty* 1, 1 (1988), 7–59.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

Amartya Sen. 2018. *Collective Choice and Social Welfare*. Harvard University Press.

Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2459–2468.

Amos Tversky and Daniel Kahneman. 1985. The Framing of Decisions and the Psychology of Choice. In *Behavioral Decision Making*. Springer, 25–41.

Tom R Tyler and E Allan Lind. 2002. Procedural Justice. In *Handbook of Justice Research in Law*. Springer, 65–92.

Menahem E Yaari and Maya Bar-Hillel. 1984. On dividing justly. *Social choice and welfare* 1 (1984), 1–24.

APPENDIX

A EXPERIMENTAL DESIGN: ADDITIONAL DETAILS

The workflow for a participant in our study is illustrated in Figure 6. Here, we provide details on the measures used in our study to ensure high quality responses from participants recruited from the Amazon Mechanical Turk platform. In addition to screening participants and qualifying responses, we discuss how participants were incentivized to provide high quality responses through our payments structure.

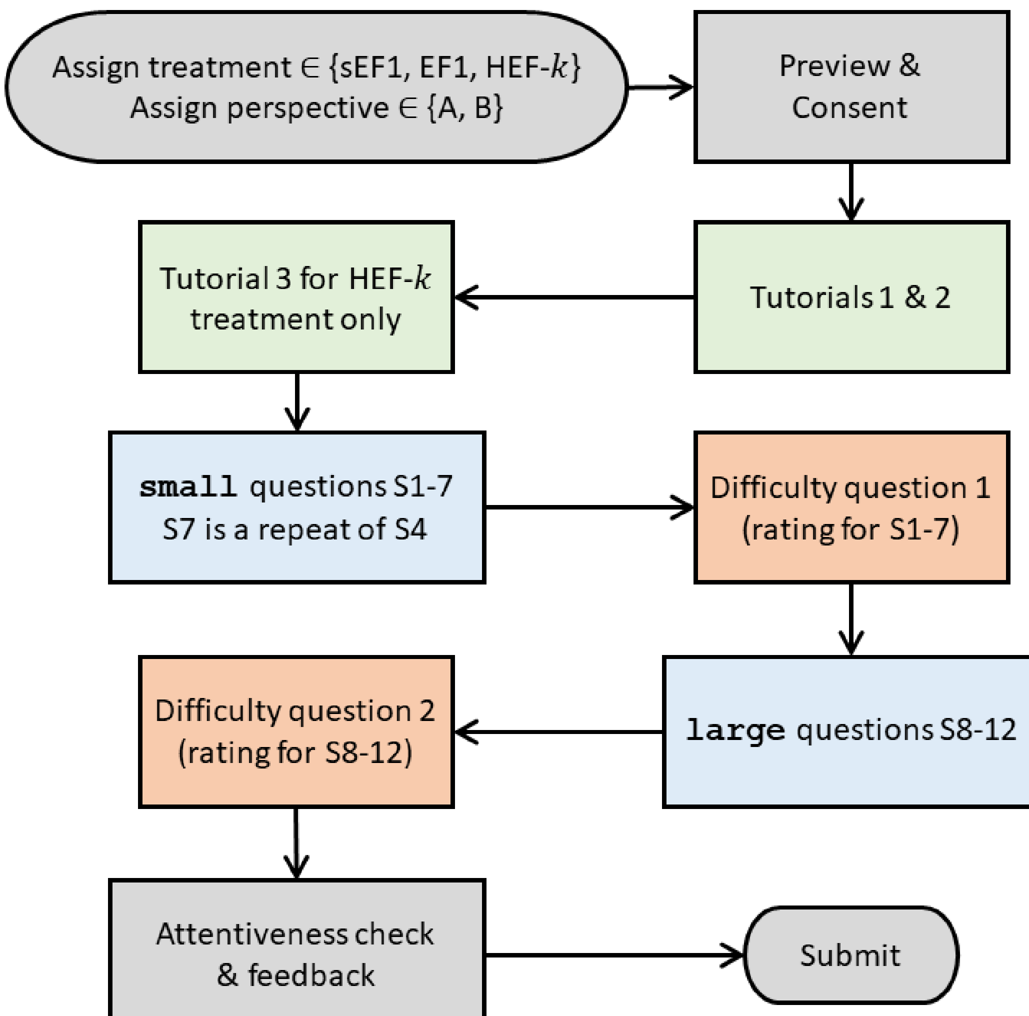

Fig. 6. The workflow of a participant.

Scenario properties. The number of allocations that satisfied each treatment is presented in Table 2.

Perspective. Participants were randomly assigned to assume the role of either the first or last (i.e., third or fifth) agent in the instance. Providing two perspectives expanded our data set and enabled

Instance Size	Allocation Balance	Treatment				
		sEF1	EF1	HEF-0	HEF-1	HEF-2
small	balanced $(3,3,3)$	15	15	5	5	5
	unbalanced $(2,2,4)$	6	6	2	2	2
large	balanced $(2,2,2,2,2)$	5	5	1	2	1
	unbalanced $(4,2,2,1,1)$	1	1	0	0	0
	unbalanced $(3,2,2,2,1)$	1	1	0	2	0

Table 2. Number of scenarios per treatment and perspective, given instance size and allocation balance (number of goods per each agent in parentheses).

participants to have different goods in their bundles for the same instances. This did not bias our results as valuations were randomly generated and allocations did not depend on agents' identities.

Incentives. In order to realize the assigned in-game valuations as real-world *value*, participants were incentivized to accumulate high-value bundles throughout the survey. Specifically, each participant was eligible to receive two payments: (1) a *base* payment of \$0.50 for completing the survey in its entirety, and (2) a *bonus* payment of \$0.50 for accumulating at least \$2000 worth of goods through all scenarios as measured by participants' assigned subjective valuations. Hence, we are able to emulate a real-world setting through our experiment with fictional pirate-related goods.

Note that within the HEF-k treatment, participants accumulate the values of any hidden goods of their chosen bundle as well. The bonus threshold was also chosen to encourage participants to pay greater attention to the study and not choose randomly for each scenario. We determined the threshold by computing the minimum and maximum total value any participant could obtain on any survey using our data set. We then chose \$2000 which falls between between 71% and 84% for these ranges.

Response qualifications. In order to obtain high quality responses, participation in our study was restricted to Mechanical Turk workers who (a) had at least an 80% approval rate on previous tasks, (b) had completed at least 100 tasks, (c) were located in either the United States or Canada[6], (d) had a Master's qualification[7] on the Mechanical Turk platform, and (e) had not attempted or taken the survey before. Through the experiment we adjusted the minimum HIT approval rate (%) and minimum number of HITs approved that were necessary in order to attract Mechanical Turk Workers to participate; see Table 3.

Attentiveness check questions. We incorporated many checks to ensure high quality responses from attentive human participants and dissuade fraud, which is a known problem for unprotected Mechanical Turk studies [Kennedy et al., 2020]. Prior to the tutorial, participants answered a simple arithmetic problem to ensure they were not bots. On the final page, they answered (1) their favorite good and (2) final comments or questions. We presumed that we could identify inattentive participants giving poor quality data, as they would not be able to answer these prompts appropriately. We did

[6]We restricted location to ensure language proficiency and prevent any potential issues due to linguistic barriers.
[7]Workers with Master's qualification, determined by Mechanical Turk, are those who "have consistently demonstrated a high degree of success in performing a wide range of HITs across a large number of Requesters." See https://www.mturk.com/worker/help.

Treatment	Minimum Approval Rate	Minimum Number Approved	Count
sEF1	95%	1000	120
EF1	95%	1000	20
	80%	100	120
HEF-k	95%	1000	20
	90%	1000	62
	80%	1000	76
	90%	500	83
	90%	100	92
	80%	100	120

Table 3. Number of participants satisfying each qualification range, per treatment, as measured by minimum approval rate and minimum approval range (not mutually exclusive).

not find any participants' responses to be of poor quality by these measures, so we did not discard any responses.

B EXPERIMENTAL RESULTS: ADDITIONAL TABLES AND FIGURES

Controlling for the choice of goods. Our scenarios presented goods related to a pirate's adventure, such as a map, rum, and a diamond. This gamified scenario stands in for a wider variety of fair division problems, such as inheritance division [Brams and Taylor, 1996], allocating medical resources [Pathak et al., 2021], and course allocation [Budish et al., 2017]. To control for any preferential bias toward these goods, we repeated a scenario and replaced the goods with identically-shaped gems of different colors. The repeated scenario (S7) was identical to the original (S4), which is `small-unbalanced` but with varying numbers of hidden goods k for the HEF-k treatment. Additionally, the pictures representing the goods were randomly permuted for every scenario.

We find that every null hypothesis that was rejected by comparing responses on all scenarios is also rejected when the test is performed only on the repeated scenario (see row labeled "Repeated scenario (S7)" in Table 1). Furthermore, the ratio of swap rates for each pair of treatments remains similar as well. Therefore, our results do not appear to be impacted by the choice of goods.

Perceived envy under different treatments. Table 4 presents the ratio of swap rates and p-values for all questions under different pairs of treatments.

Perceived envy of `swap-is-opt` versus `stay-is-opt` scenarios under different treatments. Table 5 depicts the ratio of swap rates for `swap-is-opt` versus `stay-is-opt` scenarios under different treatments. Notably, there is a statistically significant difference between `swap-is-opt` and `stay-is-opt` for all treatments (except for HEF-0, for which there are no `swap-is-opt` scenarios).

Perceived envy comparing HEF-k *treatments.* Table 6 depicts the results of our primary hypothesis test comparing the independence of swap rates and treatments, for the pairwise treatments of HEF-0, HEF-1, and HEF-2, and adjusting for different variables. These results complement Table 1 and demonstrate χ^2 statistics and associated p-values. Notably, there is a statistically significant difference between HEF-0 and both HEF-1 and HEF-2 for all questions, although there is no significant difference between the treatments conditioning on either `stay-is-opt` or `swap-is-opt`. By Figure 3, this suggests HEF-0 (i.e., envy-free) allocations are perceived as more fair than either HEF-1 or HEF-2 allocations.

Variable	**value**	**Pairs of Treatments**					
		HEF-k, sEF1	HEF-k, EF1	sEF1, EF1	HEF-0, sEF1	HEF-1, sEF1	HEF-2, sEF1
All scenarios		0.286 p : ☆☆☆	0.173 p : ☆☆☆	0.604 p : ☆☆☆	0.150 p : ☆☆☆	0.306 p : ☆☆☆	0.371 p : ☆☆☆
Optimal Choice	`stay-is-opt`	4.533 p : ☆☆☆	0.512 p : ns	0.113 †p : ☆☆	4.415 p : ☆☆☆	5.384 p : ☆☆☆	2.752 p : ns
	`swap-is-opt`	0.346 p : ☆☆☆	0.353 p : ☆☆☆	1.021 p : ☆	N/A	0.334 p : ☆☆☆	0.357 p : ☆☆☆
Instance Size	`small`	0.294 p : ☆☆☆	0.186 p : ☆☆☆	0.634 p : ☆☆☆	0.114 p : ☆☆☆	0.380 p : ☆☆☆	0.394 p : ☆☆☆
	`large`	0.268 p : ☆☆☆	0.150 p : ☆☆☆	0.561 p : ☆☆☆	0.322 p : ☆☆☆	0.253 p : ☆☆☆	0.285 p : ☆☆☆
Balance	`balanced`	0.320 p : ☆☆☆	0.167 p : ☆☆☆	0.523 p : ☆☆☆	0.246 p : ☆☆☆	0.267 p : ☆☆☆	0.426 p : ☆☆☆
	`unbalanced`	0.259 p : ☆☆☆	0.178 p : ☆☆☆	0.686 p : ☆☆☆	0.073 p : ☆☆☆	0.310 p : ☆☆☆	0.329 p : ☆☆☆
Repeated scenario (S7)		0.286 p : ☆☆☆	0.211 p : ☆☆☆	0.737 p : ☆☆☆	0.091 p : ☆☆☆	0.490 p : ☆☆	0.338 p : ☆☆☆

Table 4. Ratio of the swap rates and p-values of the test statistic for testing the independence of swap rates and treatments under different pairs of treatments, and adjusting for different variables. The χ^2 test is used except when the p-value is annotated with a "†", in which case, it is the result of the Fisher's exact test. The p-value of the test statistic is represented as follows: a cell labeled ns (not significant) implies that $p > 0.05$, ☆ for $p \in (0.01, 0.05]$, ☆☆ for $p \in (0.001, 0.01]$, and ☆☆☆ for $p < 0.001$.

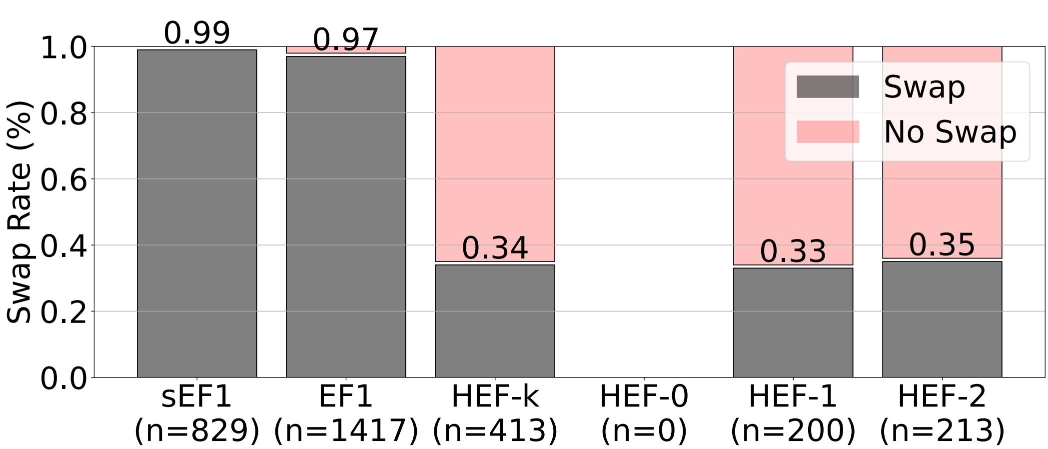

Fig. 7. Swap rates per treatment, `swap-is-opt`. Here, n is the number of questions per treatment.

Treatment	sEF1	HEF-k	EF1	HEF-0	HEF-1	HEF-2
`swap-is-opt` / `stay-is-opt`	50.241 p : ☆☆☆	3.835 p : ☆☆☆	5.559 †p : ☆☆☆	N/A	3.121 p : ☆☆☆	6.514 p : ☆☆☆

Table 5. Ratio of the swap rates and p-values of the test statistic for testing the independence of swap rates and optimal choice under different treatments. The χ^2 test is used except when the p-value is annotated with a "†", in which case, it is the result of the Fisher's exact test. The p-value of the test statistic is represented as follows: a cell labeled ns (not significant) implies that $p > 0.05$, ☆ for $p \in (0.01, 0.05])$, ☆☆ for $p \in (0.001, 0.01]$, and ☆☆☆ for $p < 0.001$.

Cognitive effort on HEF-k *allocations.* Figure 8 presents the distribution of time spent per scenario over all HEF-0, HEF-1 and HEF-2 scenarios. We find that overall, as the number of hidden goods increases, the cognitive effort, measured as the amount of time spent in order to decide which bundle

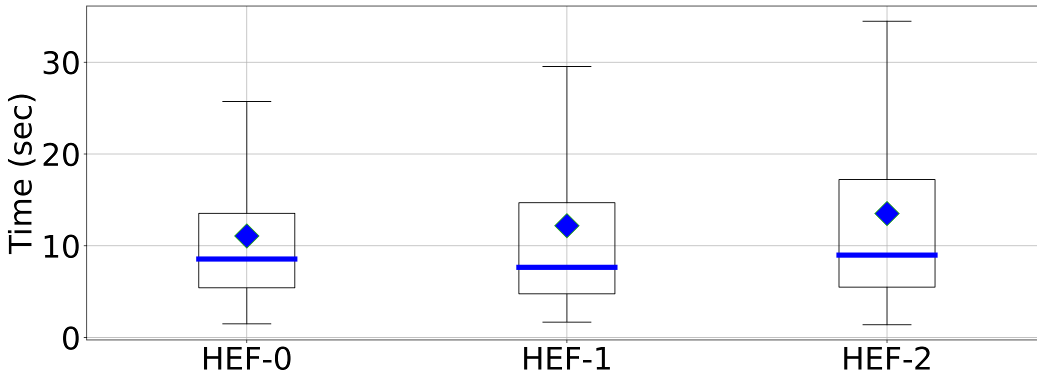

Fig. 8. Box-plot of time spent per scenario by treatment with averages shown. Outliers excluded.

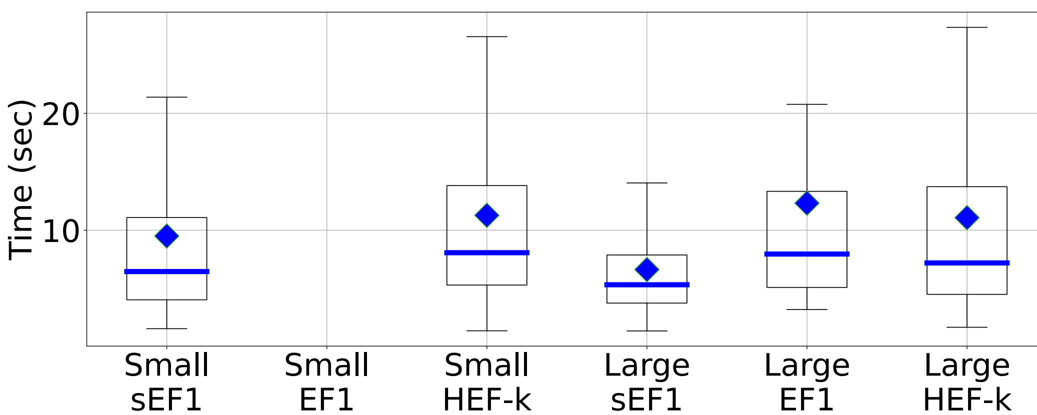

Fig. 9. Box-plot of time spent per scenario by treatment with average shown, conditioned on `stay-is-opt`. Outliers excluded. There were no such `small` EF1 scenarios.

to keep, also increases. Specifically, both the mean and variance of time spent increases as the value of k increases for HEF-k scenarios.

Notice that in an HEF-0 scenario, the participant already has the highest valued bundle and this is readily verifiable since all goods are visible. However, as k increases, the participant must reason about and form beliefs about how the hidden goods may be allocated to the other pirates. The task of computing and deciding whether it may be worth swapping for another pirate's bundle therefore becomes increasingly more complex as more goods are hidden.

Cognitive effort on `stay-is-opt` scenarios. As Figure 9 shows for `stay-is-opt` scenarios, hiding goods under the HEF-k treatment comes at the cost of an increased cognitive burden on the participants. Here, the participant's bundle has the highest value. This is evident for the sEF1 and EF1 treatments, but may not be clear under the HEF-k treatment, where goods may need to be hidden in order to eliminate envy between the other pirates.

Effect Size. We supplement our results of statistical significance with their effect sizes. Table 9, Table 10, Table 11, Table 12, and Table 13 demonstrate the effect size for each statistically significant

Variable	value	Pairs of Treatments		
		HEF-0, HEF-1	HEF-0, HEF-2	HEF-1, HEF-2
All scenarios		0.494 p : ☆☆	0.406 p : ☆ ☆ ☆	0.826 p : ns
Optimal Choice	`stay-is-opt`	0.820 p : ns	1.604 p : ns	1.956 p : ns
	`swap-is-opt`	N/A	N/A	0.937 p : ns
Instance Size	`small`	0.299 p : ☆ ☆ ☆	0.289 p : ☆ ☆ ☆	0.966 p : ns
	`large`	1.271 p : ns	1.130 p : ns	0.889 p : ns
Balance	`balanced`	0.992 p : ns	0.578 p : ns	0.627 p : ns
	`unbalanced`	0.237 p : ☆ ☆ ☆	0.223 p : ☆ ☆ ☆	0.941 p : ns
Repeated scenario (Q7)		0.186 p : ☆	0.270 p : ns	1.448 p : ns

Table 6. Ratio of swap rates and p-values of the χ^2 statistic for testing the independence of swap rates and treatments under different pairs of treatments, and adjusting for different variables.
Key: (ns : $p > 0.05$) (☆ : $p \in (0.01, 0.05]$), (☆☆ : $p \in (0.001, 0.01]$), (☆ ☆ ☆ : $p < 0.001$).

Variable	Instance Size	Pairs of Treatments		
		sEF1, EF1	sEF1, HEF-k	EF1, HEF-k
All scenarios	`small`	p : ns	p : ☆ ☆ ☆	p : ☆ ☆ ☆
	`large`	p : ☆	p : ☆ ☆ ☆	p : ☆ ☆ ☆
`stay-is-opt`	`small`	N/A	p : ☆	N/A
	`large`	p : ns	p : ☆ ☆ ☆	p : ns
Variable		**Pairs of Treatments**		
		HEF-0, HEF-1	HEF-0, HEF-2	HEF-1, HEF-2
All scenarios		p : ns	p : ☆☆	p : ns

Table 7. p-values of the t statistic for testing equal means of participant response times per scenario using Welch's t-test – for different pairs of treatments, and adjusting for different variables.
Key: (ns : $p > 0.05$) (☆ : $p \in (0.01, 0.05]$), (☆☆ : $p \in (0.001, 0.01]$), (☆ ☆ ☆ : $p < 0.001$).

Variable	Instance Size	Pairs of Treatments		
		sEF1, EF1	sEF1, HEF-k	EF1, HEF-k
All senarios	`small`	p : ☆	p : ☆ ☆ ☆	p : ☆ ☆ ☆
	`large`	p : ns	p : ☆ ☆ ☆	p : ☆ ☆ ☆

Table 8. p-values of the t statistic for testing equal means of participant reported feedback using Welch's t-test – for different pairs of treatments.
Key: (ns : $p > 0.05$) (☆ : $p \in (0.01, 0.05]$), (☆☆ : $p \in (0.001, 0.01]$), (☆ ☆ ☆ : $p < 0.001$).

test for Table 4 (and Table 1), Table 5, Table 6, Table 7, and Table 8 respectively. Effect sizes are measured with Cramer's V for χ^2 tests [Cramér, 1946] and Cohen's d for Welch t-tests [Cohen, 1992].

Variable	**value**	**Pairs of Treatments**					
		HEF-k, sEF1	HEF-k, EF1	sEF1, EF1	HEF-0, sEF1	HEF-1, sEF1	HEF-2, sEF1
All scenarios		V : ☆☆	V : ☆ ☆ ☆	V : ☆☆	V : ☆☆	V : ☆☆	V : ☆☆
Optimal Choice	`stay-is-opt`	V : ☆	ns	†OR : 0.096 95% CI : $(0.026, 0.447)$	V : ☆	V : ☆	ns
	`swap-is-opt`	V : ☆ ☆ ☆	V : ☆ ☆ ☆	V : ☆	N/A	V : ☆ ☆ ☆	V : ☆ ☆ ☆
Instance Size	`small`	V : ☆☆	V : ☆ ☆ ☆	V : ☆☆	V : ☆☆	V : ☆☆	V : ☆☆
	`large`	V : ☆☆	V : ☆ ☆ ☆	V : ☆☆	V : ☆☆	V : ☆☆	V : ☆☆
Balance	`balanced`	V : ☆☆	V : ☆ ☆ ☆	V : ☆☆	V : ☆☆	V : ☆☆	V : ☆☆
	`unbalanced`	V : ☆☆	V : ☆ ☆ ☆	V : ☆☆	V : ☆☆	V : ☆☆	V : ☆☆
Repeated scenario (S7)		V : ☆☆	V : ☆ ☆ ☆	V : ☆☆	V : ☆☆	V : ☆☆	V : ☆☆

Table 9. Effect size demonstrating the strength in statistically significant relationships between swap rates and treatments – under different pairs of treatments, adjusting for different variables, and corresponding to tests in Table 1. Not significant tests are labelled ns. Cramer's V is reported for χ^2 tests as follows: ☆ for $V \leq 0.2$, ☆☆ for $V \in (0.2, 0.6]$, and ☆ ☆ ☆ for $V > 0.6$. Odds ratio and 95% confidence intervals are reported for Fisher's exact test, annotated by "†."

Treatment	sEF1	HEF-k	EF1	HEF-0	HEF-1	HEF-2
`swap-is-opt` / `stay-is-opt`	V : ☆ ☆ ☆	V : ☆☆	†OR : 0.007 95% CI : $(0.002, 0.023)$	N/A	V : ☆☆	V : ☆☆

Table 10. Effect size demonstrating the strength in statistically significant relationships between swap rates and optimal choice, for different treatments in Table 5. Not significant tests are labelled ns. Cramer's V is reported for χ^2 tests as follows: ☆ for $V \leq 0.2$, ☆☆ for $V \in (0.2, 0.6]$, and ☆ ☆ ☆ for $V > 0.6$. Odds ratio and 95% confidence intervals are reported for Fisher's exact test, annotated by "†."

Variable	**value**	**Pairs of Treatments**		
		HEF-0, HEF-1	HEF-0, HEF-2	HEF-1, HEF-2
All scenarios		V : ☆	V : ☆	ns
Optimal Choice	`stay-is-opt`	ns	ns	ns
	`swap-is-opt`	N/A	N/A	ns
Instance Size	`small`	V : ☆☆	V : ☆☆	ns
	`large`	ns	ns	ns
Balance	`balanced`	ns	ns	ns
	`unbalanced`	V : ☆	V : ☆☆	ns
Repeated scenario (S7)		V : ☆☆	ns	ns

Table 11. Effect size measured by Cramer's V for χ^2 tests corresponding with Table 6, under different pairs of treatments and adjusting for different variables. Not significant tests are labelled as ns.
Key: (ns : $p > 0.05$) (☆ : $V \leq 0.2$), (☆☆ : $p \in (0.2, 0.6]$), (☆ ☆ ☆ : $p > 0.6$).

Variable	**Instance Size**	**Pairs of Treatments**		
		sEF1, EF1	sEF1, HEF-k	EF1, HEF-k
All scenarios	`small`	ns	d : ☆☆	d : ☆
	`large`	d : ☆☆	d : ☆☆	d : ☆☆
`stay-is-opt`	`small`	N/A	d : ☆	N/A
	`large`	ns	d : ☆☆	ns
Variable		**Pairs of Treatments**		
		HEF-0, HEF-1	HEF-0, HEF-2	HEF-1, HEF-2
All scenarios		ns	d : ☆	ns

Table 12. Effect size measured by Cohen's d for Welch t-tests corresponding with Table 7, under different pairs of treatments and adjusting for different variables. Not significant tests are labelled as ns.
Key: (☆ : $d \leq 0.3$), (☆☆ : $d \in (0.3, 0.7]$), (☆ ☆ ☆ : $d > 0.7$).

Variable	**Instance Size**	**Pairs of Treatments**		
		sEF1, EF1	sEF1, HEF-k	EF1, HEF-k
All scenarios	`small`	d : ☆	d : ☆ ☆ ☆	d : ☆ ☆ ☆
	`large`	ns	d : ☆ ☆ ☆	d : ☆ ☆ ☆

Table 13. Effect size measured by Cohen's d for Welch t-tests corresponding with Table 8, under different pairs of treatments and adjusting for different variables. Not significant tests are labelled as ns.
Key: (mp☆ : $d \leq 0.3$), (☆☆ : $d \in (0.3, 0.7]$), (☆ ☆ ☆ : $d > 0.7$).