Dynamic Preferences Elicitation Methods EXTENDED ABSTRACT

Guy Barokas* Dotan Persitz†

This version: September 13, 2025

Abstract

Eliciting individuals' preferences is vital for testing decision-making theories and informing policymakers and businesses. Traditional static methods, such as exhaustive pairwise comparisons, are often impractical and susceptible to intransitivities. We propose introducing dynamic methods that utilize information as it unfolds, offering significant benefits for experimentalists in Economics and Psychology. The design follows a between-subject "horse race" structure of ranking an even number of alternatives n, with participants visiting the laboratory twice, separated by one week. During the first visit, the alternatives are divided into $\frac{n}{2}$ distinct pairs, sequentially posed to subjects. The resulting elicited partial order, $\succ^p$, is largely free from behavioral biases. In the second visit, participants are randomly assigned a method. Each method k produces a complete strict order, $\succ^k$. We evaluate methods based on the consistency of $\succ^k$ with $\succ^p$, assuming that greater consistency indicates fewer biases. So far we executed a between-subject hypothetical online experiment comparing seven static and dynamic methods using "neutral" items (WYSIWYG, non-numeric, comparable in terms of price and utility) with 843 American subjects on Prolific. Our findings suggest that the "Removing the Best" method, which involves sequential selection of the best alternative in the set, offers a promising dynamic alternative to the static "All Pairwise" method (exhaustive pairwise comparisons), particularly for sets of alternatives that are not very small. Next, we aim to refine dynamic elicitation methods by focusing on participant heterogeneity and well-studied domains, such as risk attitudes and social preferences.

1 Introduction

Eliciting individuals' preferences is fundamental for testing decision-making theories in the laboratory and for enabling policymakers and businesses to better understand relevant individual behavior. Since the 1970s, substantial efforts have been made to study individual preferences within laboratory settings. Most of this literature employs experimental preference elicitation methods in which subjects are presented with a series of choice problems to reveal their preference ordering within a specified context.

A commonly used approach, particularly when the set of alternatives is "small enough," involves

*Economics Department, Ruppin Academic Center, Emek Hafer, Israel (e-mail: guyb@ruppin.ac.il)
†Coller School of Management, Tel Aviv University, Tel Aviv (e-mail: persitzd@post.tau.ac.il).

asking subjects to make all possible pairwise comparisons (e.g., Luce and Suppes (1965), Tversky and Kahneman (1992), Scholz et al. (2010), Manzini et al. (2010), Falk et al. (2018)). However, this method raises two significant challenges. First, as the number of alternatives increases, the number of required comparisons grows rapidly, following a $\frac{n(n-1)}{2}$ pattern, making it impractical for larger sets.[1] Second, participants' choices often violate transitivity, which prevents researchers from establishing a strict linear ordering to represent preferences (e.g., Loomes et al. (1991)). Although extensive literature exists on constructing a complete order from intransitive choices (e.g., Bouyssou (2004), Slutzki and Volij (2005), Apesteguia and Ballester (2015)), an alternative approach involves using elicitation methods that impose transitivity.

A major shortcoming of most current experimental designs for preference elicitation, including the method just described, is their static nature. We propose that the gradual accumulation of information about a participant's preferences can be leveraged dynamically to enhance the elicitation process. Specifically, a preference elicitation method is classified as static if each step in the sequence is determined independently of previous responses. By contrast, dynamic methods respond adaptively, utilizing previously obtained data to tailor subsequent questions, thereby improving the efficiency of preference elicitation. This study aims to introduce, evaluate, and compare various dynamic experimental preference elicitation methods to advance the design and execution of preference elicitation experiments.

Furthermore, preference elicitation methods often raise concerns related to bounded rationality. It has been well-documented that participants' choices can be influenced by order effects (e.g., Rubinstein and Salant (2006)), limited attention (e.g., Hirshleifer and Teoh (2003)), context effects (e.g., Huber et al. (1982), Simonson (1989), Bateman et al. (2007), Maltz and Rachmilevitch (2021)), history dependence (e.g., Brehm (1956), Arad (2013), Barokas (2024)), fatigue (e.g., Levav et al. (2010), Augenblick and Nicholson (2016)), and cognitive load (e.g., Iyengar and Kamenica (2010), Caplin et al. (2011), Bech et al. (2011)). Additionally, decision-making heuristics intended to simplify choices (e.g., Simon (1955), Gilovich et al. (2002), Halevy and Mayraz (2024)) may result in choices that do not accurately reflect participants' true preferences. Recognizing these issues, we have designed our proposed methods to account for, and mitigate, behavioral biases whenever possible. In cases where full control over expected biases is unattainable, we plan to measure and evaluate their impact, providing insights into their influence on preference elicitation.

Existing experimental designs have already explored dynamic preference elicitation, with much of this work, particularly in Marketing (though not exclusively), assuming some parametric utility representation and utilizing dynamic experimental designs to optimize the elicitation process. For example, Chapman et al. (2024) apply this approach to investigate loss aversion, Toubia et al. (2013) implement it in the contexts of risk and time preferences, Cavagnaro et al. (2013) use it for model selection in risk-related contexts, and Nguyen and Ricci (2017) apply it to group decision-making applications. Additionally, Halevy et al. (2018) compare parametric risk preference elicitation methods by generating real-time pairwise choices that differentiate between them.

[1] For instance, 10 items necessitate 45 comparisons, while 20 items require 190 comparisons, and so on.

A distinct non-parametric approach employs bisection search techniques (commonly referred to as "staircase tasks") to derive meaningful numerical values for behavioral parameters. For instance, Falk et al. (2018) use staircase tasks to elicit time and risk attitudes through a limited number of survey questions. Butler et al. (2014) adopt a non-parametric dynamic design known as "even swap" to directly gauge the strength of preferences between lotteries. Similarly, Gensler et al. (2012) apply adaptive pricing in laboratory-based willingness-to-pay elicitation tasks to mitigate "corner" subjects—those who consistently choose or avoid the no-purchase option. Our focus, however, is on a non-parametric approach that aims to recover the complete preference relation of subjects over a finite set of alternatives.

The overarching objective of this project is to transition dynamic preference elicitation from a tailored solution for specific experimental contexts to a standardized set of methods with well-known properties. Unlike much of the current literature, we will elicit ordinal, non-parametric preferences over goods that cannot be naturally represented by numerical values. Once the fundamental properties of these methods have been examined in a neutral context, the investigation could be carefully extended to include preferences over objects that can be expressed numerically (e.g., lotteries, temporal rewards, wealth distribution), which are of considerable interest to economists.

2 The Experimental Design

Suppose we have an even number of alternatives, denoted n, and m methods of elicitation. The n alternatives are divided into $\frac{n}{2}$ distinct pairs. These $\frac{n}{2}$ choice problems are then posed to the subject sequentially. This static sequence of pairwise choice problems allows for the elicitation of a partial ordering over the grand set of alternatives, with minimal exposure to the behavioral biases discussed previously. Denote this elicited partial order as $\succ^p$.

For each method k, denote the elicited complete strict order by $\succ^k$. Methods where a complete strict order is not guaranteed require appropriate adaptations, as demonstrated in subsequent examples. Our criteria for ranking different elicitation methods hinge on the consistency of $\succ^k$ with $\succ^p$. Specifically, method k is considered superior to method k' if the degree of consistency between $\succ^k$ and $\succ^p$ is greater than that between $\succ^{k'}$ and $\succ^p$. While we refrain from asserting that $\succ^p$ represents true preferences, we intentionally elicit this partial ordering to minimize known behavioral effects. Our primary assumption is that a strict order closer to $\succ^p$ is more likely to be free of such biases.

The general experimental design follows a straightforward between-subject "horse race" structure. Each participant visits the laboratory twice, with a one-week interval between sessions. During the first visit, we elicit $\succ^p$ by presenting the subject with $\frac{n}{2}$ pairwise choice problems involving distinct alternatives. In the second visit, participants are randomly assigned to one of the elicitation methods. After this visit, we elicit $\succ^k$ for the assigned method k. Possible order effects are controlled by diversifying and randomizing the sequences of alternative presentations in both parts.

The degree of consistency between a complete order and the partial order $\succ^p$ is naturally mea-

Figure 1: The alternatives.

sured by the number of inconsistencies, represented as an integer between zero and $\frac{n}{2}$. We denote this measure by α_k, referring to it as the "Hits index." [2]

An alternative, complementary measure is the "Rank index," denoted by β_k. Suppose that in pairwise choice problem t, the subject selected alternative a over alternative b. Denote the ranking of alternative a in $\succ^k$ as r_a and the ranking of alternative b as r_b. If $r_a > r_b$, then $\beta_k^t = 0$; otherwise, $\beta_k^t = r_b - r_a$. The rank index is defined as $\beta_k = \sum_{t=1}^{\frac{n}{2}} \beta_k^t$.[3] The rank index represents the sum, across the first visit's choice problems, of the minimal number of swaps of consecutive alternatives in the ranking of $\succ^k$, required for the new ranking to align with the corresponding first visit's choice. Thus, β_k quantifies the extent of inconsistency between $\succ^k$ and $\succ^p$.

3 Horse Race in a "Neutral Environment"

In this work, we focus on a between-subject hypothetical online laboratory experiment aimed at establishing the usefulness of various dynamic elicitation methods within a "neutral" environment. We conduct a between-subject "horse race" comparing several dynamic and static methods. We describe the environment as neutral because the set of alternatives, shown in Figure 1, is composed of "What You See is What You Get" items that are non-numeric and comparable in terms of price and utility. To introduce our approach, we first describe the competing methods. We then provide details of the experimental design, followed by a presentation of the results from our online experiment. We conclude the discussion by highlighting key findings that motivate the subsequent stages.

[2] In this measure, the complete ordering is derived after the partial set of problems. A similar measure, called hit rate, is used in prediction tasks where the predicting object is gathered before the predicted object. We chose this sequence of data collection to eliminate any method-specific effects on our benchmark.

[3] The rank index is inspired by the "Swaps index" introduced by Apesteguia and Ballester (2015). In Apesteguia and Ballester (2015), the swaps index for an observation is the number of elements in choice set A that are ranked higher than the chosen element $a \in A$ according to the preference relation $\succ$. In our setting, α_k corresponds to the swaps index. The rank index measures the number of elements ranked higher than the chosen alternative but not higher than the unchosen alternative within the complete strict order $\succ^k$.

The Competing Elicitation Methods

We focus on two static and five dynamic elicitation methods. The dynamic methods were specifically chosen to address various biases discussed previously. For example, all suggested methods are dynamically designed to tackle the issue of "past-dependent choices"; some methods reduce the number of choices participants need to make, thereby mitigating decision fatigue, while others focus on reducing the size of the choice set to alleviate issues related to choice overload and context effects. Table 1 below summarizes these considerations.

Static Methods

All Pairwise Choices: The subject is presented with all possible pairwise choice problems involving elements from A, with no specific order (a total of $\frac{n(n-1)}{2}$ choice problems).

Static Grand Set Ranking: The subject is presented with the complete set of alternatives and asked to rank them from the best to the worst. (See Bateman et al. (2007) for its use of Static Grand Set Ranking and Merino-Castello (2003), who refer to this method as Contingent Ranking.)

Dynamic Methods

Bottom Up: The subject begins with a pairwise choice between a_1 and a_2. If she chooses a_1, it is inferred that a_1 is ranked above a_2. The next pairwise choice problem involves the lowest-ranked alternative (a_2) and the next alternative in the predefined order (a_3). If a_2 is chosen, a_1 is ranked first, a_2 second, and a_3 last. If a_3 is chosen, the next comparison is between a_3 and a_1. If a_1 is chosen, a_1 remains first, a_3 second, and a_2 last; otherwise, a_3 is ranked first, a_1 second, and a_2 last. To incorporate a_k into the existing ranking of $k-1$ alternatives, the subject compares a_k with the lowest-ranked alternative. If not chosen, it is ranked in the k^{th} position. If chosen, the comparisons continue iteratively up the ranking until a_k is not chosen (or is chosen over the highest-ranked alternative). The new ranking will include a_k just below the alternative it was not chosen against (or first if it was always chosen). This process continues until all alternatives are ranked. While a total of $\frac{n(n-1)}{2}$ comparisons may be needed in the worst case, only $\frac{(n+2)(n-1)}{4}$ comparisons are needed on average.[4]

Top Down: The subject starts with a pairwise choice between a_1 and a_2. If a_1 is chosen, it is inferred to be ranked above a_2. The next pairwise choice involves the highest-ranked alternative (a_1) and the next alternative (a_3). If a_3 is chosen, it is ranked first, with a_1 second and a_2 last. If a_1 is chosen, the next comparison is between a_3 and a_2. The process continues similarly to place

[4] Proof by induction on n. For $n=2$, "Bottom Up" requires one comparison, giving an average of $1 = \frac{(2+2)(2-1)}{4}$. By induction, for $n+1$ alternatives, the average comparisons for the first n alternatives is $\frac{(n+2)(n-1)}{4}$, with an additional average of $\frac{n+1}{2}$ comparisons required for the $(n+1)^{th}$ alternative (between 1 and n additional comparisons in equal probability, $\sum_{t=1}^{n} \frac{1}{n} t = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$ comparisons on average). Thus, the average comparisons for $n+1$ alternatives is $\frac{(n+2)(n-1)}{4} + \frac{n+1}{2} = \frac{n(n+3)}{4}$.

Method	Limited Attention Context Effect	Fatigue	Past-dependent Winner	Past-dependent Loser
All Pairwise Choices	Controlled (pairwise choices)	Hard	Uncontrolled	Uncontrolled
Static Grand Set Ranking	Uncontrolled	Easy	Uncontrolled	Uncontrolled
Bottom Up	Controlled (pairwise choices)	Intermediate	Controlled	Uncontrolled
Top Down	Controlled (pairwise choices)	Intermediate	Uncontrolled	Controlled
Removing the Best	Uncontrolled	Easy	Controlled	Controlled
Removing the Worst	Uncontrolled	Easy	Controlled	Controlled
Iterative Categorization	Intermediate	Easy	Controlled	Controlled

Table 1: Properties of the elicitation methods.

a_k into the ranking, comparing it with the highest-ranked alternative first and progressing down the list until placement is determined (for a generalized version see Heckel et al. (2019)). As with Bottom Up, $\frac{n(n-1)}{2}$ comparisons may be needed in the worst case, but on average, only $\frac{(n+2)(n-1)}{4}$ comparisons are needed.

Removing the Best: The subject is presented with the entire set of alternatives and asked to select the best option. The process is repeated with the remaining alternatives until only one remains. The literature on choice overload (e.g., Iyengar and Kamenica (2010)) suggests that large sets may cause cognitive load and context effects in each step, especially at the beginning of the process.

Removing the Worst: The subject selects the worst alternative from the complete set, with the process repeating until only one alternative remains.[5]

Equal Size Iterative Categorization: The subject divides the entire set of alternatives into two equally sized (or nearly equal) subsets: "good" and "less good" alternatives. This process continues iteratively until each subset contains only one alternative.[6]

Summary

Table 1 provides information on how well each method handles various aspects of behavioral biases. Methods based on pairwise comparisons are expected to be robust against choice overload and context effects compared to methods that require choices from larger sets. We assume that fatigue

[5] An example of a removing-the-worst mechanism is found in British Conservative Party leadership elections (see Johnston (2024)).

[6] Variations of iterative categorization may impose different restrictions on subset sizes. In extreme cases, no restrictions are imposed.

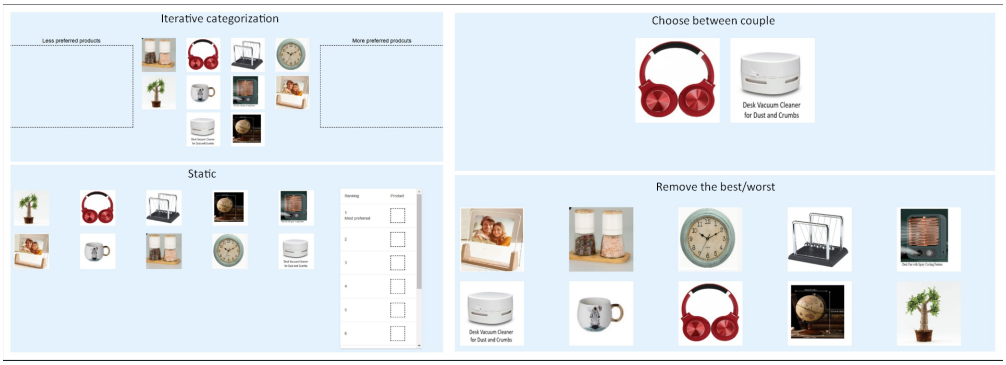

Figure 2: The main screens.

is strongly correlated with the number of choices faced by the subject. Therefore, the static pairwise choices method likely induces the highest levels of fatigue ($\frac{n(n-1)}{2}$ choices), while the Bottom Up and Top Down methods are intermediate ($\frac{(n+2)(n-1)}{4}$ choices on average), and the other methods involve at most n choices are quick to complete. Past-dependence is controlled if the subject never faces a choice between alternatives with differing histories of wins or losses (Arad (2013)). While static methods cannot mitigate this bias, dynamic methods control for at least one aspect (wins or losses). Lastly, it is important to note that the widely used static all pairwise choices method is the only method prone to intransitivities.

4 The Experiment

Between September 26, 2024, and November 8, 2024, we conducted the first-stage experiment on the online experimental platform Prolific. Figure 2 exhibits the main screens in the interface.[7] The sample consisted of 843 experienced American subjects aged 18-65.[8][9] Participants were informed that they were expected to participate in two sessions, one week apart. The experiment was hypothetical, and subjects received \$0.67 for completing the first session and \$1.83 for the second session.[10] A total of 66 subjects (7.83%) did not complete the first session, and 138 subjects (16.37%) did not return for the second session. Consequently, our final sample comprised the decisions of 639 subjects, with between 86 and 96 subjects per treatment. Table 2 provides a detailed breakdown of these numbers by treatment.

[7] A link to an app the mimics the experiment outside of Prolific can be found here.

[8] We recruited subjects who had participated in between 100 and 5,000 experiments on Prolific and had completed at least 90% of them.

[9] Due to a bug in the software implementing the Bottom Up and Top Down methods, 243 subjects from the first wave were dropped and replaced with 255 newly recruited subjects in a second wave. The numbers in this proposal refer only to the final dataset.

[10] The first session's payment was based on a rate of \$8 per hour for 5 minutes of participation, while the second session's payment was calculated at \$11 per hour for 10 minutes. Only one participant required more than the allocated time to complete the tasks.

Treatment	Quit at first stage	Didn't get to second stage	Completed the task	Perfect subjects	Average hits index	Average rank index
Remove the Best	10	16	96	31 (32.3%)	3.875	2.990
Remove the Worst	8	22	88	17 (19.3%)	3.705	3.795
Bottom Up	8	26	92	24 (26.1%)	3.783	3.196
Top Down	11	23	95	23 (24.2%)	3.758	3.463
Iterative Categorization	9	16	86	21 (24.4%)	3.535	3.767
Static Ranking	9	21	92	24 (26.1%)	3.630	3.674
All Pairwise[11]	11	14	90	31 (34.4%)	3.956	2.422
Total (843)	66 (7.83%)	138 (16.37%)	639	171 (26.8%)	3.751	3.334

Table 2: Descriptive statistics of the various treatments in the online experiment. The left-hand-side columns describe the sample and the different attrition rates by treatment. The right-hand-side describes the average performance of the subjects by treatment. Perfect subjects are those that were fully consistent between the two visits to the laboratory (hits index of five).

Table 2 also provides unconditional averages for the two consistency indices. Recall that the hits index for each subject ranges from zero to five. A score of five indicates complete consistency, meaning that every pairwise decision made during the first laboratory visit is consistent with the order elicited during the second visit. Conversely, a score of zero indicates complete inconsistency across visits. Thus, a higher hits index signifies that the partial ordering elicited during the first visit is "closer" to the complete ordering recovered in the second visit. The ranks index, on the other hand, measures the degree of inconsistencies. For subjects who are fully consistent across visits (hits index of five), the ranks index is zero (except for subjects who violated transitivity in the all-pairwise method (see Footnote 11)). For subjects displaying inconsistencies, each pairwise choice from the first visit that does not align with the ordering elicited during the second visit is assigned a value based on the number of elements separating the two alternatives in the complete ordering. Consequently, a lower ranks index indicates fewer inconsistencies and a "closer" alignment between the partial ordering elicited in the first visit and the complete ordering recovered in the second visit. Figure 3 exhibits the distribution of those indices by elicitation method.

As expected, Table 2 indicates that the "All Pairwise" method is the most effective elicitation approach. However, only 34 subjects (37.8%) satisfied transitivity, meaning that for two-thirds of the cases, researchers would need an additional method to resolve inconsistencies if a strict ordering over the alternatives is desired. It is reasonable to expect that this issue would become even more pronounced with larger sets of alternatives.

The "Removing the Best" method performs well, outperforming the other five methods. Its proportion of perfect subjects is only two percentage points lower than that of the "All Pairwise"

[11] The hits index for subjects that faced the pairwise static method counts the number of pairwise choices from the first visit to the laboratory were answered consistently in the second visit. The rank index requires that the choices made in the second visit reveal a strict linear order. However, if transitivity is violated, this is not the case. To overcome this shortcoming, following Houtman and Maks (1987), we calculated for each subject, the minimal subset of choices that should be dropped in order for the data to be consistent. After dropping those choices, a strict linear order can be formed. Note that (i) We were not able to complete this procedure for 7 subjects that require dropping at least 4 choices; (ii) The process does not guarantee uniqueness of the strict linear order. In cases there were multiple possible orders we picked one randomly.

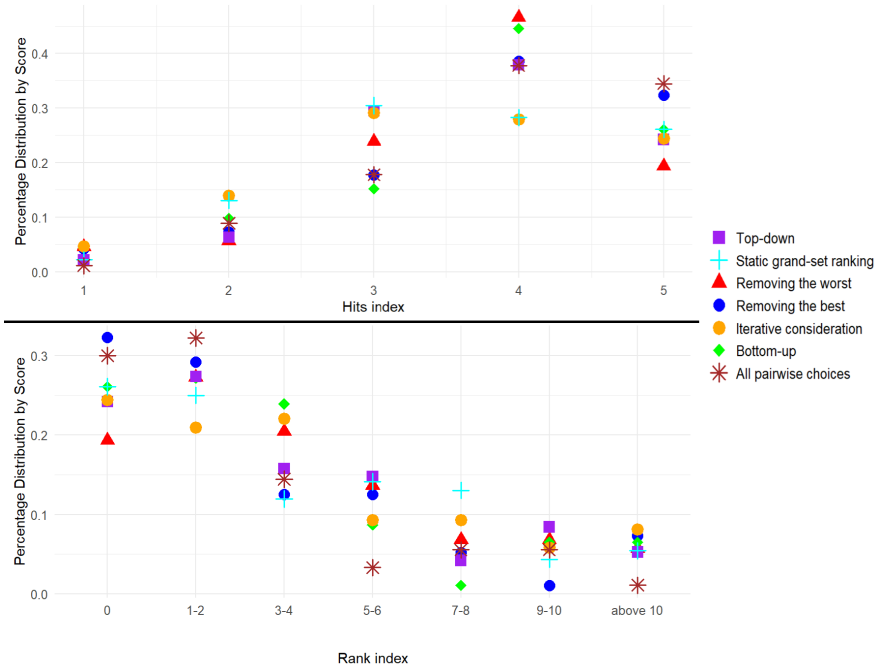

Figure 3: The upper graph represents the distribution of the hits index across treatments. The lower graph represents the distribution of the rank index across treatments.

method. While its hits index is also close, the difference in the rank index is more substantial. However, there is reason to believe that this measure may be biased downward for the "All Pairwise" method (see Footnote 11).

Table 3 presents the results of four subject-level linear regressions. In Reg(1) and Reg(2), the dependent variable is the Hits index, while Reg(3) and Reg(4) use the Rank index. We control for group allocation in the first visit, the time elapsed between visits, the relative time spent on decision-making within each method, age, gender, and whether the subject reported an ADHD diagnosis. The method dummies were coded with "All Pairwise" as the benchmark. The first and third regressions show that the consistency of choices made under the "Removing the Best" and "Bottom Up" methods is not significantly different from that of the "All Pairwise" method.

In Reg(2) and Reg(4), we added a control for the transitivity of choices made in the second visit. Since only participants assigned to the "All Pairwise" method could violate transitivity, this addition sets the benchmark as the performance of those subjects who maintained transitivity within the "All Pairwise" method. It turns out that subjects exhibiting intransitive choices in the "All Pairwise" method are significantly less consistent with their first visit choices. In the Hits index, transitive subjects average 4.206, while intransitive subjects average 3.804. In the Rank index (with seven highly intransitive subjects removed), transitive subjects average 1.559, while intransitive subjects average 3.02. Notably, in both indices, the performance of intransitive "All Pairwise" subjects (62.2% of the sample) is inferior to that of participants in the "Removing the Best" method.

Considering the high rates of intransitivity observed in the "All Pairwise" method and the

	Dependent Variable: Hits Index		*Dependent Variable:* Rank Index	
	Reg (1)	Reg (2)	Reg (3)	Reg (4)
Constant	3.415*** (0.389)	3.6248*** (0.408)	3.407** (1.361)	2.730* (1.414)
Methods				
Removing the Best	−0.072 (0.152)	−0.310 (0.207)	0.556 (0.535)	1.373* (0.712)
Removing the Worst	−0.255* (0.155)	−0.494** (0.210)	1.402** (0.545)	2.222*** (0.721)
Bottom Up	−0.158 (0.153)	−0.396* (0.208)	0.761 (0.540)	1.577** (0.716)
Top Down	−0.183 (0.152)	−0.424** (0.208)	1.005* (0.536)	1.833** (0.717)
Iterative Categorization	−0.415*** (0.156)	−0.653*** (0.210)	1.348** (0.548)	2.165*** (0.722)
Grand Set Ranking	−0.297* (0.153)	−0.535*** (0.208)	1.189** (0.540)	2.008*** (0.717)
Experimental Condition				
Group in first visit	0.079 (0.082)	0.080 (0.082)	−0.165 (0.285)	−0.166 (0.285)
Timing				
Time between visits	−0.012 (0.044)	−0.006 (0.044)	0.080 (0.150)	0.054 (0.151)
Time on second visit	0.072* (0.042)	0.077* (0.042)	−0.227 (0.153)	−0.234 (0.152)
Demographics				
Age	0.012*** (0.004)	0.011*** (0.004)	−0.028** (0.013)	−0.027** (0.013)
Gender (female = 1)	0.154* (0.084)	0.147* (0.084)	−0.509* (0.292)	−0.483* (0.292)
ADHD (not diagnosed=0, self report)	0.235** (0.108)	0.244** (0.108)	−0.730* (0.373)	−0.766** (0.373)
Transitivity				
Is transitive? (no = 1)		−0.383* (0.227)		1.388* (0.801)
R-squared	0.056	0.034	0.046	0.041
# of Observations	639	639	632	632

Table 3: Subject-level linear regressions. In Reg(1) and Reg(2), the dependent variable is the Hits index, while Reg(3) and Reg(4) use the Rank index. The methods dummy variables are coded as one if the subject encountered the particular method during their second laboratory visit and zero otherwise. The variable "Group in first visit" is a dummy distinguishing between the two potential sets of problems faced by subjects during their initial visit. "Time between visits" measures the interval between the timestamps of the first and second laboratory visits. "Time on second visit" is represented as a z-score, normalized relative to the subject's own treatment, indicating the time spent on decision-making during the second visit. The variable "ADHD" equals zero for all subjects who report being undiagnosed with ADHD (the question was "Have you ever been diagnosed with Attention Deficit Hyperactivity Disorder (ADHD)?"). "Is transitive?" is assigned a value of one if the subject's choices violated transitivity, a condition applicable only within the "All Pairwise" method.

appealing characteristics of the "Removing the Best" method—which requires only $n-1$ choice problems, eliminates past-dependence, ensures transitivity and exhibits good performance by both indices—it appears that "Removing the Best" offers a compelling dynamic alternative to the static "All Pairwise" approach, particularly for sets of alternatives that are not very small.

Two final observations regarding other methods: First, the "Bottom Up" method stands out as the second-best dynamic method, demonstrating performance comparable to that of intransitive "All Pairwise" subjects while requiring, on average, fewer than two-thirds of the choices. Second, the "Static Grand Set Ranking" method shows surprisingly weak performance, appearing inferior to all other methods except for "Iterative Categorization."

5 Next Steps

Robustness and Personalization in a "Neutral Environment"

One next step aims to deepen our investigation along two key dimensions: robustness and personalization. To enhance robustness, we will employ a pre-registered, physical, non-hypothetical within- and between-subject laboratory design. For personalization, we will explore participant heterogeneity, including variations in cognitive load management, concentration stamina, and decision-making styles. The experimental methodology will build on the general experimental design introduced earlier, closely resembling the first experiment but with two key modifications: (i) adopting a mixed design to account for individual differences among parricipants,[12] and (ii) incentivizing the experiment.

Moving into Contexts of Interest

The majority of experimental studies on individual decision-making in economics focus on domains such as risk preferences, social preferences, and time preferences. Therefore, any proposed experimental design for preference elicitation must demonstrate its effectiveness within these critical contexts.

These domains differ significantly from the neutral setting. First, alternatives within these domains can often be objectively compared based on their attributes. For example, in the risk domain, first-order stochastic dominance provides an objective criterion, whereby a lottery that dominates another is objectively preferable. Second, decision-making in these contexts is often highly sensitive to how alternatives are presented—be it textually, numerically, graphically, or otherwise. Any recommendation for an efficient elicitation method must account for these presentational nuances. Third, while not thoroughly documented, we hypothesize that the use of heuristics is more prevalent in contexts involving risk, social, and time preferences compared to the neutral environment. This increased reliance on heuristics could pose a barrier, as the extent to which individuals using heuristics reveal their true preferences remains unclear.

As noted earlier, dynamic elicitation methods have shown potential for improving preference elicitation within these domains (e.g., Toubia et al. (2013), Chapman et al. (2024), Halevy et al. (2018)). However, to the best of our knowledge, there has been no systematic investigation into dynamic, non-parametric preference elicitation methods for laboratory experiments in these areas.

[12] A mixed design involves applying a within-subject approach twice. Rather than working with a single set of n products, participants will interact with two distinct sets. During their first laboratory visit, they will make $\frac{n}{2}$ pairwise choices between $\frac{n}{2}$ distinct pairs of products for each set, completing this task twice—once for each product set. In their second visit, each participant will be randomly assigned two elicitation methods, one for each product set. The order of methods and their pairing with product sets will be randomized. This approach enables both within- and between-subject comparisons, accounting for order effects. Within-subject comparisons will allow us to control for participant heterogeneity and explore new questions related to the correlation of performance across different elicitation methods.

References

Apesteguia, Jose and Miguel A Ballester (2015) "A measure of rationality and welfare," *Journal of Political Economy*, Vol. 123, No. 6, pp. 1278–1310.

Arad, Ayala (2013) "Past decisions do affect future choices: An experimental demonstration," *Organizational Behavior and Human Decision Processes*, Vol. 121, No. 2, pp. 267–277.

Augenblick, Ned and Scott Nicholson (2016) "Ballot position, choice fatigue, and voter behaviour," *The Review of Economic Studies*, Vol. 83, No. 2, pp. 460–480.

Barokas, Guy (2024) "Positively correlated choice," *Mathematical Social Sciences*, Vol. 127, pp. 62–71.

Bateman, Ian, Brett Day, Graham Loomes, and Robert Sugden (2007) "Can ranking techniques elicit robust values?" *Journal of Risk and Uncertainty*, Vol. 34, pp. 49–66.

Bech, Mickael, Trine Kjaer, and Jørgen Lauridsen (2011) "Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment," *Health economics*, Vol. 20, No. 3, pp. 273–286.

Bouyssou, Denis (2004) "Monotonicity of 'ranking by choosing': A progress report," *Social Choice and Welfare*, Vol. 23, No. 2, pp. 249–273.

Brehm, Jack W (1956) "Postdecision changes in the desirability of alternatives.," *The Journal of Abnormal and Social Psychology*, Vol. 52, No. 3, p. 384.

Butler, David, Andrea Isoni, Graham Loomes, and Daniel Navarro-Martinez (2014) "On the measurement of strength of preference in units of money," *Economic Record*, Vol. 90, pp. 1–15.

Caplin, Andrew, Mark Dean, and Daniel Martin (2011) "Search and satisficing," *American Economic Review*, Vol. 101, No. 7, pp. 2899–2922.

Cavagnaro, Daniel R, Richard Gonzalez, Jay I Myung, and Mark A Pitt (2013) "Optimal decision stimuli for risky choice experiments: An adaptive approach," *Management science*, Vol. 59, No. 2, pp. 358–375.

Chapman, Jonathan, Erik Snowberg, Stephanie W Wang, and Colin Camerer (2024) "Looming large or seeming small? Attitudes towards losses in a representative sample," *Review of Economic Studies*, p. rdae093.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde (2018) "Global evidence on economic preferences," *The quarterly journal of economics*, Vol. 133, No. 4, pp. 1645–1692.

Gensler, Sonja, Oliver Hinz, Bernd Skiera, and Sven Theysohn (2012) "Willingness-to-pay estimation with choice-based conjoint analysis: Addressing extreme response behavior with individually adapted designs," *European Journal of Operational Research*, Vol. 219, No. 2, pp. 368–378.

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman (2002) *Heuristics and biases: The psychology of intuitive judgment*: Cambridge university press.

Halevy, Yoram, Dotan Persitz, and Lanny Zrill (2018) "Parametric recoverability of preferences," *Journal of Political Economy*, Vol. 126, No. 4, pp. 1558–1593.

Halevy, Yoram and Guy Mayraz (2024) "Identifying rule-based rationality," *Review of Economics and Statistics*, Vol. 106, No. 5, pp. 1369–1380.

Heckel, Reinhard, Nihar B Shah, Kannan Ramchandran, and Martin J Wainwright (2019) "Active ranking from pairwise comparisons and when parametric assumptions do not help," *The Annals of Statistics*, Vol. 47, No. 6, pp. 3099–3126.

Hirshleifer, David and Siew Hong Teoh (2003) "Limited attention, information disclosure, and financial reporting," *Journal of accounting and economics*, Vol. 36, No. 1-3, pp. 337–386.

Houtman, Martijn and J.A.H. Maks (1987) "The Existence of Homothetic Utility Functions Generating Dutch Consumer Data," Mimeo.

Huber, Joel, John W Payne, and Christopher Puto (1982) "Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis," *Journal of consumer research*, Vol. 9, No. 1, pp. 90–98.

Iyengar, Sheena S and Emir Kamenica (2010) "Choice proliferation, simplicity seeking, and asset allocation," *Journal of Public Economics*, Vol. 94, No. 7-8, pp. 530–539.

Johnston, Neil (2024) "Leadership elections: Conservative Party,"Technical report, House of Commons Library.

Levav, Jonathan, Mark Heitmann, Andreas Herrmann, and Sheena S Iyengar (2010) "Order in product customization decisions: Evidence from field experiments," *Journal of Political Economy*, Vol. 118, No. 2, pp. 274–299.

Loomes, Graham, Chris Starmer, and Robert Sugden (1991) "Observing violations of transitivity by experimental methods," *Econometrica: Journal of the Econometric Society*, pp. 425–439.

Luce, Robert Duncan and Patrick Suppes (1965) *Preference, utility, and subjective probability*: Wiley.

Maltz, Amnon and Shiran Rachmilevitch (2021) "A model of menu-dependent evaluations and comparison-aversion," *Journal of Behavioral and Experimental Economics*, Vol. 91, p. 101655.

Manzini, Paola, Marco Mariotti, and Luigi Mittone (2010) "Choosing monetary sequences: Theory and experimental evidence," *Theory and Decision*, Vol. 69, pp. 327–354.

Merino-Castello, Anna (2003) "Eliciting consumers preferences using stated preference discrete choice models: contingent ranking versus choice experiment." Unpublished manuscript.

Nguyen, Thuy Ngoc and Francesco Ricci (2017) "Dynamic elicitation of user preferences in a chat-based group recommender system," in *Proceedings of the Symposium on Applied Computing*, pp. 1685–1692.

Rubinstein, Ariel and Yuval Salant (2006) "A model of choice from lists," *Theoretical Economics*, Vol. 1, No. 1, pp. 3–17.

Scholz, Sören W, Martin Meissner, and Reinhold Decker (2010) "Measuring consumer preferences for complex products: A compositional approach basedonpaired comparisons," *Journal of Marketing Research*, Vol. 47, No. 4, pp. 685–698.

Simon, Herbert A (1955) "A behavioral model of rational choice," *The quarterly journal of economics*, pp. 99–118.

Simonson, Itamar (1989) "Choice based on reasons: The case of attraction and compromise effects," *Journal of consumer research*, Vol. 16, No. 2, pp. 158–174.

Slutzki, Giora and Oscar Volij (2005) "Ranking participants in generalized tournaments," *International Journal of Game Theory*, Vol. 33, pp. 255–270.

Toubia, Olivier, Eric Johnson, Theodoros Evgeniou, and Philippe Delquié (2013) "Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters," *Management Science*, Vol. 59, No. 3, pp. 613–640.

Tversky, Amos and Daniel Kahneman (1992) "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and uncertainty*, Vol. 5, pp. 297–323.