Dynamic Methods for Eliciting Preferences

A Comparative Study

Guy Barokas[1] Dotan Persitz[2]

[1]The Multidisciplinary Decision Research Center
Ruppin Academic Center

[2]Faculty of Management
Tel Aviv University

2nd Conference of the Decision Research Center
Ruppin Academic Center
May 2025

Framing the Talk

- Methodology of experiments in Economics.
- This work is still in its early stages.
- Today: main ideas and the findings of an initial study.
- Feedback and comments are very welcome.

Motivation

- Eliciting individuals' preferences is fundamental (is it?).
- An experimental **preference elicitation method** involves a sequence of questions, assignments or tasks designed to reveal an individual's preference ordering within some context of choice.
- A commonly used approach is to ask subjects to make **all** pairwise comparisons.
  - Analytic Hierarchy Process: Criteria importance.
  - Manzini et al. (2010): Preferences over monetary series.
  - Loomes et al. (1991): Preferences over lotteries.
- However, this method becomes impractical for "large" sets due to the growth in the number of comparisons ($\frac{n(n-1)}{2}$ comparisons).
- Additionally, violations of transitivity are often observed.

Issue I: Static Preference Elicitation Methods

- A preference elicitation method is **static** if each step in the sequence is determined independently of prior responses
- Most of the experimental literature on preference elicitation in economics relies on static methods.
- However, the gradual accumulation of information about a participant's preferences can be leveraged dynamically to significantly enhance the elicitation process.

Issue II: Bounded Rationality Concerns

- Key concerns in preference elicitation include:
  - Order effects.
  - Limited attention.
  - Context effects.
  - History-dependent choice.
  - Decision fatigue.
  - Cognitive load.
- 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 (or maybe heuristics are the fundamental?).

Existing Dynamic Preference Elicitation Methods

- Assume a functional form and dynamically optimize for the "best" parameters::
  - Marketing: Toubia et al. (2013) and Cavagnaro et al. (2013).
  - DOSE: Chapman et al. (2024).
  - Staircase tasks (originated in Psychology): Falk et al. (2018).
- A second stage, designed based on the first stage's results:
  - Elicit order in step 1, and cardinality in step 2: Butler et al. (2014).
  - Two elicitation methods use the same data in step 1, then race against each other using dynamically generated data in step 2: Halevy et al. (2018).
- Studying "corner" subjects: Gensler et al. (2012).
- Our goal is to employ a **non-parametric approach** to recover the complete preference relation of subjects over a **finite** set of alternatives.

Our Plan

- Step 1 (today): A between-subject, hypothetical online laboratory experiment to assess the usefulness of various dynamic elicitation methods in a “neutral” environment.
- Step 2: A within-subject, incentivized physical laboratory experiment.
- Step 3: An incentivized laboratory experiment focused on dynamically eliciting other-regarding and risk preferences.
- “The Complete Victory”: Provide economists with dynamic preference elicitation methods that outperform existing (probably static) techniques..

Our Experimental Approach

- Challenge: Design an experiment that allows for comparison and ranking of various elicitation methods.
- We elicit preferences in two stages:
  1. Partial ordering elicitation, where biases and bounded rationality are less likely to affect results.
  2. Various complete ordering elicitation methods.
- Very roughly - The complete ordering most consistent with the partial ordering is considered the “winner”.

Elicitation Methods

- In our first step experiment we use two static methods and five dynamic methods.
- The dynamic methods were selected specifically to address a range of behavioral biases.
- Some of these methods are based on pairwise comparisons, while others involve choice from sets.
- Let the set of alternatives be denoted $A = \{a_1, \ldots, a_n\}$, where n is assumed to be even for simplicity.

Static I: Static Pairwise Choices

- The participant faces all possible pairwise choice problems from the set A, without any predetermined order.
- This results in a total of $\frac{n(n-1)}{2}$ choice problems.
  - If $n = 10$ there are 45 problems.
  - If $n = 16$ there are 120 problems.

Static II: Static Grand Set Ranking

- Participants are presented with the complete set of alternatives and asked to rank them from best to worst.
- This method is used by Bateman et al. (2007) and is referred to as Contingent Ranking by Merino-Castello (2003).

Dynamic I: Bottom Up

- The participant begins by choosing between alternatives a_1 and a_2.
- If a_1 is chosen, it is ranked above a_2.
- The next pairwise choice problem involves the lowest-ranked alternative (a_2) and the next alternative (a_3).
- If a_2 is chosen, a_1 is ranked first, a_2 second, and a_3 last.
- However, 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.
- The process continues until all alternatives are ranked.
- In the worst case, this method requires a total of $\frac{n(n-1)}{2}$ choice problems but on average it requires only $\frac{(n+2)(n-1)}{4} = \frac{1}{2} \times \frac{n(n-1)}{2} + \frac{n-1}{2}$ comparisons.

Dynamic II: Top Down

- The participant begins by choosing between alternatives a_1 and a_2.
- If a_1 is chosen, it is ranked above a_2.
- The next pairwise choice problem involves the **highest**-ranked alternative (a_1) and the next alternative (a_3).
- If a_3 is chosen, a_3 is ranked first, a_1 second, and a_2 last.
- However, if a_1 is chosen, the next comparison is between a_3 and a_2.
- If a_2 is chosen, a_1 remains first, a_2 second, and a_3 last; otherwise, a_1 is ranked first, a_3 second, and a_2 last.
- The process continues until all alternatives are ranked.
- For a generalized version see Heckel et al. (2019).
- Worst-case and average number of comparisons similar to Bottom-Up.

Dynamic III: Removing the Best

- The participant is presented with the complete set of alternatives and asked to select the best option.
- The process is repeated with the remaining alternatives until only one option remains.

Dynamic IV: Removing the Worst

- The participant is presented with the complete set of alternatives and asked to select the **worst** option.
- The process is repeated with the remaining alternatives until only one option remains.
- An example of the “removing the worst” mechanism is found in the British Conservative Party leadership elections (Johnston (2024)).

Dynamic V: Equal Size Iterative Categorization

- The subject divides the set of alternatives into two equally sized (or nearly equal) subsets: “good” and ”less good”.
- This categorization process continues iteratively until each subset contains only one alternative.
- Different variations of iterative categorization may impose other restrictions on subset sizes, or, in some cases, no restrictions at all.

Performance across Key Concerns

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

- All methods guarantee completeness.
- All methods guarantee transitivity (excluding the All Pairwise Choices).

Partial Order Elicitation

- The n alternatives are divided into $\frac{n}{2}$ distinct pairs.
- These pairs are presented to the participant sequentially to elicit a partial preference order.
- This method eliminates concerns of fatigue, past-dependence, or intransitivity.
- We denote the resulting partial preference order as $\succ^p$.

The Criterion: General Discussion

- For every method k, denote the elicited complete order as $\succ^k$ (Static I requires special consideration).
- We measure consistency between $\succ^k$ with $\succ^p$.
- Method k is considered superior to method k' if the consistency between $\succ^k$ and $\succ^p$ is greater than the consistency between $\succ^{k'}$ and $\succ^p$.
- Our primary assumption is that a complete order closer to $\succ^p$ is less likely to be biased.
- We are reluctant to claim that $\succ^p$ is taken as a benchmark for the "correct preferences" since these remain unobserved (recall the heuristic concern).

The Criterion: Indices

- The "Hits index" (α_k) counts the number of inconsistencies between $\succ^k$ and $\succ^p$ (an integer between zero and $\frac{n}{2}$).
- The "Rank index" (β_k), inspired by the "Swaps index" (Apesteguia and Ballester (2015)), measures the difference in rankings between alternatives.
  - Suppose that in the pairwise choice problem t in the first visit, the subject selected alternative a_i over alternative a_j.
  - Denote the ranking of alternative a_i in $\succ^k$ as r_{a_i} and the ranking of alternative a_j as r_{a_j}.
  - If $r_{a_j} > r_{a_i}$, then $\beta_k^t = 0$; otherwise, $\beta_k^t = r_{a_j} - r_{a_i}$.
  - The rank index is defined as $\beta_k = \sum_{t=1}^{\frac{n}{2}} \beta_k^t$.
- A consistent subject achieves a high "Hits index" and a low "Rank index".

Procedures

- Each participant completes two laboratory visits, with a one-week interval between sessions.
- In the first visit, we elicit the partial preference order $\succ^p$.
- In the second visit, participants are randomly assigned one of the elicitation methods, and we elicit their complete preference order $\succ^k$ for the assigned method k.
- We randomize the order of alternative presentations in both visits.

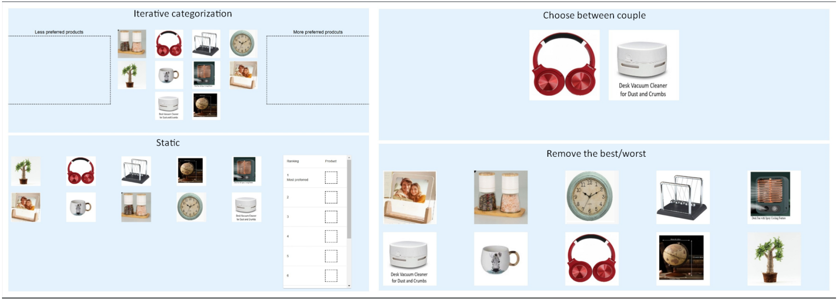
Alternatives

The alternatives used in both the first and second steps consist of non-numeric items, which are comparable in price and utility (“What You See is What You Get”).

Specifics

- Between September 26, 2024, and November 8, 2024.
- The online experimental platform Prolific.
- The sample consisted of 843 experienced American subjects, aged 18-65.
- Participants were informed that they were required to complete two sessions, one week apart.
- The experiment was hypothetical. Participants received $0.67 for completing the first session and $1.83 for the second. Only one participant took more than 5 minutes for the first session or 10 minutes for the second.
- A total of 204 subjects (24%) did not complete both sessions, leaving 639 subjects in the final sample, with between 86 and 96 participants per treatment.

Screens


Iterative categorization
Less preferred products
More preferred products
Static
Ranking
Product
1
Most preferred
2
3
4
5
6
Choose between couple
Desk Vacuum Cleaner
for Dust and Crumbs
Remove the best/worst
Desk Vacuum Cleaner
for Dust and Crumbs

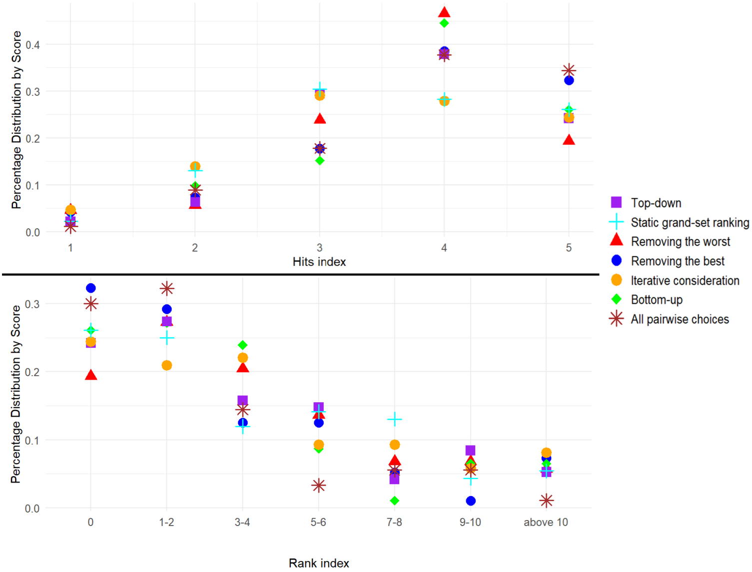
Results I

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	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

For subjects that faced the pairwise static method:

- The hits index counts the number of pairwise choices from the first visit that were answered consistently in the second visit.
- The rank index is calculated relative to the largest subset of choices that satisfies transitivity (7 subjects that require dropping at least 4 choices were dropped). The process does not guarantee uniqueness.

Results II


Percentage Distribution by Score
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
Hits index
Percentage Distribution by Score
0.0
0.1
0.2
0.3
0
1-2
3-4
5-6
7-8
9-10
above 10
Rank index
Top-down
Static grand-set ranking
Removing the worst
Removing the best
Iterative consideration
Bottom-up
All pairwise choices

Results III

	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

- The reference group in the odd regressions is all “All Pairwise” subjects while the reference group in the even regressions are the “All Pairwise” subjects that satisfy transitivity.
- “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.

Findings

1. As expected, the “All Pairwise” method is the most effective elicitation method.
2. However, only 34 subjects (37.8%) satisfied transitivity.
3. Moreover, transitive participants performed significantly better than those who were intransitive.
4. The “Removing the Best” method showed near-identical performance to the “All Pairwise” and outperformed all other methods.
5. Even compared to the transitive “All Pairwise” subjects, the difference remains insignificant (by the Hits Index).
6. The pairwise-based “Bottom Up” emerged as the second-best dynamic method, performing at par with the intransitive “All Pairwise” subjects.
7. The “Static Grand Set Ranking” method shows surprisingly weak performance, appearing inferior to all other methods except for “Iterative Categorization.”

Next Step

- Robustness:
  - Pre-registered.
  - Physical.
  - Non-hypothetical.
  - Within- and between-subject design.
- The second stage will introduce two key differences: a mixed design and the use of incentives.
- The goal of this stage is to provide a recommended experimental design for researchers aiming to elicit preferences over moderately sized, neutral sets of objects.

Mixed Design

- The mixed design applies a within-subject approach across K distinct sets of alternatives (and K methods).
- In the first laboratory visit, participants will make $\frac{n}{2}$ pairwise choices between $\frac{n}{2}$ distinct pairs of products for each set.
- In their second visit, each participant will be randomly assigned K elicitation methods, one for each product set.
- 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.

Incentives

- In the second visit, one random choice problem will be selected for implementation.
  - This is straightforward for pairwise-based methods and “removing the best”.
  - For Iterative Categorization, the reward will be chosen uniformly from the preferred set. Similarly, in “Remove the worst” (from the non-chosen alternatives).
  - For Static Grand Set Ranking, a non-uniform lottery is used, with higher chances for alternatives ranked higher.
- Incentivizing the first visit is more challenging, as we want participants to view the second visit as an independent experiment.

Third and Final Step

- The final step involves applying the methods to key economic domains, such as risk preferences, social preferences, and time preferences.
- These domains differ significantly from the neutral settings used in the earlier stages:
    - Objective comparison is frequently possible (e.g. first-order stochastic dominance in the risk domain).
    - Decision-making in these contexts is often highly sensitive to how alternatives are presented.
    - The alternatives are more suitable for numeric representation, hence heuristics may be more prevalent.
- We plan to begin with social preferences: the outcomes of a dictator game.

Personalization

- An elicitation method is considered **impersonal** all participants follow the same sequence of tasks.
- So far, we found the method that is best "on average" - Can we classify subjects by their most compatible elicitation method?
- Classification: Well-known surveys and tasks or testing the subject on various methods using a set of alternatives that has a natural ranking.
- Two reasons to be optimistic:
  1. Demographics are important.
  2. "All Pairwise" suits best only those that are consistent across visits.
- We plan to pilot this approach as part of the second step.

Conclusion

- The overarching objective of this project is to transition dynamic preference elicitation from a tailored solution in specific experimental contexts to a standardized set of methods with well-known properties.
- We believe that introducing easy-to-implement dynamic methods, which utilize information as it is gradually revealed, will be highly beneficial for experimentalists.
- In the broader context, we argue that advances in technology allow for the easy integration of dynamic and personalized elements into experimental design.
- Dynamic versus static and personalized versus impersonal should be two design questions as important as, for example, between or within subject.

Our First “Recommendation” (USE WITH CAUTION)

When conducting experiments with neutral alternatives 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.

Thanks


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
BİYAN
SHUKRIA
TINGKI
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
KOMAPSUMNIDA
GRAZIE
MEHRBANI
MAAKE
PALDIES
BOLZİN
MERCI