Intro to Conjoint Experiments
Session 3

Alberto Stefanelli

Session 2: Recap

1. Different CJ designs
   - Rating
   - Ranking
   - Choice (most used)
2. Terminology
   - Attributes: characteristics of the profile
   - Levels: characteristics of the different attributes
3. Garbage in, garbage out
   - Clarity in the attributes/levels
   - Design Complexity
   - Fatigue
   - Cognitive Burden

Session 3: Outline

1. Different CJ Randomization
   - Fully randomized uniform design
   - Randomized weighted design
   - Restricted Randomization (or nested design)
2. Assumptions
   - SUTVA
   - No profile-order effects
   - Randomization of the profiles
3. Designing a survey
   - Designing good questions
   - Response options and placement
   - Motivate respondents
   - Get feedback and pre-test
4. On-line Data Collection
   - Advantages/Disadvantages
   - Solutions (attention checks, IP checks, incentives)

Materials

- Lecture's PDF
- Lab
- Exercise

Where to find the material:

- On my GitHub/conjoint_class

Before starting

- Make sure you have access to the Qualtrics account.
- **If you have questions, shoot :)**

Section 1

Profiles construction

Profiles construction

1. Full Factorial
   1. All possible combinations of levels
2. Fully randomized design
   1. **Fully randomized uniform design**
      1. Equal probability of all levels in a given attribute
      2. E.g., Female 33%, Male 33%, Non-Binary 33%
   2. **Randomized weighted design**
      1. Unequal probability of certain levels in a given attribute
      2. Marginal distribution of levels departs from uniform distribution
      3. E.g. Female 50% and Male 49%, Non-Binary 1%
   3. **Restricted Randomization (or nested design)**
      1. Certain combinations or attributes are not allowed to happen
      2. E.g., Doctor without academic degree
3. Orthogonal and fractional factorials
4. and many more (!!)
   - Block Designs
   - Optimal Designs

Full Factorial Design (1)

1. Full Factorial
   1. Preferable design
   2. Shows every attributes level with every other attribute level
   3. All the combination are included in the design
   4. It grows geometrically with each attribute
   5. Potentially we would like to assign respondents (n=30) to each combination

2. Example
   1. Age= 30s, 40s, 50s
   2. Education= Elementary School, High School, Academic Degree
   3. Years in Congress= More than 5, between 5 and 10, more than 10
   4. Q: How many combinations?
   5. Q: What would be a reasonable sample size?

Full Factorial Design (2)

Age	Education	Job	Combinations
30s	Elementary	High Skilled	1
40s	Elementary	High Skilled	2
50s	Elementary	High Skilled	3
30s	High School	High Skilled	4
40s	High School	High Skilled	5
50s	High School	High Skilled	6
30s	Academic	High Skilled	7
40s	Academic	High Skilled	8
50s	Academic	High Skilled	9
30s	Elementary	Medium Skilled	10
40s	Elementary	Medium Skilled	11
50s	Elementary	Medium Skilled	12
30s	High School	Medium Skilled	13
40s	High School	Medium Skilled	14
50s	High School	Medium Skilled	15
30s	Academic	Medium Skilled	16
40s	Academic	Medium Skilled	17
50s	Academic	Medium Skilled	18
30s	Elementary	Low Skilled	19
40s	Elementary	Low Skilled	20
50s	Elementary	Low Skilled	21
30s	High School	Low Skilled	22
40s	High School	Low Skilled	23
50s	High School	Low Skilled	24
30s	Academic	Low Skilled	25
40s	Academic	Low Skilled	26
50s	Academic	Low Skilled	27

Full Factorial Design (3)

1. In fact, we have many more combinations
2. Q: Why is it the case?
3. We can calculate all the combination using the formula
4. L is the number of levels, M the number of alternatives, and A the number of attributes

$$Combinations = L^{M^A} = 3^{2^3} = 729$$

5. The number of alternative is thus 729
6. If we assign approximately 20 respondents per each combination
7. $729 * 20 = 14000$

Section 2

Randomized design

Randomized design

1. Developed by Hainmueller, Hopkins, and Yamamoto (2014) relaying upon the potential outcome framework
2. **Logic:** sampling each level independently such that the attributes are orthogonal
3. That is, it samples n number of rows for the full factorial

Randomized weighted design (1)

1. Level sampling performed in the same way as the fully randomized design
2. **However**, we can control the probability of a given level to appear in the CJ table
3. That is, we specify a marginal distribution for certain levels
4. The marginal distribution should correspond to the one found in the target population
5. Some levels appears more, some appears less
6. This increase the **external validity** of the CJ experiment
7. Q: Why is this the case?

Randomized weighted design (2)
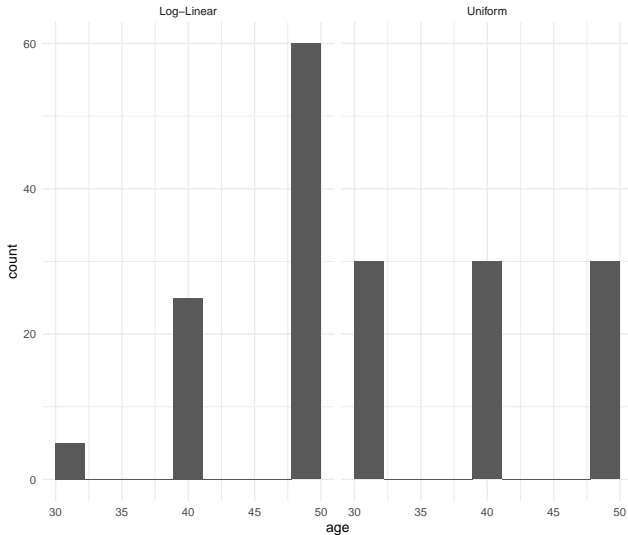
Restricted Randomization (or nested design) (1)

1. Sampling performed in the same way as the fully randomized design
2. **BUT** we impose limitation on the combination of certain attributes within a profile.
3. Technically, we set the conditional probability of the co-occurrence of two attributes/levels equal to zero
4. Meaning, we deny that two (or more) unlikely or contrasting combination of attributes/levels from appearing together
5. This improves the **internal validity** of our CJ experiment
6. Q: Why is this the case?

Assumptions (fully randomized designs)

1. Stability and no-carry over (aka SUTVA)
   - Respondents DO NOT use the information from previous questions and/or tasks in evaluation the profiles
   - If two profiles in two subsequent tasks have identical attributes, the respondents would choose the same profile.
2. No profile-order effects:
   - The ordering of profiles within a choice task does not affect responses
   - Meaning that it is possible to pool information across profiles when estimating causal quantities of interest
3. Randomization of the profiles
   - The potential choice can never be systematically related to what profiles they will actually see in the experiment.
   - Individual choices are statistically independent of the shown profiles

Section 3

Designing a questionnaire

Asking Effective Questions

- Easiest language possible yet appropriate usage of the langue (e.g., "Hookers on the streets are a threat to public safety")
- To the point (no creative writing)
- No double negatives (e.g., "Did you not use drugs in the past week?")
- No double-barrel question.s (e.g., "Do you support the legalization of street drugs and their taxation?")
- Watch out for cultural specific words or phrases (e.g., "hustler")
- Respondents have knowledge about the question's topic (e.g. "Italy has good policies regarding sex work?")
- Respondents have experience with whatever events, behaviours, or feelings you are asking them to report (i.e. feelings of nostalgia)
- Considering adding a filter question if necessary

Response options

1. Multi-choice options (e.g. Current relationship status: Married, Widowed, Divorced, Separated. . .)
2. Ordinal items
   - Distance between each point on the scale is not necessarily consistent (i.e., equidistant)
   - e.g., Multiple times a week, Twice a day, Once every day, Every other day, Less than once per week
3. Likert-Type Response

Response options: Likert-Type Response

1. Use ascending order of Likert response options (i.e. from SD to SA) (Friedman, Herskovitz, and Pollack 1993; Hartley and Betts 2010)
2. When using indices, use bidirectional response scales (half SD to SA, and half SA to SD) (Barnette 2000)
3. Include a mid-point (i.e. Neither agree nor disagree) (Nadler, Weston, and Voyles 2015)
   - *DO NOT* treat the mid-point as NA
4. Include both "I don't know" and "Not applicable" (especially for social desirable questions) (Chyung et al. 2017)
5. No formatting (italic or bold) of response scales (Hartley and Betts 2010)
6. When questions are similar, use a matrix to presenting multiple questions

Example of a Matrix-grid

	Very unsatisfied	Unsatisfied	Neither satisfied nor dissatisfied	Satisfied	Very satisfied
Reliability	○	○	○	○	○
Size	○	○	○	○	○
Ease of use	○	○	○	○	○
Portability	○	○	○	○	○

Placement and randomization

- Start with couple easy questions to allow the respondent to get familiar with the survey platform (e.g., age, gender)
- Organize the different questions in blocks (intro, demographics, moderators, experiment 1. . . N, conclusion)
- Randomize the order of the question within each block
- If a matrix is used, randomize the order of the questions within the matrix
- Place moderators ALWAYS before any experimental treatment
- Demographic can be asked at the end of the survey

Good practice

- Motivate respondents
    - Explain the purpose of the survey
    - Give back to the respondents:
        - Incentives
        - Relevant info on the topic
    - Establish trust
- Informed consent
    - Always mention the PI
    - Explain how the data will be used
    - Respondent should be 18+ (unless specific RQ)
    - Respondents can withdraw at **any** time

Fielding a survey

1. Imagine how the participants would feel responding to the questions.
2. Get feedback, especially from people who resemble those in the researcher's sample.
3. Consider performing cognitive interviews.
4. ALWAYS pre-test your instrument and survey

Section 4

Online Data Collection

Advantages

1. Relatively cheap
   - Face-to-face paper-and-pencil: ~120 euro
   - On-line probability panel: ~40 euro
2. Relatively fast
3. Very flexible
4. Far more control over experimental design

Disavantages

1. Higher "sampling error": some people are more difficult to reach than others.
2. Higher "non-response bias": some groups of people to be less likely to agree to be interviewed than others.
3. Lower data quality (Kennedy et al. 2020)
   - Participants aren't who they say they are (e.g. complete studies they wouldn't be eligible for)
   - Participants are more inattentive
   - "Server farmers" who complete studies multiple times
   - Low effort responding (it can lead to spurious correlations, Zorowitz, Niv, and Bennett (2021))

Solutions (1)

1. Choose an appropriate platform
   - If you can, use an high quality data collector (e.g., YouGov) and hold them accountable.
   - If you cannot, Eyal et al. (2021) compared (among other platforms) Prolific, Qulatrics, MTurk, and CloudResearch.
   - With platform filters, CloudResearch (MTurk toolkit) and Prolific performed equitably.

2. Put server farmer checks
   - Use an IP address detector (Waggoner, Kennedy, and Clifford 2019)
   - Language proficiency test (Gadiraju et al. 2015)
   - Cultural checks
   - Detailed open-text responses

Solutions (2)

3. Low effort respondents
    - Instructional (Oppenheimer, Meyvis, and Davidenko 2009) or nonsensical (Paolacci 2010) attention checks
    - Use multiple attention checks: 2 attentions every 5 minutes (Agley et al. 2022)
    - Comprehension checks (for a good example see, Gordon et al. 2019)
    - Bots checks (Agley et al. 2022)
    - Inspect response times (remember the first lab?)

4. Incentives
    - Pay as well as you can and consider doing a lottery (LaRose and Tsai 2014; Cobanoglu and Cobanoglu 2003)
    - Make your studies as interesting as possible
    - Be kind and respectful in participant interactions

5. Transparency (Munafò et al. 2017)
    - Register the experimental design
    - Register the quality controls
    - Register the exclusions based on these controls
    - Register the analytical approach

Session 3: Recap

1. Different CJ Randomization
   - Fully randomized uniform design
   - Randomized weighted design
   - Restricted Randomization (or nested design)

2. Assumptions
   - SUTVA
   - No profile-order effects
   - Randomization of the profiles

3. Designing a survey
   - Designing good questions
   - Response options and placement
   - Motivate respondents
   - Get feedback and pre-test

4. On-line Data Collection
   - Advantages/Disadvantages
   - Solutions (attention checks, IP checks, incentives)

References I

Agley, Jon, Yunyu Xiao, Rachael Nolan, and Lilian Golzarri-Arroyo. 2022. "Quality Control Questions on Amazon's Mechanical Turk (MTurk): A Randomized Trial of Impact on the USAUDIT, PHQ-9, and GAD-7." *Behavior Research Methods* 54 (2): 885–97. https://doi.org/10.3758/s13428-021-01665-8.

Barnette, J. Jackson. 2000. "Effects of Stem and Likert Response Option Reversals on Survey Internal Consistency: If You Feel the Need, There Is a Better Alternative to Using Those Negatively Worded Stems." *Educational and Psychological Measurement* 60 (3): 361–70. https://doi.org/10.1177/00131640021970592.

Chyung, Seung Youn (Yonnie), Katherine Roberts, Ieva Swanson, and Andrea Hankinson. 2017. "Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale." *Performance Improvement* 56 (10): 15–23. https://doi.org/10.1002/pfi.21727.

Cobanoglu, Cihan, and Nesrin Cobanoglu. 2003. "The Effect of Incentives in Web Surveys: Application and Ethical Considerations." *International Journal of Market Research* 45 (4): 1–13. https://doi.org/10.1177/147078530304500406.

References II

Eyal, Peer, Rothschild David, Gordon Andrew, Evernden Zak, and Damer
Ekaterina. 2021. "Data Quality of Platforms and Panels for Online Behavioral
Research." *Behavior Research Methods*, September.
https://doi.org/10.3758/s13428-021-01694-3.

Friedman, Hershey H, Paul J Herskovitz, and Simcha Pollack. 1993. "The Biasing
Effect of Scale-Checking Styles on Response to a Likert Scale," 4.

Gadiraju, Ujwal, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015.
"Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of
Online Surveys." In *Proceedings of the 33rd Annual ACM Conference on
Human Factors in Computing Systems*, 1631–40. CHI '15. New York, NY,
USA: Association for Computing Machinery.
https://doi.org/10.1145/2702123.2702443.

Gordon, Andrew, Susanne Quadflieg, Jonathan C. W. Brooks, Ullrich K. H. Ecker,
and Stephan Lewandowsky. 2019. "Keeping Track of 'Alternative Facts': The
Neural Correlates of Processing Misinformation Corrections." *NeuroImage* 193
(June): 46–56. https://doi.org/10.1016/j.neuroimage.2019.03.014.

References III

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22 (1): 1–30. https://doi.org/f5qzwp.

Hartley, James, and Lucy R. Betts. 2010. "Four Layouts and a Finding: The Effects of Changes in the Order of the Verbal Labels and Numerical Values on Likert-Type Scales." *International Journal of Social Research Methodology* 13 (1): 17–27. https://doi.org/10.1080/13645570802648077.

Kennedy, Ryan, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. 2020. "The Shape of and Solutions to the MTurk Quality Crisis." *Political Science Research and Methods* 8 (4): 614–29. https://doi.org/10.1017/psrm.2020.6.

LaRose, Robert, and Hsin-yi Sandy Tsai. 2014. "Completion Rates and Non-Response Error in Online Surveys: Comparing Sweepstakes and Pre-Paid Cash Incentives in Studies of Online Behavior." *Computers in Human Behavior* 34 (May): 110–19. https://doi.org/10.1016/j.chb.2014.01.017.

References IV

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (1): 1–9. https://doi.org/10.1038/s41562-016-0021.

Nadler, Joel T., Rebecca Weston, and Elora C. Voyles. 2015. "Stuck in the Middle: The Use and Interpretation of Mid-Points in Items on Questionnaires." *The Journal of General Psychology* 142 (2): 71–89. https://doi.org/10.1080/00221309.2014.994590.

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko. 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45 (4): 867–72. https://doi.org/10.1016/j.jesp.2009.03.009.

Paolacci, Gabriele. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5 (5): 9.

References V

Waggoner, Philip, Ryan Kennedy, and Scott Clifford. 2019. "Detecting Fraud in Online Surveys by Tracing, Scoring, and Visualizing IP Addresses." *Journal of Open Source Software* 4 (37): 1285. https://doi.org/10.21105/joss.01285.

Zorowitz, Samuel, Yael Niv, and Daniel Bennett. 2021. "Inattentive Responding Can Induce Spurious Associations Between Task Behavior and Symptom Measures." PsyArXiv. https://doi.org/10.31234/osf.io/rynhk.