Factorial Survey Designs

Carsten Sauer
Zeppelin University, Friedrichshafen

Social Science Data Lab, MZES
09.12.2020

Idea of Factorial Surveys

Experiments and Surveys

Lab experiments

+ Control of unobserved influences
+ Precision of measurement
– Artificial situation
– Selectivity of samples

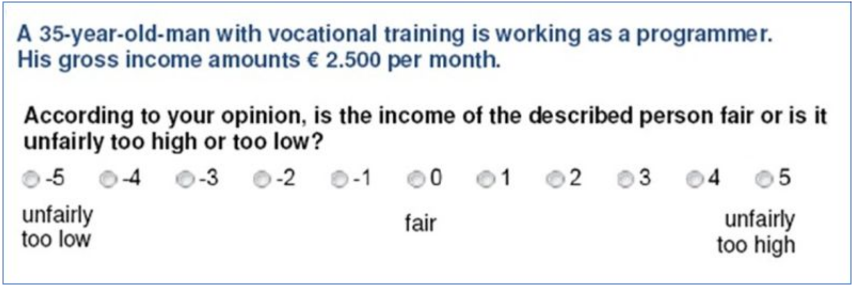
Surveys

+ Sampling based on sampling plan
+ Heterogeneous samples realizable
– Unobserved heterogeneity
– No actual behavior

- **Survey-based experiments**
  - Control of unobserved influences
  - Sampling based on a sampling plan
- **Factorial survey experiments**
  - Multi-factorial design
  - Simultaneous variation of attributes
  - Multi-causal relationships can be modeled

Idea and Goals

- Respondents evaluate short descriptions of objects or situations (*vignettes*)
- Within these vignettes attributes (*dimensions*) experimentally vary in their *levels*
- *Within-subject design*: Each respondent evaluates several vignettes
- *Between-subjects design*: Each respondent evaluates one vignette


A 35-year-old-man with vocational training is working as a programmer.
His gross income amounts € 2.500 per month.

According to your opinion, is the income of the described person fair or is it unfairly too high or too low?

-5 -4 -3 -2 -1 0 1 2 3 4 5

unfairly too low — fair — unfairly too high

Idea and Goals

- Testing theoretical models with, e.g., additive, multiplicative, etc. relations.
  - $Y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_3 + ...$
  - $Y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times (x_1 \times x_2) + ...$
  - $SEU = \beta_1(q \times B) - \beta_2(p \times C)$
    with SEU = subjective expected utility; q and p = probabilities;
    B = benefit; C = costs
- Measuring influence of:
  - Single dimensions on evaluations (e.g., attitudes, decisions)
  - Interactions between dimensions
  - Respondents‘ characteristics on vignette evaluations
- Less appropriate for exploratory analyses!

Example: Just Gender Pay Gap

- Observation: Persisting gender gaps in pay
- Caused by discrimination/underlying norms? Under which conditions? How do justice/fairness perceptions differ?
- Fairness evaluations of earnings of female and male vignette persons with the same characteristics (Auspurg et al. 2017)
- Different evaluations caused by gender-specific evaluations ("discrimination")

A 40-year-old **man** with university degree is working as a programmer. Her monthly gross earnings total 2,500 euros.

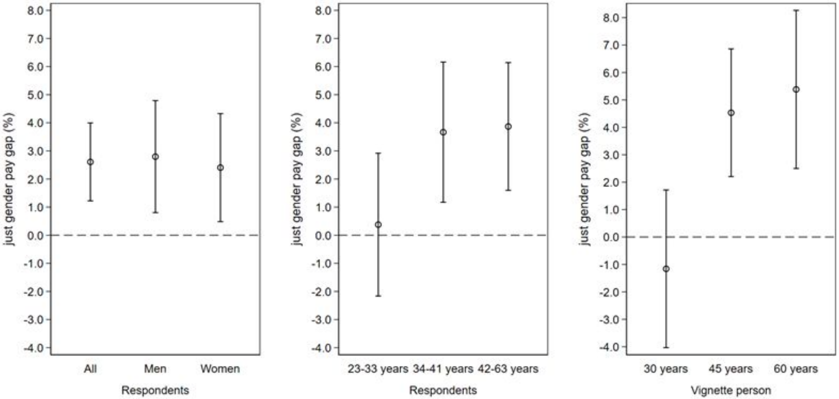
Are the earnings of this person fair or are they, from your point of view, unfairly high or low?

Unfairly low					Fair					Unfairly high
-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

A 40-year-old **woman** with university degree is working as a programmer. Her monthly gross earnings total 2,500 euros.

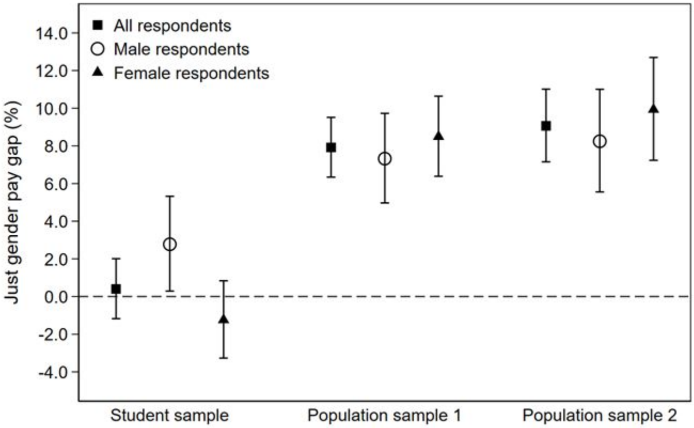
Are the earnings of this person fair or are they, from your point of view, unfairly high or low?

Unfairly low					Fair					Unfairly high
-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐



just gender pay gap (%)
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
All
Men
Women
Respondents
23-33 years
34-41 years
42-63 years
Respondents
30 years
45 years
60 years
Vignette person

Example: Just Gender Pay Gap (Sauer 2020)



Construction of Factorial Surveys

- Theoretically driven selection of dimensions and levels
- Calculation of all possible combinations (vignette universe/full factorial)
  - Orthogonal crossing of all dimensions $\longrightarrow$ Cartesian product of levels. For example, 4 dimensions with 3, 3, 4, and 4 levels: $3 \times 3 \times 4 \times 4 = 3^2 \times 4^2 = 9 \times 16 = 144$ vignettes
- Trade-offs between methodological decisions, e.g.:

  \+ More dimensions and levels $\longrightarrow$ more information on judgment principles

  – Higher cognitive burden for respondents

  – Bigger vignette universe: complex sampling or more respondents?

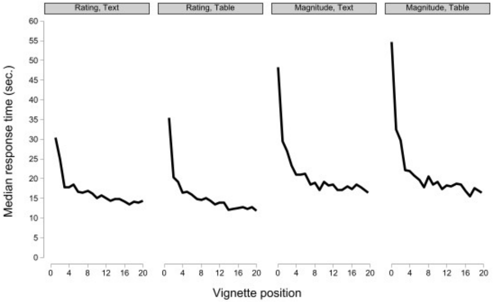
Complexity

- Number of dimensions: about 5-9 dimensions
  - Even more complex vignettes get evaluated (e.g.: 30 vignettes with 12 dimensions: very few missing values)
  - However, high complexity (> 10 vignettes, ≥ 8 dimensions) $\longrightarrow$ lower consistency of responses of older or less-well educated respondents
  - Evidence for unwanted method effects such as simplifying heuristics and order effects (Sauer et al. 2011)
  - Information that is not varying across the vignettes: should be specified in the introductory part of FS modules / general instructions for respondents
- Number of levels
  - similar numbers of levels across dimensions to avoid "number-of-levels-effects".
  - High numbers of levels only if it is necessary for theoretical or methodological reasons
  - If possible, avoid the occurrence of many illogical and implausible combinations

Presentation Style and Answering Scale

- Text vs. table
  - First comparisons only find little differences in response quality (e.g., Shamon and Dülmer 2019, Sauer et al. 2020)
  - Tabular presentations might help overcoming effects of dimension order (Auspurg and Jäckle 2017)

- Random-order vs. extreme cases first
- Answering scales: ratings, nominal categories, amounts, probabilities, open (magnitude) scales



- The Value of Good Experimental Designs
  - ”Researchers should recognize that the designs chosen [...] are at least as, if not more important than, the models that one uses to analyze the resulting data” (Louviere 2006: 177).
  - The design determines the parameter identification, model flexibility, and the statistical efficiency of resulting estimates.
  - There are no ways to improve an experimental design once you conducted your survey!
- Example
  - Imagine a factorial survey with 4 dimensions 2, 3, 4, 6 levels per dimension
  - Full factorial: $2 \times 3 \times 4 \times 6 = 144$ possible combinations
  - Sample n = 20

Experimental Design: Orthogonality and Balance

	x1	x2	x3	x4
x1	1.0000			
x2	0.0000	1.0000		
x3	0.0000	0.0000	1.0000	
x4	0.0000	0.0000	0.0000	1.0000

Correlations, full factorial

	x1	x2	x3	x4
x1	1.0000			
x2	0.1455	1.0000		
x3	0.0306	-0.1919	1.0000	
x4	0.1619	0.2297	-0.0693	1.0000

Correlations, sample

x1

		Freq.	Percent	Valid	Cum.
Valid	1	72	50.00	50.00	50.00
	2	72	50.00	50.00	100.00
	Total	144	100.00	100.00	

Frequencies of x1, full factorial

x1

		Freq.	Percent	Valid	Cum.
Valid	1	6	30.00	30.00	30.00
	2	14	70.00	70.00	100.00
	Total	20	100.00	100.00	

Frequencies of x1, sample

- D1, D2 and D3, each having two levels, coded +1 resp. -1 ("orthogonal polynomial" or "effect coding")

	Main Effects			Two-Way Interactions			Three-Way Interaction
#	D_1	D_2	D_3	$D_1 \cdot D_2$	$D_1 \cdot D_3$	$D_2 \cdot D_3$	$D_1 \cdot D_2 \cdot D_3$
1	- 1	- 1	+1	+1	- 1	- 1	+1
2	- 1	+1	- 1	- 1	+1	- 1	+1
3	+1	- 1	- 1	- 1	- 1	+1	+1
4	+1	+1	+1	+1	+1	+1	+1
5	- 1	- 1	- 1	+1	+1	+1	- 1
6	- 1	+1	+1	- 1	- 1	+1	- 1
7	+1	- 1	+1	- 1	+1	- 1	- 1
8	+1	+1	- 1	+1	- 1	- 1	- 1

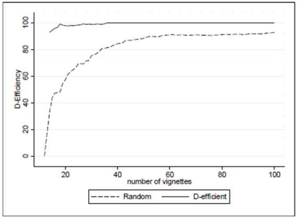
- With the first half-fraction, you cannot empirically distinguish these two models (see, e.g., Auspurg and Hinz 2015):
  $Y = \beta_0 + \beta_1 \times D_1 + \beta_2 \times D_2 + \beta_3 \times D_3$
  $Y = \beta_0 + \beta_1 \times D_2 \times D_3 + \beta_2 \times D_1 \times D_3 + \beta_3 \times D_1 \times D_2$

- By means of sampling techniques one should try to preserve the desirable characteristics of the vignette universe:
  - First: All meaningful parameters are identifiable (not confounded)
  - Only parameters that can be expected to show negligible effects (effect sizes close to zero) might be confounded (e.g. higher-order interactions)
  - In social sciences: 2-level interactions often have impact. Take care that they get not confounded!
- Second (and less important): Minimal correlations ("orthogonality") and maximal variance or "level balance" of all dimensions, therewith maximum statistical power ("efficiency")
- In particular when working with small fractions of the vignette universe one should take care of adequate sampling techniques

- Random sampling: simple random sampling, stratified random sampling, ...
  $\longrightarrow$ one leaves confounding patterns to chance
- D-efficient sampling: Maximizes orthogonality and level balance by means of computer algorithms $\longrightarrow$ control over which effects become confounded (Dülmer 2007)
- D-efficiency: measurement that jointly captures both criteria (orthogonality and level balance):
  $D_{eff} = 100 \times \frac{1}{n_s |X'X|^{-1}|^{\frac{1}{p}}} = 100 \times (\frac{1}{n_s}|X'X|^{\frac{1}{p}})$
  n_s: Sample size of the design
  $|X'X|$: Determinant of the Fisher information matrix of the vignette dimensions
  p: Number of b-coefficients that have to be estimated

- Realization of D-Efficient Designs: Specification of relevant parameters that should not be confounded, then looking for a D-efficient design that allows identifying all parameters (e.g., with SAS, see Kuhfeld et al. 1994).
- All dimensions should have similar numbers of levels or should be multiples of each other (e.g.: better 2,4,8 than 2,3,8 levels)
- Exclusions of illogical cases reduce D-efficiency. Consider such restrictions in advance.
- Often it makes sense to orthogonalize at least all two-way interactions ("Resolution V-Design").

Figure 3-2. *D*-efficiencies of *D*-efficient and Random Samples by Size of Vignette Fraction (Orthogonal Main Effect $3^3 2^4$ Designs)



Note: For each size of the vignette fraction, one random and one *D*-efficient sample was generated. For the *D*-efficient sample, a Stata programming code was kindly provided by Carsten Sauer.

Data Collection within Surveys

- Vignettes are in most cases too complex to be understood by respondents when the interviewer only reads them. At least the vignette module should be self-administered even in CAPIs
- For that reason telephone interviews are not advisable
- CAPIs and web surveys allow easily to implement different questionnaire versions for different respondents
- For all modes it is crucial to ensure that:
  - Vignettes are randomly allocated to respondents
  - There is a unique identification of the vignettes rated by the different respondents that is included into the response data

- With or Without Interviewer?
  - Interviewers can be helpful, but FS-modules are not standard $\longrightarrow$ interviewers should get special training.
  - Even with interviewers, the FS-module should be self-administrated
  - Interviewer can give an introduction and assistance if problems occur
  - Respondents seem to have no problems to answer vignettes self administrated in online surveys or PAPIs, but difficulties of handling depends on the research design

Survey Times

- Random allocation of vignettes to respondents
- Ensure a random order of vignettes for each respondent Response time: 10 vignettes with 8 dimensions, simple rating scale (as shown in our example) takes respondents about 4 minutes (median)

Table A1: Median response times by experimental split

	Number of dimensions		
Number of vignettes	5	8	12
10	2.85	3.52	3.85
20	4.97	5.72	6.33
30	6.87	8.37	9.55

Source: Project "Factorial Survey Design", general population sample

Analysis of Factorial Surveys

- Data preparation: reshaphing and merging
- Data quality checks: realized design, outcome measure, plausibility (hints for merging problems), missing values
- Regression analysis
  - Between-subjects design: OLS, logit, etc.
  - Within-subjects design: clustered observations (there might be an intra-rater correlation) $\longrightarrow$ cluster-robust regressions, multi-level regressions (fixed or random effects regressions), etc.
- Group differences (cross-level interactions): Statistical significance of group differences can be assessed by interaction terms between vignette dimensions and binary (0-1) coded group-variables that are separately (t-test) or jointly tested (F-test, “Chow”-test)

- Trade-Offs, part-worth, cross-elasticity
- Mean vignette evaluation:
  $Y = \beta_0 + \beta_A \times Dimension_A + \beta_M \times Dimension_M + \beta_k \times X_{ki}$
  $Dimension_M$: monetary dimension
  $Dimension_A$: binary coded dimension
- The „part-worth“ (PW) or willingness to pay for dimension A calculates as:
  $\beta_A \times Dimension_A + \beta_M \times Dimension_M = 0$ cross-elasticity of a 1-unit-change:
  $PW = -(\beta_A/\beta_M)$.
- All estimations rely on correct specifications of the regression model!
- Formulas for logged variables can be found in Auspurg and Hinz (2015)
- In Stata, you may use the ado wtp (willingness to pay)

Concluding remarks

Summary

- Construction of vignettes and cognitive limitations to process information
- Vignette sampling, efficiency and confounding
- Random allocation to respondents and random order of vignettes
- Prevention of implausible combinations
- Considering hierarchical data structure

Adriaans, J., Sauer, C., & Wrohlich, K. (2020). Gender Pay Gap in den Köpfen: Männer und Frauen bewerten niedrigere Löhne für Frauen als gerecht, DIW Wochenbericht 87 (10), 147-152.
Auspurg, K., & Hinz, T. (2015). Factorial survey experiments (Vol. 175). Sage Publications.
Auspurg, K., Hinz, T., & Sauer, C. (2017). Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments. American Sociological Review, 82(1), 179-210.
Auspurg, K., & Jäckle, A. (2017). First equals most important? Order effects in vignette-based measurement. Sociological Methods & Research, 46(3), 490-539.
Dülmer, H. (2007). Experimental plans in factorial surveys: random or quota design?. Sociological Methods & Research, 35(3), 382-409.
Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. Journal of Marketing Research, 31(4), 545-557.
Louviere, J. J. (2006). What you don't know might hurt you: some unresolved issues in the design and analysis of discrete choice experiments. Environmental and Resource Economics, 34(1), 173-188.
Sauer, C. (2020). Gender Bias in Justice Evaluations of Earnings: Evidence From Three Survey-Experiments. Frontiers in Sociology, 5, 22.
Sauer, C. G., Auspurg, K., Hinz, T., & Liebig, S. (2011). The application of factorial surveys in general population samples: The effects of respondent age and education on response times and response consistency.
Sauer, C., Auspurg, K., & Hinz, T. (2020). Designing multi-factorial survey experiments: Effects of presentation style (text or table), answering scales, and vignette order. Methods, Data, Analyses: mda, 14(2), 195-214.
Shamon, H., Dülmer, H., & Giza, A. (2019). The factorial survey: the impact of the presentation format of vignettes on answer behavior and processing time. Sociological Methods & Research.