methods, data, analyses | Vol. 14(2), 2020, pp. 195-214 DOI: 10.12758/mda.2020.06

Designing Multi-Factorial Survey Experiments: Effects of Presentation Style (Text or Table), Answering Scales, and Vignette Order

Carsten Sauer[1], *Katrin Auspurg*[2] *& Thomas Hinz*[3]

[1] *Department of Political and Social Sciences, Zeppelin University Friedrichshafen*

[2] *Department of Sociology, LMU Munich*

[3] *Department of History and Sociology, University of Konstanz*

Abstract

Multi-factorial survey experiments have become a well-established tool in social sciences as they combine experimental designs with advantages of heterogeneous respondent samples. This paper investigates three under-researched design features: how to present vignettes (running text vs. table), how to measure responses (rating vs. open scale), and how to sort vignettes (random vs. extreme-cases-first, to prevent censored responses). Experiments were conducted in a 2 x 2 x 2 between-subject design with 408 university students rating decks à 20 vignettes. Analyses of 7,895 ratings showed no differences of whether vignettes were presented as running texts or tables. Open scales revealed more measurement problems, e.g., missing values, than rating scales. Finally, vignettes presented randomly sorted produced similar results compared to sorting extreme vignette cases first. Recommendations based on the findings are to use random orders of vignettes and rating scales. Table vignettes provide an alternative to text vignettes but should be further evaluated with heterogeneous samples.

Keywords: Multi-factorial survey, vignette presentation, response scale, vignette order, ceiling effects

Multi-factorial survey experiments have become a well-established tool in the social sciences, mostly because they combine experimental design features (i.e. randomization) with the advantages of heterogeneous respondent samples (i.e. large and/or random samples that enable the estimation of heterogeneous treatment effects). In these survey experiments participants respond to descriptions of hypothetical objects or situations (vignettes). Within the vignettes, factors (dimensions) vary experimentally in their levels. The experimental variation allows an analysis of dimensions' causal influence on the responses (normative judgments or hypothetical decisions). At the same time, as the experiment is embedded in a survey, it is a tool to reach heterogeneous respondent samples and to analyze differences in attitudes or behavioral intentions across social groups. During the last years increasing numbers of studies have been published indicating that multi-factorial survey experiments became more and more a standard tool in social sciences (Auspurg & Hinz, 2015; for multifactorial, "conjoint" survey experiments in political sciences: Hainmueller, Hangartner, & Yamamoto, 2015).

Whenever implementing such experiments, researchers make decisions about multiple design features. Previous research focused on the complexity (number of dimensions and vignettes; see Auspurg, Hinz, & Liebig, 2009; Sauer, Auspurg, Hinz, & Liebig, 2011), sampling techniques (Atzmüller & Steiner, 2010; Dülmer, 2007, 2016), survey mode (Weinberg, Freese, & McElhattan, 2014), methods of data analyses (Hox, Kreft, & Hermkens, 1991), and external validity (Hainmueller et al., 2015; Petzold & Wolbring, 2019). Our study extends this literature by inves-

Acknowledgements

The data collection reported in this paper was supported by a grant funded by the German Research Foundation (DFG) (Hi 680/4-1). The paper was presented at the workshop "A perfect match? Comparative Political Economy and conjoint analysis" at the University of Zurich in 2017. We thank the organizers, in particular Silja Häusermann, and the participants, especially Devin Caughey, for their valuable comments and suggestions. Moreover, we thank participants of the colloquium of the Department of Sociology in October 2016 at Radboud University for many helpful comments. For contributing to this paper, Thomas Hinz was supported by EXC2035 "The Politics of Inequality". Carsten Sauer acknowledges support by the Dutch Research Council (NWO, Veni grant number: 4510-17-024).

Data Note

This article uses data from the project "The Factorial Survey as a Method for Measuring Attitudes in General Population Surveys". Replication files (Stata do-files and the used data) can be found on the following webpage: https://dx.doi.org/10.7802/2011

Direct correspondence to

Carsten Sauer, Lehrstuhl für Soziologie mit Schwerpunkt Sozialstrukturanalyse, Fakultät für Staats- und Gesellschaftswissenschaften, Zeppelin Universität, Am Seemooser Horn 20, 88045 Friedrichshafen
E-mail: carsten.sauer@zu.de

tigating the effects of three fundamental design features on the data quality which received little attention so far: first, presenting the vignettes in a running text vs. table format; second, using open response scales vs. rating scales with closed ends; and third, a random or systematic (extreme-cases-first) order of the vignettes presented to the respondents. The first two design features are crucial for all researchers in the field as they must decide how to present information and choose (at least) one answering scale. The third question about the vignette order is additionally important for the large bulk of applications with multiple vignettes per respondent: Researchers typically ask respondents to evaluate several (e.g., 10 or 20) vignettes (for a review of applications, see Wallander, 2009). As we will explain in more detail below, in these cases ordering the vignettes in systematic (instead of random) way is seen as a promising tool to avoid censored responses, but there are so far no empirical evaluations.

In the literature, there are some guidelines for the construction of multifactorial experiments to gather most reliable and valid results (Auspurg & Hinz, 2015; Jasso, 2006; Sauer, Auspurg, Hinz, Liebig, & Schupp, 2014). The findings of our study provide additional insights as so far only few studies contrasted a text and tabular format (Shamon, Dülmer, & Giza, 2019), an open and a rating scale (Auspurg & Hinz, 2015), and/or different vignette orders within the same experimental design.

Background: Why Should the Design Features Make a Difference?

Presentation Style. Vignettes used in multi-factorial survey experiments typically describe hypothetical situations or persons by a running text, i.e. a paragraph of one or several full sentences (see, Auspurg & Hinz, 2015, pp. 69-72). By doing so, the vignettes describe short scenarios close to 'real-life-stories,' which is seen as a main advantage of this presentation style. Moreover, it allows for a very subtle, indirect question format that can be useful to investigate sensitive topics (Auspurg, Hinz, Liebig, & Sauer, 2015). An alternative style would be a table format that only shows the dimensions and levels and avoids additional text. This presentation style is frequently used in conjoint studies and choice experiments, i.e. multi-factorial survey experiments that prevail in marketing research and economics. Critical about this tabular presentation style might be the more abstract question format which is not embedded in a story. Further possible limitations exist with respondents more likely using heuristics or being more prone to social desirability bias when the dimensions are presented more evidently in tables instead of being 'hidden' in smooth stories. However, there are also lots of advantages of tables: The format might minimize respondents' cognitive burden by reducing the reading task. Information presented

in tables can be assessed faster and should therefore economize on survey time. Additionally, table formats provide an appealing alternative to running text if one wants to randomize the order of the dimensions to neutralize potential effects of the dimension order (such as primacy and recency effects, see Auspurg & Jäckle, 2017). Vignette dimensions can more easily be rotated in a tabular format, as the order is no longer specific to the syntax of a language. In text vignettes, moreover, respondents might simply overlook some dimensions, which would obviously invalidate results gained by such experiments. Thus, even though running texts are mostly used in multi-factorial survey experiments so far, table formats may be a versatile alternative. So far, one study investigated differences between tabular vignettes and text vignettes using an online quota sample (Shamon et al., 2019) and finds no differences between the two methods regarding response inconsistency and processing time but more missing values (including refusals to answer any vignette at all) for text vignettes especially for respondents with lower educational degrees.

Response Scales. There are several ways to measure the responses to the vignette stimuli (see Auspurg & Hinz, 2015, pp. 64-67; Wallander, 2009). We tested the most frequently used response scales of vignette studies in the social sciences, an ordered rating scale (in our case an 11-point scale) against an open scale, also known as magnitude scale (Jasso, 2006; Sauer et al., 2011). The advantage of rating scales is that they are easily accessible for respondents as they are frequently used in various types of survey questions and, therefore, represent a standard tool of survey research. However, obviously, the range of values is restricted by the predefined minimum and maximum of such a scale. For this reason, ceiling effects might occur: In particular, when respondents have to rate multiple vignettes, they might not be able to express a more nuanced judgement that is located between to scale points or that goes beyond the scale's minimum or maximum. The resulting censored responses would lead to a systematic underestimation of the effects of vignette dimensions (i.e. there is a lower statistical power to detect the vignette dimensions' impact). Open (magnitude) scales that have no limits are deemed to overcome such ceiling effects and also to provide more fine-grained, metric values (Jasso, 2006). The drawback is that these scales likely cause a higher cognitive burden for the respondents. Open (magnitude) scales have been frequently used and recommended for multi-factorial survey experiments and conjoint analyses (for an overview, see Liebig, Sauer, & Friedhoff, 2015), but tests of their reliability are missing. (To best of our knowledge, the only systematic evaluation for multi-factorial survey experiments exists with a small marketing survey, a conjoint analysis, with 100 respondents in the U.S.; see Teas 1987.)

Vignette Order. The use of a random order of vignettes allows neutralizing possible effects of a fixed vignette order (such as carry-over, learning or fatigue effects). However, to avoid ceiling effects, some authors alternatively presented the vignettes in a systematic order, starting with the most extreme vignette cases. The

reason for this recommendation is that beginning with the vignettes likely to provoke the most extreme reactions could help to calibrate respondents regarding the end points of closed-ended rating scales (Auspurg & Hinz, 2015). Yet one drawback is that the researchers must decide which vignettes respondents may perceive as extreme cases. Systematic comparisons of both orders are lacking.

Interactions between the design features. Although it is not the core question of this study, our orthogonal, multi-factorial experimental design also allows us to test interaction effects between all three design features. The vignette order and response scales might have a different impact for tabular vignettes with a clear-structured presentation format compared to text vignettes, where respondents might be less aware of all dimensions. Similarly, the use of an extreme-case-first order might be especially effective in combination with closed-ended rating scales that are more prone to ceiling effects.

Data and Methods

We fully crossed all three design features (text/tables, response scales, and vignette order), leading to a 2 x 2 x 2 between-subject experiment (the between-subject design was chosen to not distract the respondents with changing scales or presentation styles). The substantive issue of the factorial survey module was the fairness of earnings of hypothetical full-time employees. The analysis sample consisted of 408 bachelor students of social sciences, 177 men and 231 women. All participants were recruited in 2008 in social science courses at 27 German universities and then randomly allocated to one of the 8 different experimental cells.[1] Depending on the local conditions, respondents could answer to the online survey (CASI) either during their course or afterwards in their free time. The questionnaire started with some socio-demographic questions, e.g., about the field of studies. The vignette module started with an introductory screen that provided shortly some general information on the hypothetical employees that was held constant for all vignette persons, such as their weekly working hours (40 hours). The following vignette module included 20 vignettes for each respondent. Table 1 provides the realized numbers of observations (rated vignettes) and number of participants per experimental cell.

In the vignettes, information on hypothetical employees participating in the German labor market was presented. The 8 dimensions (including the gross earnings) were selected close to prior factorial survey studies in the substantive field

1 The data collection was part of a larger project that investigated multiple methodological issues of multi-factorial survey experiments such as effects of the number of dimensions and levels, mode effects, and the reliability of measurement. Participating universities were recruited via personal contacts to the PIs.

Table 1 Number of Vignettes and Respondents (in Parentheses) per Experimental Cell

	Type of scale				
	Rating scale		Open scale		
Presentation of dimensions	Random order	Extreme cases first	Random order	Extreme cases first	Total
Text	1,087 (56)	1,044 (53)	916 (47)	839 (45)	3,886 (201)
Table	1,159 (58)	1,099 (55)	886 (47)	865 (47)	4,009 (207)
Total	2,246 (114)	2,143 (108)	1,802 (94)	1,704 (92)	7,895 (408)

Table 2 Vignette Dimensions and their Levels

#	Dimensions	(Number of) levels
1	Age	(4) 30, 40, 50, 60 years
2	Sex	(2) male, female
3	Degree	(3) without degree, vocational degree, university degree
4	Occupation	(10) unskilled worker, door(wo)man, engine driver, clerk, hairdresser, social worker, software engineer, electrical engineer, business manager, medical doctor
5	Experience	(2) short on, much
6	Tenure	(2) entered recently, entered a long time ago
7	Children	(5) no child, 1 child, 2, 3, 4 children
8	Earnings	(10) values from 500 to 15.000 Euros

(e.g., Jasso & Rossi, 1977; Shepelak & Alwin, 1986). Table 2 shows all dimensions and levels. Each vignette was presented on a single screen page. The task for the respondents was to assess the justice of the gross earnings. Respondents had the possibility to skip evaluations (no forced evaluations) and to return to vignettes evaluated before if they wanted to change their evaluation. About 92 percent of vignettes were visited only once, thus, respondents did not change their ratings. In 8 percent of the cases people went back to previous screens to change their judgements. Screenshots of some exemplary vignettes are provided in Online-Appendix, Part A.

We used a sample of vignettes as the full-factorial of all combinations of dimension levels would yield 48,000 vignettes. Our selection of 240 vignettes (12 decks à 20 vignettes) was based on the D-efficiency criterion (Kuhfeld, Tobias,

& Garratt, 1994). With this sampling method, it is possible to find a selection of vignettes in which correlations between dimensions are minimized (overall and within the different decks; criterion of orthogonality). At the same time, it is ensured that all levels of each dimension appear similarly often (criterion of level balance). Both criteria ensure that one receives a sample that allows to estimate coefficients efficiently and unbiased. Illogical and very implausible combinations were excluded, like medical doctors without a university degree.[2] (For a detailed description of the sampling method and comparisons with alternative designs, see Auspurg & Hinz, 2015).

The experimental manipulations were set-up as follows: The running text vignettes were programmed as shown in the sample vignette presented in Figure A1 in the Appendix A. The table format was programmed with 4 rows and 2 columns showing the dimensions and their levels (Figure A2_1 and A2_2). In these table vignettes, the order of dimensions was fixed to have equivalent conditions as in the text vignettes.

The answering scales were programmed in two versions with an 11-point rating scale versus an open (magnitude) scale. The rating scale had the standard format used in previous vignette studies with the scale running from -5 (unfairly too low) over zero (fair) up to +5 (unfairly too high). For the magnitude scale, we implemented a design very similar to that described in a prominent instruction on factorial surveys (Jasso, 2006).[3] This answering scale followed a three-step procedure (shown in Figure A2_1 and A2_2) as it is recommended in the literature (Jasso, 2006). First, respondents evaluated if the earnings of the vignette person were just or unjust. If respondents rated the earnings to be just, they approached to the next vignette. If respondents evaluated the earnings to be unjust, they answered in a second step whether the earnings were too high or too low. In a third step the participants were asked to specify the amount of injustice. Respondents could use their own unrestricted continuum of numbers that express their perception of injustice best for this evaluation step. Based on the insights of psychophysics (Stevens, 1975) these numbers are deemed to be metric evaluations. To have a reference point for these evaluations across respondents, a calibration vignette, which was the same for all respondents, was added in front of the vignette decks in the magnitude-split; i.e. all respondents first had to evaluate this calibration vignette (see Jasso (2006) for an in-depth description of this approach). For data analyses, these three response variables were transformed into one joint measurement following Jasso (2006): First, the ratings were combined within one numeric scale with zeros describing perfect justice, negative numbers describing under-reward and positive numbers describing

2 Plausible interaction terms have been orthogonalized (Resolution-IV-design). The D-efficiency of the 240 vignettes sampled was 91.

3 The method is based on psychophysics (Stevens, 1975) and has been applied in many factorial survey studies (for an overview in the justice literature, see Liebig et al., 2015).

over-reward. Second, the number continuums used by different respondents were calibrated by dividing these numbers by the rating of the calibration vignette.[4]

Regarding the variation of the vignette order, respondents evaluated in the first condition vignettes that were ordered randomly. For each respondent, the random order of the 20 vignettes in their deck was generated by a random number generator (we used the statistical software Stata). The second condition was an extreme-cases-first order. In this split, first, again for each respondent a random order of the twenty vignettes was generated. After the randomization, the order was manipulated by moving the two most extreme vignette cases (high underpayment and high overpayment) to the beginning of the vignette module. The driving dimensions for the selection of these extreme cases were the "gross earnings" and "occupation": We selected the two vignette cases that showed the highest (lowest) earnings given what is common in Germany for the respective occupations. To determine these cases' earnings, we used official information about the actual earnings by occupation from labor market data in Germany.[5] Information on earnings per occupation was chosen because existing surveys (and also our survey) showed that respondents in Germany account in their justice evaluations very strongly for what people realistically earn in different occupations. Therefore, these two vignettes could be expected to evoke extreme ratings in both directions (over- and underpaid). Putting them first is thought to lessen ceiling effects in later judgments of less extreme vignettes (Garret, 1982; O'Toole, Webster, O'Toole, & Lucal, 1999).[6]

Data Analyses. Data were analyzed using linear multi-level (random-intercept) regressions, with vignette evaluations at level 1 and respondents at level 2. The outcome variable was the vignette ratings of the respondents. To make estimates based on the open (magnitude) scale comparable to those based on the rating scale, all ratings were z-standardized. As input variables we used the vignette dimensions described in Table 2. The dimensions "degree" and "occupation" were included as dummy sets.

To identify if design features affected the importance of different dimensions for the judgements, we chose the following strategy: For each experimental split, the 17 coefficients were interacted with a binary-indicator for the two design variants (text vs. table format, rating vs. open scale, random order vs. extreme cases

4 The calibration has the drawback that one needs valid values in these first judgments. In our study 11 respondents produced missing values and 9 respondents evaluated the first vignette as just (0) and could therefore not be used for the calibration.

5 When there were several extreme vignette earnings in a deck (i.e. vignette earnings were at least for two vignettes twice or even three times the mean actual earnings for this occupation) we additionally used information on the educational degrees to determine the two most "extreme" under-/overpaid vignette cases.

6 Extremely under-rewarded vignette persons were, e.g., medical doctors with meagre earnings; extremely over-rewarded vignettes persons were, e.g., unskilled workers with top-earnings.

first) to test for significant differences. Control variables included the respective other design features as well as respondent's sex and the university where the survey took place (26 dummies). We estimated linear multi-level regressions,[7] post-estimation tests were used to assess differences by our three experimental conditions. We employed χ^2-tests for the null hypotheses that the interaction terms of vignette dimensions with the binary design indicator are (jointly) zero (this "omnibus" hypotheses test of that there are no differences at all is known as "Chow test", see Wooldridge, 2003). We report Sidak-adjusted *p*-values to account for multiple comparisons.

To check how design features affected response quality, we evaluated standard parameters to assess the response quality, such as the proportions of missing values with logistic regressions. In these analyses, we also explored two-way interactions between the different design features (e.g. between style of presentation and response scales). Moreover, we investigated response times and response consistency. General criteria to evaluate design features refer to the cognitive burden they impose on respondents. Obviously, the time respondents need to provide vignette evaluations serve as a proxy for the cognitive effort needed. We compare response times (measured during data collection for each of the 20 vignettes) by design splits and expect the scales to make a difference. For the analysis of response times we used median regression (Parente & Santos Silva, 2016). The consistency of responses is measured by another proxy, namely, the squared residuals following the procedure of Shamon et al. (2019) and Sauer et al. (2011). Lower values of squared residuals (given the same set of vignette dimensions for all respondents) are equal to a higher consistency in evaluations. While there are inter-individual differences (which are not at focus of this paper but see Auspurg, Hinz, & Liebig 2009) we assume again that the open scale is accompanied by less consistency. All data analyses were done with the statistical software Stata version 14.2 (StataCorp., 2013). The graphs were created with the user-written Stata ado *coefplot* (Jann, 2014).

Results

Before we report the results of the methods experiments, we take a quick look at the substantive results to check their plausibility based on the empirical justice literature. Respondents' evaluations led to plausible effects of vignette dimensions on justice evaluations and were in line with prior factorial survey experiments in the field of pay fairness: E.g., vignette persons were considered as being the more

7 Note, we used a Generalized Least Square (GLS) estimation that leads to approximately similar results as Maximum Likelihood (ML) estimation but makes no assumption about the distribution of the unit-specific error term. The results reported here are not affected by the estimation algorithm (GLS or ML) and lead to the same results.

likely underpaid, the higher their educational degree, labor market experience, and occupational prestige; and the lower their gross earnings. The substantive findings of these regression results are presented in Appendix B.

Effects on the Impact of Vignette Dimensions

What is more interesting for the study at hand: Did the results (effect sizes of dimensions) depend on the experimentally varied method features like the way vignettes were presented or had to be evaluated by respondents? Table 3 shows the differences across our three experimental splits (the underlying, substantive regression models and their interpretation are provided in Appendix B). Model 1 reports the results for table vs. text vignettes. The non-significant χ^2-values indicate that there are no differences in the effects of vignette dimensions on respondents' judgements between the two presentation styles. Moreover, the insignificant joint test reported in the last row of the table suggests that the two design variants (text or tables) produce similar results. Model 2 shows the differences in coefficients for open vs. rating scales. Of the 8 dimensions, 5 were found to show significant differences between the two scales and the highly significant joint test at the bottom of the table also indicated that the two scales produced strikingly different results. This difference will be analyzed in more detail in the subsequent paragraph. Note, even with an alternative categorical coding of the dependent variables (with three categories: under-rewarded, fair, over-rewarded) differences remained (see Appendix C), meaning that differences were not driven by outliers of the open (continuous) scale. Model 3 focuses on the splits in which the order of the vignettes was varied. Results show that differences (interaction effects) – both being tested separately or jointly – are statistically insignificant. That is, we did not observe any significant differences between coefficients estimated with a random order of vignettes or with extreme cases first. This result remains stable also in case of restricting the analysis sample only to respondents who did not change previous ratings (92 percent of the sample).

Response Quality of Response Scales and Vignette Orders

The analyses so far showed that only the choice of the answering scale had a significant impact on the regression results. The question follows, which scale performed better? Additional analyses revealed that the number of missing values was remarkably higher in evaluations made with the open scale than with the rating scale. Within the rating split, 4,389 vignettes were evaluated and 131 (2.9 %) vignettes were not. Within the magnitude split, 3,816 vignettes were rated and 344 (8.3 %)

Table 3 Tests for Design Effects on the Impact of Vignette Dimensions

		M1 Presentation: table vs. text	M2 Open vs. rating scale	M3 Extreme cases vs. random
	df	χ^2	χ^2	χ^2
Experimental variation x sex	1	3.170	0.399	0.004
Experimental variation x age	1	2.381	0.521	1.221
Experimental variation x degree	2	0.823	6.111*	1.402
Experimental variation x children	1	0.219	5.716*	2.319
Experimental variation x experience	1	0.386	10.454**	0.095
Experimental variation x tenure	1	1.370	0.001	1.312
Experimental variation x earnings	1	0.003	75.107***	1.613
Experimental variation x occupation	9	9.356	37.828***	5.011
Overall	17	22.497	177.836***	17.766

Notes. Tests after multi-level (random-intercept) regressions with interaction terms; df: degrees of freedom of the respective vignette dimension; reference category M1: text vignettes; M2: rating scale; M3 random order; Controlled for further experimental manipulations, respectively, and respondents' sex and place of survey (26 dummies for the universities). N_vignettes = 7895; N_respondents = 408; Sidak-adjusted p-values; * $p < .05$; ** $p < .01$; *** $p < .001$.

vignettes were not.[8] This difference indicates that the respondents had more problems (or were less cooperative) with the open scale with its three-step rating procedure. Table 4 shows the coefficients of a logistic regression on the probability of missing values and reveals that missing values were only significantly more likely with open scales (Model 1). As shown in Models 2-4, there were also no significant interactions between the type of scale and presentation style or vignette order, indicating the open scale to be the main driver of missing values.

Besides the probability of missing values, the share of explained variance (overall R^2 in Stata) of the linear multiple regression model (see Appendix B2) – as another measure of response quality – was remarkably lower with the open scale ($R^2 = .11$) than with the rating scale ($R^2 = .51$) indicating that a lot of noise in the data collected with the open scale affected the precision of estimation.

8 Note: 8,680 potential judgments = 4,389 valid rating scale judgments + 131 missing rating scale judgments + 3,506 valid open scale judgements + 344 missing open scale judgments + 310 missings because of failed calibration. The analysis of missing values only includes missings (131 + 344) that were produced by the respondents. The actual missings for the analysis of the open scale split were even higher due to the lost cases through the calibration.

Table 4 Logistic Regressions of the Probability of Missing Values (1 = yes) in Dependence of Design Features

	(1)	(2)	(3)	(4)
Style (ref. text)	-0.039 (0.391)	-0.165 (0.736)	-0.023 (0.562)	-0.042 (0.390)
Answering scale (ref. rating scale)	1.104* (0.432)	1.019 (0.572)	1.104* (0.431)	0.932 (0.592)
Order (ref. random order)	0.042 (0.390)	0.041 (0.390)	0.058 (0.539)	-0.214 (0.737)
Style * answering scale		0.176 (0.867)		
Style * order			-0.033 (0.780)	
Order * answering scale				0.359 (0.869)
Constant	-3.512*** (0.458)	-3.450*** (0.500)	-3.520*** (0.483)	-3.391*** (0.503)
McFaddens Pseudo R^2	0.034	0.034	0.034	0.035
$N_{vignettes}$	8680	8680	8680	8680
$N_{respondents}$	434	434	434	434

Notes. β-coefficients (log-odds) with cluster-robust (cluster=respondent) standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Proposed advantages of open scales are that they allow for more nuanced, fine-grained ratings of respondents. However, it is unclear if the respondents use the scale in the intended (metric) way. Table 5 shows the 10 most frequent values gained from the open scale. As it can be seen, respondents frequently used rough, rounded numbers (such as 100, 1000) to express their perception of injustice and did not fully exploit the open continuum of the scale.

Open scales are particularly deemed to perform better regarding the prevention of ceiling effects that could occur especially in a random order design. Table 6 provides the tests for differences in regression coefficients by vignette order separately for both scales. We use the multi-level linear regression models (shown in Model 1 and Model 3) and compare them to Tobit regressions that are regularly used to account for censored data (shown in Model 2 and Model 4). The joined test for differences across design features shows insignificant χ^2-values for the linear models and insignificant F-values for the interactions specified via Tobit regression models. Thus, the more nuanced regression analyses correcting for a possible censoring of responses that are presented in Table 6 are in line with the more general results reported in Table 3, Model 3: Overall, the differences between the modes

Table 5 Ten Most Frequent Values Indicated by Respondents on the Open Scale

Value	*N*	Percent
0	1282	36.57
100	319	9.10
10	201	5.73
1000	199	5.68
50	164	4.68
5	99	2.82
20	75	2.14
3	70	2.00
500	70	2.00
1	69	1.97

Table 6 Tests for Vignette Order Effects on Vignette Evaluations

		Rating scale		Open scale	
		(1) Linear regression	(2) Tobit regression	(3) Linear regression	(4) Tobit regression
	df	χ^2	F	χ^2	F
Extreme cases first x sex	1	0.024	0.002	0.059	0.106
Extreme cases first x age	1	0.374	0.184	2.937	1.044
Extreme cases first x degree	2	6.404*	2.849	8.887*	1.792
Extreme cases first x children	1	4.075*	3.256	0.591	0.576
Extreme cases first x experience	1	0.002	0.001	0.186	0.161
Extreme cases first x tenure	1	0.821	1.247	0.871	2.998
Extreme cases first x earnings	1	1.255	1.149	2.099	1.799
Extreme cases first x occupation	9	11.451	1.252	8.385	0.919
Overall	17	22.735	1.348	21.116	1.016
$N_{vignettes}$		4389	4389	3506	3506
$N_{respondents}$		222	222	186	186

Notes. Tests after multi-level estimation with interaction terms; df: degrees of freedom; reference category: random order; Sidak-adjusted *p*-values; * $p < .05$; ** $p < .01$; *** $p < .001$.

of sorting are marginal for both types of answering scales. A closer look on the coefficients shows that with the rating scale there are two vignette dimensions (educational degree and children) that significantly differ depending on the order of the vignettes (Model 1). In case of extreme-cases-first ordering, the coefficients of

these dimensions are bigger in absolute size compared to those in the mode of random order, indicating potential ceiling effects. However, we also find one significant difference (again, for educational degree) with the open scale (Model 3). This is, however, only one positive finding within 17 tests. Performing Tobit regressions to account for ceiling effects (with cluster-robust standard errors accounting for the nested data structure) completely vanishes the significant differences between the experimental splits (Models 2 and 4).

In a final step, the experimental splits are evaluated regarding response times and response consistency (based on the squared residuals, see Table 7). Model 1 shows the results of a median regression of response time on the design features. The constant indicates that on the average, respondents needed about 17 seconds to evaluate a single vignette. While there were no differences for table vs. text vignettes and for different order, the use of open answering scale took on average about 3.5 seconds longer than the rating scale. This seems obvious since the evaluation using the open scale is based on a three-step process. A more nuanced picture of the response time by vignette position offers Figure 1 and shows a well-known pattern. Respondents need more time during the first vignettes in all experimental splits to get used to the task. They speed up until the fourth vignette and have a roughly stable response time then. When comparing different modes, it becomes obvious that the respondents using the open scale need always some seconds more due to the more complex rating task. Besides this difference, the patterns are similar in all experimental splits. The analysis of response consistency shown in Model 2 of Table 7 highlights differences between the answering scales with open scales producing higher squared residuals. We find no differences between other design features and also no interaction effects between design features (not shown).

Table 7 Response Time and Response Consistency (Squared Residuals) by Experimental Variation

	(1) Response time	(2) Residuals sq.
Style (ref. text)	-0.984	-0.0593
	(0.593)	(0.167)
Answering scale (ref. rating scale)	3.531***	0.556**
	(0.618)	(0.185)
Order (ref. random order)	0.312	-0.0689
	(0.589)	(0.168)
Constant	17.12***	0.527*
	(0.989)	(0.244)
N	7895	7895
N_respondents	408	408

Note: Coefficients of Model 1 are based on a median regression with cluster robust standard errors. Coefficients of Model 2 are based on a multi-level regression (GLS) with robust standard errors. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

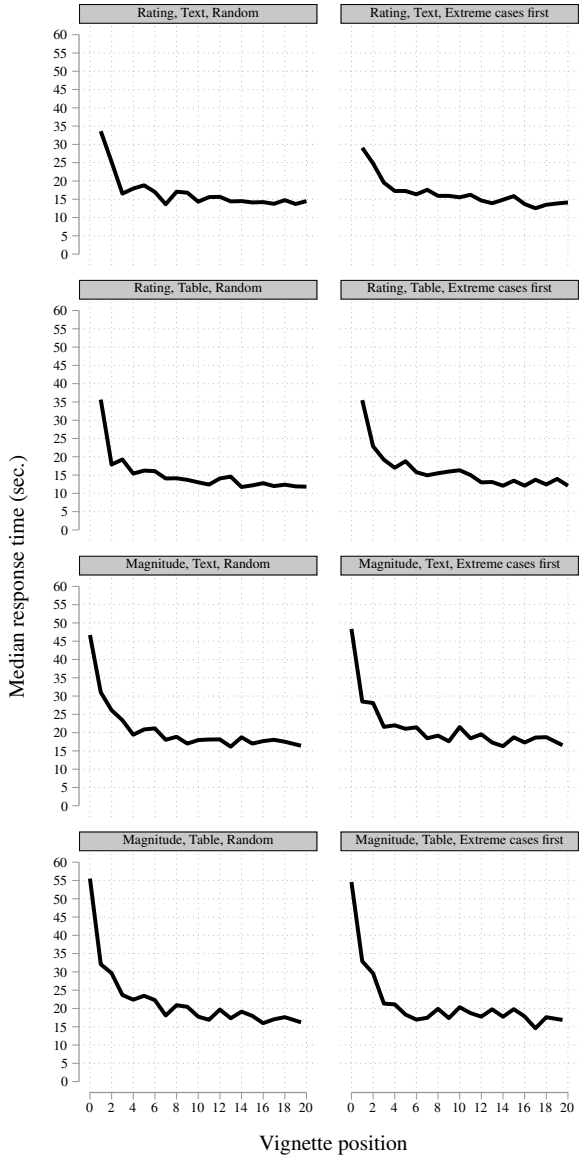

Figure 1 Median response time in seconds per experimental variation (rating vs. magnitude answering scale, text vs. table, random order vs. extreme cases first) and vignette position. Note, in the vignettes with the open (magnitude) scale every respondent rated the same vignette (vignette position = 0) before the deck with 20 vignettes started. Therefore, the figures for the rating task start at vignette position 1 and the others at vignette position 0.

Summary

This study analyzed the effects of design features of factorial surveys that have not been systematically evaluated so far, although these features are often varied across applications. We summarize the main findings in three implications and recommendations:

1. The presentation of dimensions in a running text – as it is done in most factorial surveys – did not produce significantly different results compared to a presentation in a table format. Our findings are in line with the study of Shamon et al. (2019) that also finds no differences between texts and tables focusing on response inconsistency and response time. However, their study finds differences between the two styles regarding the prevalence of missing values while we do not find differences. Shamon et al. (2019) find significantly lower total non-response (including refusals, break-offs, and vignette non-response) for table vignettes compared to text vignettes. They report about 24.1 percent of missing values for the vignette evaluations with most of them (18.3 percent) occurring due to refusals (i.e. respondents produced only missings in the vignette module or answered with a constant rating pattern). Focusing only on vignette non-response (without refusals) they report similar non-response numbers as we have (about 3.5 percent) and find support for text vignettes compared to table vignettes (less missing values). In our study we have only vignette non-response (2.9 percent with rating scales and 8.1 percent with magnitude scales) as nobody refused to fulfill the task. One explanation for different findings might be the different sample populations in both studies (in our study university students vs. quota sample of German population in Shamon et al. 2019) as well as the survey mode. We would expect that this difference is related both to the difference in population and survey mode, as well as the difference in the evaluation task. Taken together, we conclude that researchers might use tables instead of running texts, specifically if they want to neutralize possible effects of dimension order (see Auspurg & Jäckle, 2017), as tables allow for a more flexible (random) ordering of dimensions.
2. The rating scale clearly out-performed the open scale in many terms, e.g., in the number of missing values, and probably also produced more valid regression estimates. The open scales are more time consuming as a thorough introduction into the procedure and a calibration vignette is needed and, in our case, a three-step scale was necessary. In addition, the open scales did not come with the benefits of true metric scales. The findings are in line with other research indicating weak performance of metric scales with extensive response options (Sauer et al., 2014). We therefore recommend using standard, one-step rating scales. As we compared rating scales to three-step open scales, future research

should investigate potential differences between rating scales and one-step open scales used by Shamon et al. (2019).

3. The variation of the vignette order (random vs extreme-cases-first) did not yield to substantive differences in the overall estimation of regression coefficients. Only when splitting the analysis additionally by response scales, results slightly differed. Given these small differences, the easier and more flexible random sorting of vignettes seems quite more advisable. In case there occur ceiling effects, these can still be adjusted by means of specific econometric regression methods (cf. Auspurg & Hinz, 2015). Moreover, if ceiling effects occur in pre-tests, one might lessen them by switching to a broader rating scale (e.g., 11 points instead of 7) or lower numbers of vignettes.

Conclusions

Our study found only few method effects, which is good news: Factorial survey results seemed to be very robust against the tested variations of design features. However, an exception existed with open (magnitude) scales, which performed on many parameters worse than standard rating scales. Given the relatively common usage (and recommendation) of these response scales, this is an important finding. In standard survey research, these response scales were already abandoned due to similar problems as the ones found in our study (see, e.g., Schaeffer & Bradburn, 1989). However, in multi-factorial survey designs they have been still used until today to prevent censored responses. The latter were, however, hardly spotted in our survey. This makes us even more confident in our recommendation that also in multi-factorial survey experiments one should in future better rely on standard rating scales.

Our study also has limitations. The most important one is certainly that the participants were throughout university students. This standardization enabled us to have more power to detect pure effects of design features. But this specific population also impacts the generalizability of our findings to other samples, as this population is particularly used to read and process complex information (provided in tables). Thus, additional research with general population surveys is needed. In addition, one should test applications that are more prone to social desirability bias. Therewith, one could explore whether the evident presentation of dimensions in tables triggers more socially desirable evaluations as when potentially sensitive dimensions are embedded in a short story. Finally, we only tested one variant of open response scales that was bound to a three-step response procedure, and one specific survey mode (an online survey). Evaluations of other design variants are certainly desirable although they occur less likely in practice as we tested the most common designs.

In sum: The study shows that multi-factorial survey designs are robust against variations in presentation style and kind of vignette order but answering scales should be selected carefully.

References

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology, 6*(3), 128-138.

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments* (Vol. 175). Los Angeles: Sage Publications.

Auspurg, K., Hinz, T., & Liebig, S. (2009). *Complexity, learning effects, and plausibility of vignettes in factorial surveys.* Paper presented at the 104th Annual Meeting of the American Sociological Association, San Francisco.

Auspurg, K., Hinz, T., Liebig, S., & Sauer, C. (2015). The Factorial Survey as a Method for Measuring Sensitive Issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving Survey Methods. Lessons from recent Research* (pp. 137-149). New York: Routledge.

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.

Dülmer, H. (2016). The Factorial Survey Design Selection and its Impact on Reliability and Internal Validity. *Sociological Methods & Research, 45*(2), 304-347.

Garret, K. (1982). Child abuse: Problems of definition. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments. The factorial survey approach* (pp. 177-204). Beverly Hills: Sage.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences, 112*(8), 2395-2400.

Hox, J. J., Kreft, I. G., & Hermkens, P. L. (1991). The analysis of factorial surveys. *Sociological Methods & Research, 19*(4), 493-510.

Jann, B. (2014). Plotting regression coefficients and other estimates. *Stata Journal, 14*(4), 708-737.

Jasso, G. (2006). Factorial Survey Methods for Studying Beliefs and Judgments. *Sociological Methods & Research, 34*(3), 334-423.

Jasso, G., & Rossi, P. H. (1977). Distributive justice and earned income. *American Sociological Review, 42*(4), 639-651.

Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research, 31*(4), 545-557.

Liebig, S., Sauer, C., & Friedhoff, S. (2015). Using factorial surveys to study justice perceptions: five methodological problems of attitudinal justice research. *Social Justice Research, 28*(4), 415-434.

O'Toole, R., Webster, S. W., O'Toole, A. W., & Lucal, B. (1999). Teachers' recognition and reporting of child abuse: a factorial survey. *Child Abuse and Neglect, 23*(11), 1083-1101.

Parente, P. M. D. C., & Santos Silva, J. M. C. (2016). Quantile Regression with Clustered Data. *Journal of Econometric Methods, 5*, 1-15.

Petzold, K., & Wolbring, T. (2019). What can we learn from factorial surveys about human behavior? *Methodology, 15*(1), 19-30.

Sauer, C., 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. *Survey Research Methods, 5*(3), 89-102.

Sauer, C., Auspurg, K., Hinz, T., Liebig, S., & Schupp, J. (2014). *Method effects in factorial surveys: An analysis of respondents' comments, interviewers' assessments, and response behavior.* SOEP papers on Muldisciplinary Panel Research, No. 629/2014. German Socio-Economic Panel Study (SOEP). Berlin.

Schaeffer, N. C., & Bradburn, N. M. (1989). Respondent behavior in magnitude estimation. *Journal of the American Statistical Association, 84*(406), 402-413.

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*, doi: 0049124119852382.

Shepelak, N. J., & Alwin, D. F. (1986). Beliefs about inequality and perceptions of distributive justice. *American Sociological Review, 51*(1), 30-46.

StataCorp. (2013). *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual neural and social prospects*. New York: Wiley.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research, 38*(3), 505-520.

Weinberg, J. D., Freese, J., & McElhattan, D. (2014). Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample. *Sociological Science, 1*(19), 292-310.

Wooldridge, J. M. (2003). *Introductory Econometrics. A Modern Approach*. Mason, OH: South Western.