Auspurg, Katrin; Hinz, Thomas; Walzenbach, Sandra

Book Part — Accepted Manuscript (Postpint)

Are Factorial Survey Experiments Prone to Survey Mode Effects?

Suggested Citation: Auspurg, Katrin; Hinz, Thomas; Walzenbach, Sandra (2019) : Are Factorial Survey Experiments Prone to Survey Mode Effects?, In: Lavrakas, Paul J. Traugott, Michael W. Kennedy, Courtney Holbrook, Allyson L. de Leeuw, Edith D. West, Brady T. (Ed.): Experimental Methods in Survey Research. Techniques that Combine Random Sampling with Random Assignment, ISBN 978-1-119-08374-0, Wiley, Hoboken, pp. 371-392

This Version is available at:
https://hdl.handle.net/10419/231408

Are Factorial Survey Experiments Prone to Survey Mode Effects?

Katrin Auspurg[1], Thomas Hinz[2], and Sandra Walzenbach[2]

[1] Ludwig Maximilian University of Munich, Germany
[2] University of Konstanz, Germany

Suggested citation:
**Katrin Auspurg, Thomas Hinz, & Sandra Walzenbach (forthcoming): Are Factorial Survey Experiments Prone to Survey Mode Effects? In: Paul J. Lavrakas, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw, & Brady T. West (eds): Experimental Methods in Survey Research.Techniques that Combine Random Sampling with Random Assignment, Wiley, 371-392.*

Are Factorial Survey Experiments Prone to Survey Mode Effects?

Katrin Auspurg (Ludwig Maximilian University of Munich, Germany)
Thomas Hinz (University of Konstanz, Germany)
Sandra Walzenbach (University of Konstanz, Germany)

Abstract

Drawing on data from a nation-wide survey on the fairness of earnings, this piece of work examines the susceptibility of factorial survey experiments to mode effects. It uses multilevel models to compare the results from a completely self-administered questionnaire to those of an experimental condition in which interviewers were present.

From a theoretical point of view, interviewers might affect measurement error in different directions. On the one hand, they can provide explanations and assistance if respondents feel unsure about answering the rather complex vignette questions, or even motivate them to put more effort into the response process. As a result, respondents should be less inclined to use satisficing as a response strategy. On the other hand, respondents might feel pressured to provide answers quickly and in line with what is socially desirable.

We compare the two survey modes with regard to a set of indicators for data quality: nonresponse, common heuristics (ticking the middle and straight-lining), response consistency throughout the vignette set and social desirability bias. The latter is examined by analysing substantive results from the vignette module, which allows for quantifying the gender wage gap that respondents perceive as just.

The study discusses to what extent interviewers can relieve the cognitive burden associated with factorial surveys and positively affect data quality. Do the favourable effects of interviewers outweigh the negative ones? The study can help practitioners to decide if the gains in data quality are great enough to justify investment in a more expensive personal survey mode.

Keywords: factorial survey, vignettes, mode effects; interviewer effects; sensitive questions, gender wage gap

1) Introduction

Over the past decades, factorial survey experiments that ask respondents to evaluate experimentally varied scenarios (i.e. *vignettes*) have become an increasingly popular tool in many subfields of the social sciences (for overviews, see Mutz 2011; Wallander 2009). While there is some literature on experimental design issues, such as the ideal number of experimental factors and levels, the effects of survey modes and respondent samples have not received much attention. In this chapter, we will use key concepts of the total survey error framework (Biemer 2010; Groves & Lyberg 2010) to study possible mode effects. We will compare two survey modes, namely, a mode with interviewer presence vs. a completely self-administered mode. Both survey modes have been applied to study the same substantive issue (fairness of earnings) and have been used in random respondent samples of the German residential population in 2009. We are mainly interested in the effects of interviewer presence on item-nonresponse, inconsistency of responses and measurement errors such as response sets, and how these issues and possible further mode effects affect the substantive results gained by the factorial survey experiment in our case study.

We begin with a brief illustration of the factorial survey method (Section 2) and continue with a discussion of typical modes, design issues, and their connection to mode effects (Section 3). In the case study described in Section 4, two different survey modes are analysed: a *face-to-face* interview where respondents filled in the factorial survey module themselves, but an interviewer was present for optional support, and a completely *self-administered* mode (where respondents could choose between completing a mail and a web survey). Do these modes differ in regard to data quality issues such as the proportion of nonresponse and response consistency? Does this lead to different substantive results? We conclude with a short summary and discussion of the practical implications (Section 5).

1.1) Idea and Scope of Factorial Survey Experiments

Factorial survey experiments go back to Peter Rossi who originally used the method to study normative judgements (Rossi & Andersen 1982). Since then, they have been applied to a wide range of research topics, including punishment preferences for deviant behaviour, measurements of social status, normative perceptions of fair earnings, and attitudes toward immigration. In factorial survey experiments, respondents' opinions and attitudes are retrieved by asking participants to evaluate scenarios (*vignettes*) in which the values (*levels*) of several characteristics (*dimensions*) are experimentally varied.[1] Respondents are frequently asked to evaluate the vignettes on an ordinal rating scale. For illustration, we refer to the factorial survey experiment used in our case study: respondents assessed the fairness of earnings of hypothetical employees with varied characteristics on an eleven-point rating scale ranging from *unfairly low* to *unfairly high* (see Figure 1).

Figure 1. Vignette Implementation: Example (Varied Dimensions Underlined)

A 60-year-old woman with vocational training works as a social worker.
Her monthly gross earnings total 2500 Euros (before tax and extra charges).

Are the gross monthly 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
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

By including various dimensions at once, vignettes can account for the fact that (normative) judgements typically require a simultaneous consideration of several factors. An independent but joint experimental variation of the factors allows the researcher to quantify their impact on the requested evaluations and to estimate trade-offs and interactions between the different factors. For instance, one could find out if respondents believe that men or women should get higher earnings, and if this is true even if they show the same levels of education, or if this holds particularly true in some occupational fields, such as typical male or female occupations.

[1] Note that there are also "vignette studies" where researchers do not vary any characteristics. We use the term *factorial surveys* for all designs where the vignette scenarios are varied in the form of a multifactorial experiment.

Because of the multifactorial design, it is also possible to estimate (monetary) trade-offs, such as just gender pay gaps (for more details, see Auspurg & Hinz 2015a).

Typically, respondents are asked to evaluate 10–30 vignettes, because these multiple ratings allow economizing research resources (many vignette judgements can be gathered with few respondents). However, there are cognitive restrictions: methodological studies have found that fatigue effects occurred when respondents had to evaluate more than 10 vignettes (Sauer, Auspurg, Hinz, & Liebig 2011). To get higher numbers of vignettes evaluated in a survey, researchers commonly work with different subsets of vignettes that are included into different versions of a questionnaire (these different vignette subsamples are called *decks* or *sets*; for more information on such design issues, see Section 3.2).

A crucial advantage of factorial survey experiments is that they combine randomised treatments with the benefits of surveys. Based on the randomised assignment of respondents to experimental stimuli (vignettes), the method offers a high internal validity for testing causal hypotheses. But while laboratory experiments usually suffer from small sample sizes and homogeneous groups of participants (Henrich, Heine, & Norenzayan 2010), factorial survey modules can be implemented in any (large scale) survey.[2]

1.2) Mode Effects

1.2.1) Typical Respondent Samples and Modes in Factorial Survey Experiments

Factorial survey experiments have frequently used small convenience samples, such as university students (Wallander 2009). This might be adequate if one is solely interested in testing causal mechanisms based on the experimental vignette factors that do not involve any moderator (or mediator) variables on the part of respondents. More insights on potential moderator variables, and consequently more general conclusions, can, however, be gained with more heterogeneous respondent samples (Auspurg & Hinz 2015a). Random respondent

[2] Although not the primary focus of this chapter, there are other survey methods apart from factorial surveys that rely on multifactorial experimental plans, namely, choice experiments and conjoint analyses (Auspurg & Hinz 2015b). In contrast to factorial surveys, they usually do not use full textual descriptions of scenarios (vignettes) but sum up characteristics in short keywords that are arranged in a table format. The crucial feature of choice experiments is that respondents make choices between options. Their answer is not a gradual one but a definite "yes" statement for the preferred option. In conjoint analyses, the typical task consists of simultaneously ordering options according to the respondent's preferences.

samples provide more firm ground for inference statistics, and they are needed if one wants to describe distributions such as the amount of social consensus across a population. For instance, do older and younger respondents share the same (justice) principles? Are there indicators of a social cleavage, e.g. between respondents with leftwing or rightwing political positions (for such an application, see Hermkens & Boerman 1989)?

Empirical social research often builds on telephone surveys, as these provide a less cost-intensive alternative to face-to-face interviews. However, although there are a few factorial survey studies using this mode (e.g. Pager & Quillian 2005), most applications seem to be too complex to be completely administered on the phone. Typical applications involve various experimental dimensions that are combined into paragraph descriptions spanning several text lines. A full processing of so much information probably requires the respondents to read the vignettes themselves. One has at least to expect *recency-effects* (for those effects, see Krosnick & Alwin 1987) when vignettes are only presented orally – that is, respondents will likely only recall the last pieces of information given to them, which would give vignette dimensions presented later in the vignettes disproportionate weight on evaluations.

For all these reasons, it is clearly advisable to allow respondents to assess the vignette information themselves. This can be achieved using different modes, such as computer-assisted personal or self-administered interviews (CAPI and CASI), or paper and pencil interviews (PAPI). In personal interviews, the interviewer should be instructed to hand the questionnaire (computer) over to the respondents so that they can read and evaluate the vignettes themselves, while the interviewer remains present in the room to provide optional support. Following this advice, factorial survey modules are nearly always given in a self-administered manner. Researchers have nevertheless to make a decision which mode to use for their survey (completely self-administered or not).

Often this decision is linked to the techniques used to recruit respondents. We will focus on two different options that are frequently applied in social science research, and which we have used in our case study. (i) *Random route* procedures are often employed in countries where countrywide registers of the residential population do not exist and are implemented as a multistage sampling technique. Starting with randomly selected addresses within randomly selected geographical areas, interviewers walk from house to house following a prescription that should ideally lead to a "random route" through the area (Bauer 2014). During this walk, dwelling units are listed. The final sampling step consists of a random selection of a person within the household, which is commonly realized by a Kish-selection grid. This sampling technique is typically combined with face-to-face interviewing, as the interviewer is already

present with “a foot in the door.” (ii) Respondents can alternatively be recruited via a short screening interview on the phone, where a *random digit dialing* technique is combined with a technique (such as a Kish-selection grid) to randomly select one target person inmultiperson households. After that, for factorial surveys, all recruited respondents get invited to a self-administered mail and/or online survey (as indicated, it is not recommended to do factorial surveys completely on the phone).We will refer to this technique hereafter as the completely *self-administered* mode.

This second option, the completely self-administered mode, is in general cheaper than the random walk procedure in combination with a personal interview, but is often accompanied with lower response rates and more severe coverage problems (such as underrepresentation of less educated people or low-income households; see de Leeuw 2008: 127 et seq; Groves et al. 2009: 153 et seq; Sue & Ritter 2007: 7–8). However, factorial surveys are an experimental method, and the key factor for good experimental designs is not the random selection of respondents, but rather a well-established experimental design, combining an adequate set-up of vignettes which contain the experimental stimuli and are randomly allocated to respondents. If this is done well, coverage and nonresponse problems should not strongly impair valid conclusions on the causal impact of vignette dimensions (Mutz 2011). Sampling errors might still affect descriptive statistics, which one might try to correct through weighting procedures. So far, there seems to be nothing specific about factorial survey methods. Thus, the discussion boils down to the question of whether the presence of an interviewer has an impact on measurement issues, and if yes, what that impact is.

1.2.2) Design Features and Mode Effects

When setting up factorial surveys, researchers have to plan different steps. They have to select a number of experimental dimensions and levels for the vignette characteristics, and they have to choose a strategy for sampling vignettes and splitting them into different subsamples (i.e., *decks*). Some further design issues, such as use of response scales, also have to be decided. In the following, we will review the most important design features and discuss how they might be related to survey modes, with a focus on the presence of interviewers.[3]

[3] For more detailed instructions see, for instance, Auspurg & Hinz (2015a) and Mutz (2011).

Selection of Dimensions and Levels

Factorial surveys are primarily designed to test theories on the causal impact of experimental factors. Therefore, selecting the vignette dimensions should be done carefully. The recommendation is to use theories to derive the dimensions and specify the regression equations one wants to identify (including possible interactions effects). This helps not only to ensure that all relevant dimensions are included, but also to optimize the experimental design. Regarding the number of dimensions, the general advice is to use a mid-level of complexity of about 6–8 variable dimensions that all have about 2 or 3 different levels.[4] Theoretical assumptions that respondents are sensitive to complexity have been put forward under the term *satisficing*: respondents do not necessarily use maximum effort to answer survey questions, but might shortcut the process to provide satisfactory answers requiring less effort (Krosnick 1991; originally Simon 1957). According to Krosnick, satisficing is more likely when the respondents' cognitive ability and motivation are low, but also when the task complexity is high.[5] For the data used in our case study, Sauer et al. (2011) only found a lowered response consistency for very complex vignette modules (consisting of 12 instead of 5 or 8 variable dimensions). According to their results, participants first learn and get better at dealing with the question format, until a fatigue effect sets in after about the 12th vignette (the authors tested at most 30 vignettes per respondent).[6]

However, it was not investigated if these effects interact with the survey mode. In completely self-administered modules, respondents can simply drop out if they are overburdened. In contrast, in face-to-face interviews, social cooperation norms make it very unlikely that, once the interview has started, respondents will refuse to finish the questionnaire or one of its core modules (Meulemann 1993: 113 et seq; Schnell 1997: 122 et seq, 144). This is also because respondents can easily rely on interviewer assistance in case they have problems fulfilling the task. Particularly if there are many vignettes and dimensions, however, respondents in a face-to-face mode also might suffer from fatigue effects or boredom. In that case, respondents might

[4] Lower numbers of levels help minimize the number of vignettes that are needed to achieve a certain amount of statistical power (for details see Auspurg & Hinz 2015a). Yet, one needs at least 3 different levels to estimate nonlinear relationships, and there might be further substantial reasons to specify more different vignette levels. Similar numbers of levels across dimensions help to avoid the so-called *number of levels effects* that have been found in choice experiments (De Wilde, Cooke, & Janiszewski 2008; Verlegh, Schifferstein, & Wittink 2002; Wittink, Krishnamurthi, & Reibstein 1990): dimensions varying on a higher number of levels attracted disproportional attention, probably because they have been more eye-catching than dimensions showing less variation.

[5] However, too few dimensions might also cause problems, because respondents might lack information to evaluate the vignettes and undercomplexity might risk causing boredom and fatigue effects (Auspurg et al. 2009).

[6] These results are somewhat confirmed by Teti, Groß, Knoll, & Blüher (2016). The authors report that even elderly respondents aged up to 90 had no difficulty in responding consistently to a vignette module on relocation preferences that consisted of 10 vignettes with 6 (dichotomous) dimensions.

provide satisficing responses (i.e. select any level on the response scale or fade out some dimensions). Although there might be lower nonresponse in cases of interviewer assistance, at least for complex modules the data quality might be lower. This is also because respondents might experience pressure to evaluate the vignettes within a short time, so as not to keep the interviewer waiting. However, one might also argue that the absence of anonymity in the face-to-face situation makes respondents invest more cognitive effort into answering the vignette questions. It is therefore still an open question which mode shows a higher data quality.

Generation of Questionnaire Versions and Allocation to Respondents

In the overwhelming majority of factorial survey modules, the full *universe* of all possible combinations of dimensions and levels is too large to be completely administered. This means that the researcher needs techniques for selecting vignettes to be used in the survey and to allocate these vignettes to the different questionnaire versions. What is important is that, first, the chosen vignettes show minimum correlations of vignette dimensions (including their interaction effects), because otherwise one would no longer be able to disentangle their effects. Second, it is desirable that one is able to estimate the effects of the vignette dimensions independent of deck effects (which might stem from the specific composition of vignettes within the different decks) or characteristics of respondents. If characteristics of respondents get confounded (i.e., get strongly correlated) with vignette dimensions, the effects of these characteristics or respondents' cognitive abilities might be mistaken for substantial effects of the vignette dimensions (Sauer et al. 2011).

So far, most researchers have used random selections of vignettes that leave confounding structures simply to chance. One should better use *quota designs* that allow the researcher to plan the confounding structures beforehand (Atzmüller & Steiner 2010; Dülmer 2007; 2016). In particular, D-*efficient designs* are recommended. They combine the vignette dimensions in a way that makes the vignette sample as orthogonal (meaning that vignette dimensions are maximally uncorrelated) and balanced (in that all dimension levels occur with about the same frequency) as possible, while at the same time ensuring that all important parameters (effects of vignette dimensions) can be identified. Optimizing these features allows maximizing the precision (or power) with which effects of the vignette dimensions can be identified. The higher this statistical efficiency, the lower the number of respondents needed to achieve a given amount of precision (for details, see Auspurg & Hinz 2015a). Regarding mode effects, such sampling techniques might also help to decrease correlations of decks with interviewer effects in personal interviews. To be able to separate substantial from possible interviewer effects, one

should at least avoid strong correlations of decks with interviewers. This can be achieved by randomly blocking decks to interviewers or by using *D*-efficient sampling techniques again (where the interviewers have to be treated as another blocking factor; the later technique would again increase the statistical efficiency of the experimental design in terms of a maximum orthogonality of substantive *and* methodological factors).

Whatever way is chosen to select vignettes, it is absolutely crucial that they are *randomly* assigned to respondents. In addition, when working with fixed decks, it is advisable to randomise the vignette order across respondents to avoid possible order effects.[7] All these designs likely pay out particularly in the case of heterogeneous respondent samples, as these are more prone to a confounding of respondents' effects with order or deck effects.

There is one additional issue to be considered. Often, combining all vignette dimensions and levels leads to implausible or even unrealistic vignette scenarios (for instance, in our case study, there could have been medical doctors without a university degree). It has been shown that too unrealistic combinations of levels lead respondents to ignore the respective dimensions for future evaluations (Auspurg, Hinz, & Liebig 2009: 86). Very unrealistic combinations should therefore be dropped. But there may nevertheless be some implausible combinations: eliminating all such combinations often triggers strong correlations of vignette dimensions, and it might also be interesting to get some evaluations on scenarios that expand reality (Rossi & Anderson 1982). In case respondents get irritated, the interviewer might inform them that those vignette cases were deliberately designed. In that regard, we expect interviewer presence to lead to lower nonresponse rates and/or to evaluations based on more different dimensions.

Number of Vignettes per Respondent and Response Scales

As already mentioned, it is recommended to restrict the number of vignettes per respondent to no more than 10 vignettes if one wants to prevent fatigue effects. Another decision parameter is survey time, particularly in face-to-face interviews, meaning that longer vignette modules increase the costs per interview ("time is money"). We can only provide some rules of thumb. According to our experience, for vignette modules with mid-complexity (8 dimensions and about 10 vignettes), respondents need an average of about five minutes to complete the vignette module (for more detailed statistics, see Auspurg & Hinz 2015a). In personal interviewers,

[7] Such random orders can be achieved by means of software in computer-assisted surveys, but it is also possible to implement them in paper questionnaires (for some instructions, see Auspurg & Hinz 2015a). Methodological research showed that the effects of the order of vignette dimensions can also occur in complex designs (containing 12 variable dimensions), or if respondents feel unsure about the topic (Auspurg & Jäckle 2017). Randomizing the order of vignette dimensions across respondents – which can especially be implemented in computer-assisted modes – can help to neutralize such order effects.

respondents might show a higher willingness to evaluate long vignette modules, while at the same time counterweighting this lower nonresponse by satisficing strategies. That is, we expect effects that are similar to those assumed for the number of dimensions.

Regarding response scales, sometimes two-step answering scales were employed (where, for example, respondents first rate whether the vignette earnings are fair or not, and then rate the amount of unfairness only if earnings are perceived as being unfair). Such scales, however, offer easy opt-out alternatives to respondents and thus enhance satisficing (Sauer, Auspurg, Hinz, Liebig, & Schupp 2014). For this and other reasons, it is advisable to use standard response scales like the ordinal rating scale employed in our case study (see Figure 1 again), although alternatives have been implemented by some researchers (Jasso 2006; Wallander 2009: 511). Complex response scales (such as magnitude scales) probably work with interviewer assistance; they can definitely not be recommended for self-administered modes (Auspurg & Hinz 2015a: 64 et seq)

Interviewer Effects and Social Desirability Bias

Turning more specifically to mode effects, the role of the interviewers has been widely debated (Groves et al. 2009: 141 et seq; Loosveldt 2008). On the one hand, interviewers can clarify the respondents' role and provide explanations and assistance, which might at least be important in case of more demanding response tasks (as reported by Holbrook, Green, & Krosnick 2003 for a standard survey). On the other hand, data collection in a social situation offers manifold ways for interviewers to compromise data quality and introduce systematic interviewer bias. Apart from conscious falsifications, misinterpretations of vague responses and inappropriate feedback, more subtle cues such as the interviewer's gender or dialect can also play a role. What is also important is that in face-to-face interviews the clustering of respondents within interviewers and sample points causes autocorrelations of error terms (Snijders 2005). These autocorrelations commonly decrease the efficiency (precision) of estimates (Groves et al. 2009: chapter 4.4; Auspurg & Hinz 2015a: chapter 5). However, an important concern of self-administered modes is the inability to control the circumstances under which the survey is completed (Zwarun & Hall 2014). Research with factorial surveys is often targeted at sensitive issues, such as respondents' discriminatory attitudes or the reported likelihood of engaging in criminal behaviour (for some applications, see Auspurg, Hinz, & Sauer 2017; Pager & Quillian 2005; Graeff, Sattler, Mehlkop, & Sauer 2014). Factorial surveys are assumed to be less prone to social desirability bias than single-item questions, or even randomised response techniques that are particularly designed to reduce this bias. Auspurg et al. (2015) were the first to report

promising empirical evidence on this topic. Compared to a direct question format, the vignette module yielded less socially desirable answers concerning a just gender wage gap.[8] However, it was not possible to assess if the results generalized to modes other than the employed face-to-face mode, and if there was still some social desirability bias left. Such bias might be induced by the interviewers' presence, even when respondents were allowed to answer the vignette questions themselves, because of a latent fear of negative consequences within a personal situation of low anonymity, and a higher awareness of social norms due to the presence of another person. There is so far only one study focusing on the second mechanism: in a split-ballot experiment, in which a student sample answered a factorial survey module on the fairness of earnings, the presence of an experimenter led to stronger preferences for equality (Liebig, May, Sauer, Schneider, & Valet 2015). Therefore, particularly in face-to-face interviews, one might use techniques to reduce social desirability bias. For instance, one could vary the sensitive dimension only between (and not within) the vignettes presented to single respondents.[9]

1.2.3) Summing Up: Mode Effects

All in all, we expect higher nonresponse rates in the completely self-administered mode. This is because it is more difficult for respondents to ask for assistance if they are, for instance, irritated by some implausible vignette combinations, or have any other problems with the response task. In addition, the risks of being sanctioned for quitting the questionnaire are lower. For other data quality issues, it is more difficult to make clear predictions on mode effects. Respondents might be more likely to take shortcuts in more anonymous situations, but there are also good reasons to assume that respondents speed up and satisfice if an interviewer is sitting next to them. Social desirability bias should be more present in the face-to-face mode; therefore, in our application, we expect lower effects for the most sensitive dimension, that is, the vignette person's sex, in the face-to-face mode. In addition, the stronger clustering of responses in this

[8] Similar results in a related scenario question format have previously been found in articles by Armacost, Hosseini, Morris, & Rehbein (1991) and Burstin, Doughtie, & Raphaeli (1980). However, the authors varied none or only one dimension of their vignettes.

[9] In a study by Auspurg et al. (2015), using an application on fairness of earnings, the results of a split with a pure between variation of the sensitive dimension (the vignette person's gender) did not differ from the split employing a within variation. This finding might, however, not generalize to more sensitive topics: in a factorial survey on discriminatory attitudes towards Muslims, answers from a between condition suffered less from social desirability bias than the ones from a within condition (Walzenbach & Hinz 2019).

mode (evaluations are nested within interviewers) should lead to a lower statistical efficiency (i.e. higher standard errors).

1.3) Case Study

While mode effects are a well-investigated field in general survey methodology (Berrens, Bohara, Jenkins-Smith, Silva, & Weimer 2003; Carini, Hayek, Kuh, Kennedy, & Ouimet 2003; Fisher & Herrick 2013; Malhotra & Krosnick 2007), empirical research on mode effects for factorial survey experiments is, to the best of our knowledge, completely lacking (with the only exception being the cited study of Liebig et al. 2015 on interviewer presence). All in all, it is an open question whether the higher survey costs related to personal interviews pay out in higher data quality. To find out, we run a case study with different survey modes.

1.3.1) Data and Methods

Survey Details

Our case study consists of two factorial survey experiments on the fairness of earnings in Germany that were run simultaneously in 2009, using the same questionnaire and factorial survey module. However, the two surveys differed in recruitment procedure and data collection mode. Respondents either completed the questionnaire in a personal interview (*face-to-face mode* with an interviewer present) or as a totally self-administered survey, which could take place online or as a paper and pencil questionnaire (*self-administered mode*). Recruitment into those survey modes took place by drawing two separate random samples from the adult residential population in Germany. For the face-to-face mode, 129 regional sample points in Germany were selected. Sampled subjects were chosen by applying a random route strategy in combination with a Kish-selection grid. For this face-to-face sample, there were 82 interviewers, each conducting between 2 and 28 interviews (median value: 7). For the evaluation of the factorial survey module, the interviewers handed the computer over to the respondents so that they could read and evaluate the vignettes themselves, but the interviewers remained present in the room. After the factorial survey module, interviewers took over the computer again and went on with asking questions. For the self-administered data collection mode, respondents were recruited by telephone using random digit dialing. Within this group,

respondents selected via the Kish-selection grid were offered two options to complete self-administered questionnaires.[10]

To test if mode effects interact with the complexity of factorial survey modules, the length of the sequence (10, 20, or 30 vignettes) and the number of dimensions (5, 8, or 12) were experimentally varied using split ballot experiments in both survey modes. Figure 1 shows a vignette example with 5 dimensions. Table 1 contains an overview of all experimentally varied dimensions and levels.

Table 1. Vignette Implementation: Overview of Dimensions and Levels

		Dimensions	**Levels**
exp. split with 5ddimensions	1	Age	30/40/50/60 years
	2	Sex	Male/female
	3	Vocational training	Without degree/vocational degree/university degree
	4	Occupation (ordered according to their Magnitude Prestige Scores)	Unskilled worker/door(wo)man/engine driver/clerk/hairdresser/social worker/software engineer/electrical engineer/manager/medical doctor
	5	Monthly gross earnings	500/950/1200/1500/2500/3800/5400/6800/10000/15000
experimental split with 8 dimensions	6	Experience	Little/much
	7	Job tenure	Entered recently/entered a long time ago
	8	Children	None/1/2/3/4
experimental split with 12 dimensions	9	Health status	No health problems/long-term health problems
	10	Performance	Below average/above average
	11	Economic situation of the firm	High profits/threatened by bankruptcy/solid
	12	Firm size	Small/medium/large

For implementation, a *D*-efficient sample of 240 vignettes was drawn from the vignette universe (which contained more than 1 million possible vignettes).[11] To standardize the experimental design across all experimental splits, the same vignette sample was used for all complexity conditions and irrelevant surplus dimensions were deleted for the splits with 5 and 8 dimensions. The number of vignettes, as well as the order in which the scenarios were

[10] Respondents in the self-administered mode could decide themselves whether they preferred the paper or the online questionnaire. In the analyses, we will combine both variants, because we are mainly interested in the effects of sampling compositions resulting from the two different recruiting techniques and in interviewer effects.

[11] Some implausible combinations such as medical doctors without a university degree were deleted (for details, Sauer et al. 2011).

displayed, was randomly assigned to respondents, and in the mode with interviewers, decks were randomly blocked to the different interviewers.[12]

Analysis Techniques

In the first step (Section 4.2), we will assess if the different modes are accompanied by different data quality issues: Do both modes differ in regard to nonresponse rates and indicators for measurement errors, such as respondents engaging in satisficing strategies (signaled by response sets or inconsistent judgements)? Even if there are some data quality issues, results from the factorial survey experiment, with its high potential to increase the internal and construct validity (Auspurg & Hinz 2015a), might turn out to be robust. Therefore, in the second step (Section 4.3), we explore differences in substantive results. These sections also include some analyses on social desirability bias, respondents' use of heuristics, and the possible loss of statistical efficiency in the face-to-face mode that is caused by interviews not being independent but clustered within interviewers and sample points.

When setting up our case study, we decided to apply the most typical survey settings, where data collection modes go hand-in-hand with different recruitment strategies (random walks in the personal interviews and phone recruitments in the completely self-administered mode). This makes it necessary to disentangle the possible mode effects we are interested in from sampling errors: our two survey splits might not only differ because of mode effects but also because of the composition of respondents. To separate both effects, we did some analyses on the composition of respondents (for details see Appendix 6.1). Both modes/recruitment strategies showed some indications for sampling errors: in both splits, some groups of respondents were under- or overrepresented in comparison to the sociodemographic distributions expected in the general German population (these "true" distributions at the time when the survey took place were captured by census data; see Appendix 6.1). We used different techniques to adjust for these sample selection errors and, more importantly, to adjust the composition of respondents in both survey splits. Respondent characteristics were used as controls in multivariate regressions, and in the analyses on the substantive effects of the vignette dimensions we tested if results changed when using poststratification weights to adjust both samples to the general population. More details on the employed (regression) techniques are provided in the respective subsections.

[12] However, this was done in a way that makes the data mostly balanced, meaning that, across interviewers, all different decks occurred with about the same frequency.

1.3.2) Mode Effects Regarding Nonresponse and Measurement Errors

Nonresponse

Table 2 shows descriptive statistics about item-nonresponse by survey mode. If we only compare the portions of respondents with at least one missing value in the vignette module, numbers first seem pretty similar in face-to-face interviews and the self-administered mode (10.3% vs. 9.4%). However, a closer look shows that there is no complete refusal to answer the factorial questionnaire if an interviewer was present, whereas 1.4% of the respondents in self-completion modes left out the whole module. Partial nonresponse is a bit higher for face-to-face respondents (10.3%) than for respondents who answered in the self-administered mode (8.0%).

Table 2. Item-Nonresponse by Survey Mode

		Face-to-face	**Self-administered**
Data: respondents who at least saw the first vignette		777 respondents	844 respondents
Respondents refusing to answer any vignette question		0% (0)	1.42% (12)
Respondents partly refusing to answer vignette questions		10.3% (80)	8.0% (67)
Missing vignettes in case respondent partly refused answers (absolute numbers in parentheses)	Overall	10.6% (1.9)*	17.3% (2.9)*

*Difference significant on 5%-level.

Given that partial item-nonresponse takes place, the quantity of missing values also differs by sampling mode. Respondents who answered the questionnaire in self-administered modes show higher proportions of item-nonresponse than the face-to-face participants. This is true both for absolute frequencies (face-to-face: on average 1.9 vignettes per respondent; self-administered: 2.9) as well as for proportions relative to the number of displayed vignette questions (face-to-face: 10.6%, self-administered: 17.3%, difference significant according to undirected *t*-test: $p = 0.011$).

Table 3 shows the results of binary regressions on the missing values. We display average marginal effects (AMEs) under control of the respondents' sex, age, and educational background. Apart from the survey mode, the number of vignettes (10, 20, or 30) and the

number of vignette dimensions (5, 8, or 12) were included as explanatory variables, both measuring the complexity of the factorial survey module. Model 1 contains only the cases with complete nonresponse. Since completely missing vignette modules only occurred in self-completion questionnaires, the corresponding mode dummy predicts the dependent variable perfectly, and the model is restricted to the self-administered sample. For the models on partial nonresponse, the dependent variable was coded 1 as soon as at least one of the vignette questions remained unanswered, irrespective of the exact amount of item-nonresponse. Cases in which vignettes are partly missing are compared with those in which respondents showed no nonresponse whatsoever, meaning that all the vignette modules with complete nonresponse were excluded from the analysis.

Table 3. Item-Nonresponse in the Vignette Module – Regression Tables

	Logistic Regressions (AMEs) (controlled for respondents' sex, age and education)			
	(1) Complete nonresponse (self-admin)	**(2) Partial nonresponse (both samples)**	**(2a) Partial nonresponse (face-to-face)**	**(2b) Partial nonresponse (self-admin)**
Face-to-face (ref: self-administered)	*omitted*	0.018 (0.015)		
20 vignettes (ref: 10)	-0.013 (0.008)	0.038* (0.017)	0.057* (0.026)	0.021 (0.023)
30 vignettes (ref: 10)	-0.006 (0.011)	0.055** (0.021)	0.089** (0.032)	0.025 (0.026)
8 dimensions (ref: 5)	0.001 (0.010)	-0.011 (0.017)	-0.024 (0.027)	0.001 (0.022)
12 dimensions (ref: 5)	0.004 (0.010)	0.009 (0.018)	-0.009 (0.028)	0.024 (0.023)
Intercept	-2.908* (1.262)	-2.911*** (0.539)	-2.932*** (0.724)	-2.823*** (0.797)
Pseudo R^2	0.113	0.031	0.037	0.041
N respondents	767	1593	771	822

Standard errors in parentheses.
+ $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$
Data basis: respondents who saw at least the first vignette; models on partial nonresponse do not consider cases with complete nonresponse and models on complete nonresponse ignore cases with partial nonresponse; additional drop-outs occur due to missings on the control variables.

Contradictory to what we found for complete nonresponse, the data collection mode has no noteworthy effect on partial nonresponse (Model 2, summarizing both modes). Concerning the

complexity measures, the number of presented vignettes shows significant effects. Respondents seem to get tired of the relatively complex vignette questions if they become too numerous. In sets with 20 vignettes, nonresponse is already more likely than in sets with only 10 vignettes, and the effect is even larger and more significant for the 30 vignette questionnaire versions. If we consider partial nonresponse separately for the two modes (Models 2a and 2b), we can see that this result is mainly driven by the respondents who completed the questionnaire in a face-to-face situation. This might be a hint that respondents feel rushed when an interviewer is waiting for them to finish the vignette task. If respondents do not want to continue answering, producing missing values might simply be a more socially acceptable way to get through the vignette task than completely refusing it in front of the interviewer.

All in all, the presence of an interviewer seems to be a powerful strategy to prevent complete dropouts, although the exact same personal interview situation is likely to foster item-nonresponse if the vignette module is too long. Irrespective of data collection mode, we did not find any effect of the number of displayed vignette dimensions on nonresponse.[13]

Response Quality

Table 4 sums up some indicators for response quality by survey mode. It is based on the cases that will be considered as valid from now on. This means that all unanswered vignette questions were dropped. Additionally, two more respondents (one from every sample mode) were excluded from further analyses since they always ticked the middle category on the answer scale, so that it can be doubted that they took the task seriously. We are therefore left with 12,699 vignette judgements from 776 face-to-face interviews and 13,537 evaluations from 831 completely self-administered questionnaires.

There are no differences in the proportion of vignettes rated with the middle category fair but the standard deviation of vignette evaluations is a bit higher in the face-to-face mode and this difference is highly significant in a two-tailed t-test ($p = 0.000$). These results are further illustrated in Figure 2, which shows the distribution of the vignette evaluations by survey mode. It is interesting to note that the extreme categories of the ordinal scale were more likely to be ticked in the face-to-face interviews than in the self-administered modes – which is probably the main mechanism responsible for the reported differences in standard deviations.

[13] Additional analyses showed that there were no significant interaction effects between the number of vignettes and the number of dimensions.

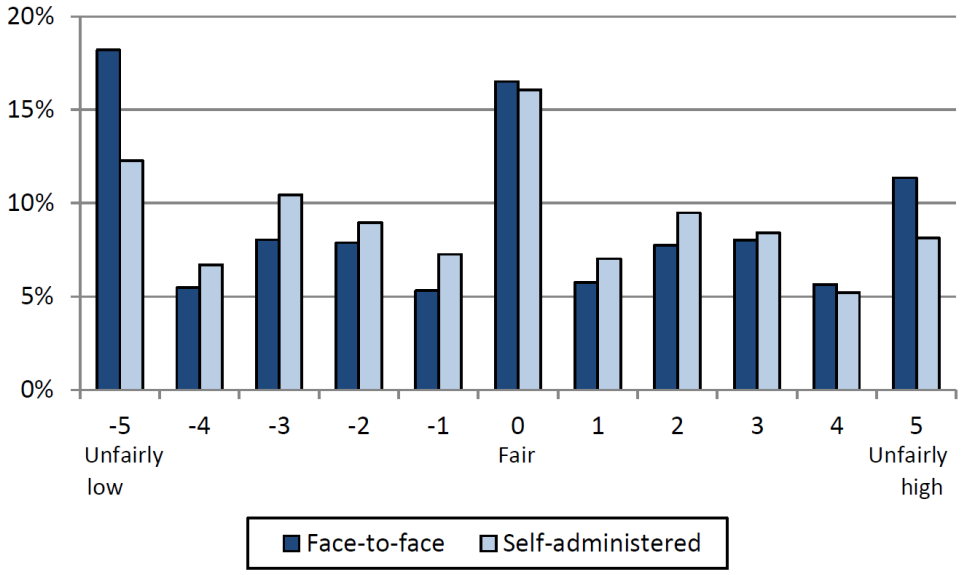
Table 4. Response Quality by Survey Mode

		Face-to-face	Self-administered
Data: non-missing vignette evaluations		12,699 (by 776 respondents)	13,537 (by 831 respondents)
Proportion of vignettes rated as "fair"		16.5%	16.1%
Standard deviation of vignette evaluations[2]	mean	3.30*	3.05*
	median	3.35	3.07
Proportion of unexplained variance if vignette judgements are regressed on first five vignette dimensions	mean	7.5%*	8.8%*
	median	4.8%	5.7%

*Difference significant on 5%-level.

[2] There was one respondent in each survey mode with only one valid judgement, meaning that the standard deviations of their vignette evaluations could not be calculated.

Figure 2. Vignette Evaluations by Survey Mode



Traditionally, higher standard deviations are considered an indicator for good data quality in item batteries, suggesting that respondents do not satisfice by ticking the same response level many times. Frequently selecting only the two most extreme categories, however, might represent another heuristic behaviour. Apart from that, differences between survey modes are unsystematic, and even the proportion of vignettes rated as fair is almost equal in both samples.

To obtain another indicator of response quality, the vignette judgements were regressed on the five vignette dimensions that were evaluated by each respondent.[14] This procedure was undertaken separately for each respondent, so that the residual sum of squares could be calculated as an indicator of individual response consistency. Consistent behaviour is expected to result in a good model fit, that is, a little proportion of variance left unexplained by the vignette dimensions. As can be seen in Table 4, the average proportion of unexplained variance was significantly higher in the self-administered mode (p =0.000), suggesting that respondents answered more consistently if an interviewer was present.

The proportion of unexplained variance on the respondent level can be used as the dependent variable in a regression model to see how the vignette module's complexity and survey mode influence the consistency of responses. Since the proportion of unexplained variance did not at all follow a normal distribution, we used a fractional regression model. In such models, no assumption about the distribution is made and effects are estimated on the conditional mean (for details, see Papke & Wooldridge 1996). The results of these regressions are shown in Table 5. To detect potential interaction effects of complexity and mode, the analyses were also run separately for the face-to-face interviews and the self-administered survey (Models 1 and 2).

An interesting first result is that the length of the factorial survey module has a negative effect on response consistency in both survey modes. However, the number of displayed vignette dimensions only seems to matter in the self-administered surveys. Compared to the experimental split with 5 dimensions, those respondents who had to evaluate vignettes with 12 dimensions show significantly higher shares of unexplained variance, and thus probably more inconsistent responses.[15]

The pooled model containing both samples (Model 3) confirms the mode effect already reported in the descriptive statistics in Table 4. Also under the control of respondent characteristics and task complexity, the result holds that the face-to-face group leaves lower proportions of variance unexplained and hence answers more consistently throughout the factorial survey module.[16] A higher consistency might, however, also be bought by using some heuristics, such

[14] The vignette dimensions 6–12 that were only answered by some of the respondents were intentionally omitted from the regression models for two reasons: to ensure comparability across the experimental splits and to avoid the systematic loss of cases with a high number of dimensions but with vignette sets that were too small to provide enough degrees of freedom to estimate the full model.

[15] Additional analyses showed that there were no significant interaction effects between the number of vignettes and the number of dimensions.

[16] Inconsistent responses certainly lead to larger residuals. However, there might also be other reasons for large residuals, such as misspecification errors (e.g. when respondents react in a nonlinear way to the age of the vignette persons).With the low number of degrees of freedom in respondent-specific regressions, it is hardly possible to test for such misspecification errors. We therefore use the proportion of unexplained variance only as a first hint for a lower response consistency.

as fading out dimensions. We will check in Section 4.3 if the effect sizes of vignette dimensions depend on data collection modes.

Table 5. Response Consistency in the Vignette Module – Regression Tables

	Fractional response regression (AMEs) Dep.var: proportion of unexplained variance (controlled for respondent sex, age and education)		
	(1) **Face-to-face**	**(2)** **Self-admin**	**(3)** **Both samples**
20 vignettes (ref: 10)	0.096*** (0.007)	0.106*** (0.006)	0.100*** (0.005)
30 vignettes (ref: 10)	0.114*** (0.007)	0.159*** (0.010)	0.138*** (0.006)
8 dimensions (ref: 5)	-0.002 (0.006)	0.009 (0.006)	0.004 (0.004)
12 dimensions (ref: 5)	0.000 (0.006)	0.022** (0.007)	0.012** (0.005)
Face-to-face (ref: self-administered)			-0.014*** (0.004)
Intercept	-3.412*** (0.172)	-3.259*** (0.249)	-3.218*** (0.162)
Pseudo R^2	0.082	0.102	0.092
N respondents	743	788	1531

Standard errors in parentheses.
+ $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$
Data basis: respondents who answered at least ten vignettes.

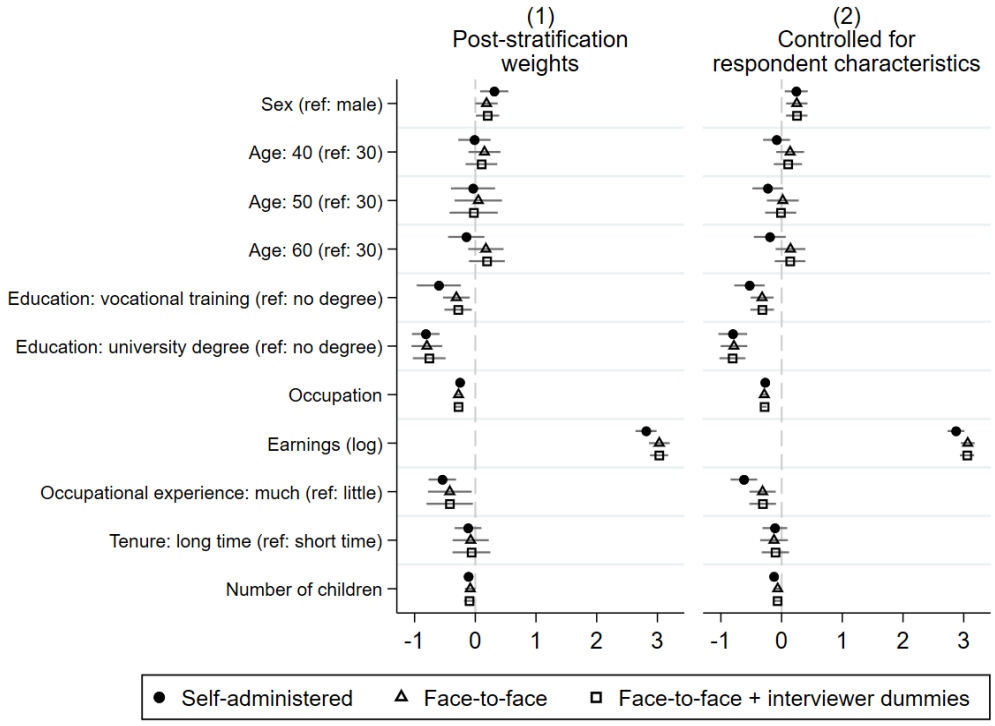
1.3.3) Do Mode Effects Impact Substantive Results?

Point Estimates and Significance Levels

Leaving potential differences arising from varying complexity levels aside, the following analysis refers only to the group of respondents who answered 10 vignettes with 8 dimensions. As the main target of factorial surveys is to estimate the causal effects of vignette dimensions (Auspurg & Hinz 2015a; Rossi & Anderson 1982), measurement errors affecting these estimates would obviously be particularly problematic.

Figure 3 shows how the varied dimensions in the vignette texts influence the respondents' judgements on the justice of earnings (for the underlying regression models, see Appendix 6.2). The plotted coefficients were obtained by regressing the respondents' vignette evaluations on the vignette dimensions of sex, age, educational background, occupation (higher numbers are related to a higher prestige score), earnings, occupational experience, length of tenure, and number of children. This was done separately for the self-administered survey (circles) and the face-to-face condition (triangles and squares). Since various judgements are nested within respondents (Hox, Kreft, & Hermkens 1991), cluster-robust standard errors (on the level of respondents) were applied.

Figure 3. Regressions of Fairness Evaluations on Vignette Dimensions by Survey Mode



Data: only respondents evaluating 10 vignettes with 8 dimensions; regression coefficients with 95% confidence intervals. For the 10 different occupations, prestige scores as measured by a common prestige scale in Germany were used (magnitude prestige scores; for these, see Christoph 2005). Positive (negative) values mean that earnings are evaluated as being unfairly too high (low) compared to the reference group.

To differentiate between genuine data collection mode effects and mere sampling problems, it is necessary to take the sociodemographic discrepancies across the two samples into account. A common but controversial method to deal with this issue is the application of post-stratification weights to adjust the drawn sample to the distribution of certain sociodemographic features known for the general population. An alternative approach would be to include only respondent characteristics as control variables in the regression model (for a discussion see

Snijders & Bosker 2011: chapter 14). Both strategies were applied to our data. Model 1 (the left column in Figure VIII.3) uses post-stratification weights, while Model 2 (the right column) controls for the respective respondent characteristics (sex, age, and educational background). *R*2-values were similar across models for both modes, with slightly higher shares of explained variance in the face-to-face condition (adjusted *R*2 = 0.712 and 0.734) compared to the self-administered survey (adjusted *R*2 = 0.704 and 0.710; see the full regression table in Appendix 6.2).

First of all, one can see that results are plausible. For instance, vignette persons who have a higher educational degree, whose occupations show higher prestige, and who have more work experience should get higher gross earnings: the earnings are more likely to be evaluated as being too low (see the negative effects of these vignette variables). All these effects are at least significant at a 5% level, meaning that the 95% confidence intervals do not intersect with the zero-line. A comparison between the self-completion (circles) and the face-to-face mode (triangles) shows that the effects of most regressors differ either not at all or only moderately between the two modes. This is particularly true for the weighted data, where we only find one noteworthy difference: higher amounts of income are penalized more clearly in the face-to-face mode. However, the difference between modes marginally fails to reach the conventional 5% significance level (p =0.077).[17] Once we control for respondent characteristics (see the second column of Figure 3), we observe some more severe modes effects. The interaction between income and mode reaches statistical significance (p =0.037), with the income having more impact in the face-to-face mode, while job experience was more important in the self-completion mode (p =0.026).[18]

Interestingly, we do not find any significant mode effects concerning the most sensitive of our vignette dimensions: the sex of the hypothetical employee. Respondents approve lower earnings for women, which is evidence for some kind of discrimination going on (for more details, see Auspurg et al. 2017), but they do so consistently and irrespective of interviewer presence. Only after post-stratification weighting do we find a subtle cue that, in some cases, a reactive face-to-face mode might increase social desirability bias or activate norms of equality. If we translate the respective coefficients to the gender wage gaps perceived as just, respondents in the self-administered surveys would penalize women with a 10.5% deduction in mean wages

[17] The significance level was obtained by specifying a full model containing the data from both survey modes, meaning that apart from the main effect for mode, interactions between mode and all the vignette dimensions could be estimated.

[18] This is evidence against self-completers fading out more dimensions. Dimensions were not consistently rated more important in the face-to-face mode. This only was true for the earnings while the contrary occurred for job experience.

but only 5.9% in the face-to-face sample.[19] However, we could not observe those differential estimations for the models that include respondent characteristics as controls. In contrast, some rather robust differences were found in the regression intercepts: respondents seem to agree to higher baseline wages if an interviewer is present (p =0.077 for the weighted data and p =0.003 if respondent characteristics are controlled).

According to an overall Chow test, the null hypothesis that the effect sizes of vignette dimensions in both samples are equal was not rejected for Model 1 (p =0.238), but Model 2 only barely failed to reach statistical significance (p =0.055).[20] When the same analysis was run with the most and least complex vignette modules (the conditions that assessed 30 vignettes with 12 dimensions and 10 vignettes with 5 dimensions) as a robustness check, overall Chow tests indicated mode effects at least on the 10% significance level in all cases. They were particularly driven by differences in the baseline wages perceived as just: in the case of interviewer presence, respondents considered higher base wages as just (see the more negative intercepts in the regression table in Appendix 6.2). Moreover, above-average effort was rewarded significantly more in the self-administered condition than when an interviewer was present. (This finding could only be observed in the high complexity specification with 30 vignettes and 12 dimensions, because this dimension was not included in the less complex vignette versions with only 8 dimensions.)

Measurement Efficiency and Sources of Unexplained Variance

So far, we have ignored the fact that vignette judgements in the face-to-face mode are not only clustered within respondents but also within sample points and interviewers on a still higher level. This means that one can expect cluster (*design*) effects that decrease measurement efficiency. To address this issue, we specified an additional model that captures the hierarchical data structure more adequately by adding dummy variables for the different interviewers (see squares in Figure 3). In our case, however, the nested data structure does not seem to be accompanied by a high loss of efficiency: standard errors change only marginally when taking all three levels into account, and, as a consequence, estimates with and without fixed effects (dummies) for interviewers are very similar.

[19] The just gender wage gaps were computed with the formula mean[income]*(exp(-_b[sex]/_b[income])-1) (for further details, see Auspurg & Jäckle 2017 or Auspurg & Hinz 2015a).

[20] Chow tests are Wald tests for the null hypothesis that several interactions between survey modes and vignette dimensions are jointly zero (Wooldridge 2003). In our case, the Chow test comprised interactions of the mode with all vignette variables and the main effect of the mode. When testing without the main effect, the p-value was 0.467 for the weighted data and 0.036 for Model 2.

Beside a possibly biased estimation of significant effects, another potential consequence of ignoring the hierarchical data structure could be an erroneous attribution of the variance in vignette evaluations to respondents, although the observed results would partly describe differences between interviewers. To check how much variance can be ascribed to interviewer effects, we estimated random effects models for the experimental condition in which respondents answered 10 vignettes with 8 dimensions (random effect models allow for the estimation of the amount of unexplained variance that can be attributed to higher levels, i.e. respondents or interviewers). In a first step, random effects were specified only on the respondent level (similar to what was done in Figure 3). The second step enhanced the model for the face-to-face group to a three-level specification that also takes interviewer random effects into account. As before, the vignette evaluations were regressed on the vignette dimensions separately for both modes. Apart from the respondents' sociodemographic characteristics (sex, age, and educational background), however, two regional variables were added to the models: a dummy variable specifying if a respondent lived in the western or the eastern part of Germany and an ordinal indicator if their place of residence was located in a rural or an urban area. This was done to disentangle possible interviewer effects from confounded variables indicating regional variance.

Table 6. Random Effects (RE) Models: Variance due to Interviewers

	Self-admin RE respondents (2 levels)	**Face-to-face** RE respondents (2 levels)	**Face-to-face** RE respondents + interviewer (3 levels)
ICC (respondents):	0.062	0.154	0.040
ICC (interviewers):			0.127

Data: only respondents evaluating 10 vignettes with 8 dimensions.
ICC = intra-class-correlation.

Table 6 shows the calculated intra-class-correlations (ICCs) for these models. The ICCs indicate the fraction of unexplained variance that can be attributed to the respondent and interviewer levels. At a first glance, the ICC for respondents is considerably lower in the self-administered surveys (6%) than in the face-to-face group (15.4%; see the first row in Table 6, which shows results of two-level models that only include random effects for respondents). However, as the three-level specification suggests, the lion's share of this variance is caused by differences in interviewers (or by other regional variables we have not controlled for). These

results suggest that interviewers (or the small regional districts they cover with their random walks) cause much additional noise to the data that has to be accounted for to be able to achieve the main goals of factorial survey research: exploring the impact of vignette dimensions and the heterogeneity of respondents in regard to their judgement rules.[21]

1.4) Conclusion

This chapter has contributed to fill the void in methodological research on factorial surveys. Based on a total survey error approach, we discussed different design characteristics of factorial survey experiments as well as decisions about data collection mode and sampling with respect to possible measurement and representation errors. Empirically, the data analysed came from a factorial survey module on the justice of earnings obtained through two different data collection modes. Respondents either completed the factorial survey module while an interviewer was present or as a self-administered survey. Recruitment into those modes took place after drawing two separate random samples.

Regarding mode effects, there were some differences in the prevalence of nonresponse. Nonresponse was particularly low in the face-to-face sample, but also the self-administered mode produced low nonresponse rates: in both modes, less than 3% of vignette evaluations were missing. It is noteworthy that interviewers in the face-to-face mode were very successful in convincing *all* respondents to answer at least parts of the vignette module and avoiding complete dropouts.

Regarding possible measurement errors, evidence was mixed. Respondents did not differ across mode in their use of evasive response sets (that is, ticking the middle or always the same response category). Small but significant differences were found with regard to the standard deviation of responses, such that face-to-face participants tended to make greater use of the extreme categories on the answering scale. These small effects do not suggest substantial differences in heuristic response behaviour across survey modes. However, response consistency was significantly higher if an interviewer was present during data collection.

[21] Obviously, respondents in self-administered data collection modes are also clustered in regional units, such as the (city) districts they live in. However, sampled with a random dialing procedure, the respondents are more evenly distributed across the whole regional area where the target population lives. Adding random effects for regional units (such as federal states) hardly changes the ICC measured for respondents. In contrast, the random walk procedure leads to a narrow clustering of respondents within smaller geographical units. In addition, interviewer effects cause autocorrelations of observations collected by single interviewers.

Sequences containing more than 10 vignettes were, irrespective of survey mode, related to higher proportions of unexplained variance and thus inconsistent responses, suggesting that the number of evaluation tasks should not exceed this threshold in any mode of data collection. At the same time, higher numbers of vignette dimensions were only problematic in terms of lower response consistency if there was no interviewer present.[22]

We also tested whether substantive results differ across modes: the experimental design might make factorial survey methods immune against the found measurement errors. First, concerning the causal impact of vignette dimensions, it has to be said that, overall, there were very few differences across data collection mode. There was some evidence for mode-specific perceptions of just baseline wages (in the case of interviewer presence, respondents supported somewhat higher base wages), whereas most of the vignette dimensions did not show any differential evaluations across mode. Depending on the model specification, either 1 or 2 out of the 8 vignette dimensions were found to differ by survey mode on a 10% significance level. Respondents in the self-administered mode more strongly acknowledged the job experience of the vignette person, while respondents in the face-to-face mode reacted more strongly to the earnings of the vignette person. An explanation for this latter finding could be that the presence of others motivates respondents to support higher earning equality (Liebig et al. 2015). However, it is difficult to decide if this higher preference for equality is triggered by the frame of a more cooperative situation when others are present or by social desirability bias. Obviously more research on the sources of this mode effect is needed.

Finally, adding interviewer dummies or random effects to account for the nested data structure hardly changed the importance of the vignette dimensions on the final judgement. Multilevel analyses, however, showed that the face-to-face survey mode induced some random noise due to observations not being independent from each other but interviews being clustered into interviewers and/or the small geographical areas where the interviewers recruit their respondents by means of the random walk. Following our random effects estimates, more than 10% of unexplained variance can be attributed to the level of interviewers. When not taking this source of heterogeneity into account, one might erroneously conclude that there are large differences between respondents' judgement styles.

To sum up, effect sizes in factorial surveys are somewhat sensitive to survey mode, whereas the distinction between significant and nonsignificant effects seems to be more robust across

[22] Further research should be undertaken to analyse whether part of the higher response consistency in the face-to-face mode is caused by respondents using heuristics, such as fading out some dimensions. Our first analysis on the aggregate level provided no evidence for this being the case; effect sizes of vignette variables were found to be very similar across both modes.

modes and hence more reliable. Self-administered and face-to-face interviews showed little difference in terms of item-nonresponse and response consistency, and (apart from the possible social desirability bias due to the interviewer presence) both survey modes hardly differed in meaningful results. That is, the causal impact of vignette dimensions on the vignette evaluation was mostly the same.

Nonetheless, we refrain from generally recommending the use of less expensive online surveys without interviewer assistance. It is true that in the case of regional two-stage sampling, interviewers go hand in hand with an artificial between-respondents variation due to the clustering of respondents, and survey costs for this mode are also higher because of the higher unit costs for completed interviews. However, face-to-face surveys – if implemented properly – should still produce respondent samples with lower representation errors (Biemer 2001, see also our analysis in Appendix 6.1), and they yielded somewhat more consistent responses in our study. In addition, replications with further factorial survey modules would be desirable, dealing with more sensitive topics or being better designed to disentangle sample selection and mode effects. Although we used different techniques, such as poststratification weights, to adjust both survey splits with regard to the sociodemographic composition of respondents, some differences might have remained.

We conclude with a very general statement on factorial surveys. Due to their experimental setup, factorial survey modules produce high internal validity regarding the causal influence of vignette variables whenever the randomisation of vignettes across respondents worked properly. This is true for both random and non-random samples of respondents, even though large random population samples increase the possibilities for testing possible moderator variables on the level of respondents (Aronson, Wilson, & Brewer 1998). This feature makes factorial survey designs attractive even if researchers have a somewhat biased random sample at hand or cannot build on a random sample of respondents at all.

1.5) Appendix

1.5.1) Analyses on Respondents' Compositions and Recruitment Errors

This section is dedicated to the following question: do results from the factorial survey module vary due to the different sampling procedures in both samples and, if so, how? As indicated in our book chapter, the samples for the face-to-face interviews and the self-administered mode were drawn separately using two different sampling strategies: a random route procedure and random digit dialing via phone. Although both procedures aimed to draw on a random sample of the population, the resulting respondent groups were—due to differential coverage and nonresponse—likely to differ in their sociodemographic characteristics (namely sex, age and educational background). This matters because respondent traits are likely to come with specific response behaviours. In this case, it becomes essential to take those respondents' characteristics into account when analysing the vignette module if we do not want to mix up survey mode effects with sample differences resulting from the applied sampling procedures.

Table 7. Demographic Characteristics by Sampling Mode
Before and After Design Weighting (for Different Household Sizes)

Demographic characteristics		**Census data** (Mikrozensus 2009)	Unweighted		Design weighted	
			Self-admin	Face-to-face	Self-admin	Face-to-face
Sex	Male	**48.71**	47.94	46.65	50.76	46.23
	Female	**51.29**	52.06	53.35	49.24	53.77
Age	18 to <25	**10.04**	9.02	9.63	12.79	10.90
	25 to <35	**14.28**	11.46	15.04	10.93	13.81
	35 to <45	**18.38**	16.71	20.58	17.19	19.55
	45 to <55	**18.64**	17.07	20.18	18.39	22.67
	55 to <65	**14.40**	16.95	18.21	15.06	18.68
	65 to <75	**14.53**	21.83	11.87	19.85	10.61
	75 and older	**9.74**	6.95	4.49	5.80	3.78
Highest educational level	- Pupil / left school without degree	**4.74**	3.39	4.12	4.56	4.41
	- Lower secondary	**39.66**	22.49	40.08	22.01	38.69
	- Middle secondary	**29.02**	30.11	36.21	29.28	35.99
	- Technical school	**6.21**	9.67	3.48	10.77	4.20
	- Upper secondary	**20.37**	34.34	16.11	33.38	16.71
Household size (> 18 years)	1	**24.73**	33.49	35.57	18.24	19.63
	2	**54.48**	53.20	50.90	57.93	56.19
	3	**14.06**	9.43	10.31	15.40	17.07
	4	**6.73**	3.87	3.22	8.43	7.11

All values are relative frequencies.

To quantify the potential problem, Table 7 compares both samples to the sociodemographic distributions expected in the general German population when the data collection took place. We capture those "true" distributions by means of own calculations on the basis of census data from the Federal Statistical Office, which contains a 1%-random sample of the German population with a participation rate of almost 100% ("census data", shown in the boldly printed column).

When analysing Table 7, it is difficult to make clear claims about which sampling method results in the higher sample quality because both groups deviate to some extent from the general population. Older people (aged 55 and older) seem easier to reach by phone than younger people and are therefore overrepresented in the self-administered sample. However, the most severe deviations from the general population distributions are observable for educational background. The random digit dialing sample contains a 14 percentage points higher share of highly educated people than the general population, while respondents with a lower secondary degree are undersampled by 17 percentage points. Overall, educational background is better captured by the random route sample, although this also comes with some noticeable deviations, particularly for the middle secondary degrees.

As both samples were drawn using procedures where first households and then respondents were selected, respondents in smaller households had a higher individual probability to become part of the sample. Consequently, Table 7 indicates that individuals from single-person households are clearly overrepresented in both samples, while individuals living with more than one other person participate too rarely. Even the commonly accepted practice of correcting for unequal sampling probabilities by applying design weights does not leave us with perfectly representative samples (see the "weighted" columns in Table 7. It at least does not solve the most severe problems and sometimes overcorrects existing deviations. For some of the analyses in the book chapter, we therefore decided use post-stratification weights to separately adjust our samples to the general population.

1.5.2) Full Regression Models

Table 8. Regressions of Fairness Evaluations on Vignette Dimensions by Survey Mode

	(1a) Self-admin	**(1b)** Face-to-face	**(1c)** Face-to-face + FE interviewer	**(2a)** Self-admin	**(2b)** Face-to-face	**(2c)** Face-to-face + FE interviewer
	Clustered robust standard errors (resp) ---- Post-stratification weights ----			**Clustered robust standard errors (resp) -- Controlled for respondent sex, age, educ --**		
Respondents' characteristics	-	-	-	yes	yes	yes
Sex (ref: male)	0.312**	0.185*	0.202*	0.241*	0.251**	0.250**
	(0.118)	(0.093)	(0.097)	(0.095)	(0.089)	(0.089)
Age: 40 (ref: 30)	-0.014	0.151	0.102	-0.082	0.141	0.105
	(0.134)	(0.133)	(0.132)	(0.112)	(0.116)	(0.118)
Age: 50 (ref: 30)	-0.038	0.050	-0.027	-0.228+	0.020	-0.013
	(0.184)	(0.197)	(0.201)	(0.128)	(0.133)	(0.130)
Age: 60 (ref: 30)	-0.148	0.175	0.191	-0.194	0.148	0.139
	(0.151)	(0.147)	(0.150)	(0.133)	(0.124)	(0.128)
Educ: vocational training (ref: no degree)	-0.601**	-0.311**	-0.284*	-0.529***	-0.321**	-0.319**
	(0.184)	(0.111)	(0.113)	(0.127)	(0.095)	(0.098)
Educ: university degree (ref: no degree)	-0.816***	-0.797***	-0.759***	-0.804***	-0.785***	-0.810***
	(0.115)	(0.128)	(0.137)	(0.120)	(0.111)	(0.108)
Occupation	-0.253***	-0.276***	-0.280***	-0.272***	-0.285***	-0.283***
	(0.024)	(0.027)	(0.025)	(0.020)	(0.019)	(0.019)
Earnings (log)	2.813***	3.031***	3.028***	**2.872*** **	**3.066*** **	3.055***
	(0.088)	(0.086)	(0.076)	(0.070)	(0.058)	(0.059)
Occ. experience: much (ref: little)	-0.544***	-0.420*	-0.422*	**-0.622*** **	**-0.313** **	-0.311**
	(0.115)	(0.182)	(0.193)	(0.111)	(0.109)	(0.111)
Tenure: long time (ref: short time)	-0.120	-0.076	-0.064	-0.111	-0.125	-0.103
	(0.111)	(0.150)	(0.157)	(0.103)	(0.115)	(0.114)
Number of children	-0.113**	-0.083+	-0.098*	-0.127***	-0.064+	-0.069*
	(0.039)	(0.045)	(0.044)	(0.028)	(0.033)	(0.033)
Intercept	**-3.243*** **	**-3.973*** **	-2.076***	**-2.424** **	**-4.024*** **	-1.891***
	(0.195)	(0.220)	(0.434)	(0.838)	(0.538)	(0.545)
N vignette evaluations	1438	1342	1342	1438	1342	1342
N respondents	145	135	135	145	135	135
r2	0.707	0.714	0.778	0.715	0.738	0.795
r2_a	0.704	0.712	0.765	0.710	0.734	0.781

Cluster-robust standard errors in parentheses.
Coefficients printed in **bold and highlighted in grey** differ at a 5% level; coefficients only highlighted in grey differ at a 10% level.
Only respondents evaluating 10 vignettes with 8 dimensions.
+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$