The Factorial Survey: The Impact of the Presentation Format of Vignettes on Answer Behavior and Processing Time

Hawal Shamon[1)], Hermann Dülmer[2)], and Adam Giza[3)]

1) Forschungszentrum Jülich, Institute of Energy and Climate Research - Systems Analysis and Technology Evaluation (IEK-STE), D-52425 Jülich, Germany
2) Institute of Sociology and Social Psychology (ISS), University of Cologne, Cologne, Germany
3) Independent Scholar, Cologne, Germany

Abstract

The factorial survey is an experimental design in which the researcher constructs varying descriptions of situations or individual persons (vignettes) which will be judged by respondents with regard to a particular aspect. Some researcher present vignettes in text format as short stories, others present the central information of vignettes in a tabular format. To date only a few sentences have been published, by Auspurg and Hinz (2015), on the impact of the presentation format (text vs. table) on the answer behavior of students. Empirically, no differences were found between either format. Based on an internet experiment conducted with a quota sample we find evidence that ordinary tabular formats outperform text vignettes in terms of total vignette nonresponse, but not when it comes to processing time. The former result especially applies in the case of less well educated people. We further find that tabular format do not perform worse than text-format regarding response inconsistency.[2]

Keywords

Factorial Survey, Satisficing, Text Vignette, Table Vignette, Response Time, Nonresponse, Cognitive Scheme, Design, Methodological Study

Please use following reference for citation purposes:
Shamon, H., Hermann Dülmer, Adam Giza (2019). "The Factorial Survey: The Impact of the Presentation Format of Vignettes on Answer Behavior and Processing Time." Sociological Methods & Research. Firstonline. **https://doi.org/10.1177/0049124119852382**

[1] h.shamon@fz-juelich.de, phone: +49 2461 61-3322, fax: +49 2461 61-2540
[2] The authors thank the Research Training Group 'Social Order and Life Chances in Cross-National Comparison' (SOCLIFE) at the University of Cologne, which is funded by the German Science Foundation, (DFG) for funding this study. We thank the anonymous reviewers for their constructive comments and suggestions.

1. Introduction

The question-answer-process for standardized interviews starts with reading or listening to a question. It ends when an answer is given by a respondent. The survey response process consists of separate steps which can be arranged into four main categories (cf. Tourangeau et al. 2000): 1. Understanding a question, 2. Retrieving from memory information subjectively perceived as relevant, 3. Forming a judgment based on the retrieved information, and 4. Communicating the judgment to the interviewer or researcher, which in the case of a closed answer scale has to be adapted to the given answer format.

In this survey response process, the second step (i.e., retrieval of relevant information) forms a crucial basis for a stable judgment. Its importance becomes clear by elaborating on the retrieval step in case of attitude questions. When asked for their attitudes, respondents are sometimes confronted with a topic that they have rarely or even never been concerned about before. Under these conditions it is rather unlikely that a preconceived attitude that can be easily retrieved from memory exists (Porst 2008; Tourangeau et al. 2000). When respondents have no preconceived attitude it can be assumed that they arrive at an answer by, for instance, using available information from diverse sources (Porst 2008; Saris and Gallhofer 2007). Tourangeau et al. (2000) highlight three compatible ways in which respondents can come to an answer when they are asked for their attitude. Firstly, without detailed knowledge of a topic, people can base their judgment on vague impressions or existing stereotypes. Secondly, they can align their judgment with their own value orientations. Finally, in order to reach a judgment respondents can also search for concrete information which is personally considered to be relevant.[3] Based on these considerations, for conventional questions about attitudes that are asked of the same respondents repeatedly across time, a high variability can be expected in the answers (Converse 1964; Van der Veld and Saris 2003). The less a respondent has a preconceived attitude towards an issue or object, and the more the person's attitude starts to form within the interview situation, the more plausible this expectation becomes. Moreover, these elaborations indicate that respondents may significantly differ with regard to the subjective basis of their judgment (hereafter called *reference frame*).

Instead of measuring judgments directly via general survey questions, judgments can be measured indirectly by using a factorial survey where the survey response process does not require information retrieval from memory. The factorial survey (Rossi and Anderson 1982; Jasso 2006; Auspurg and Hinz 2015) consists of a number of descriptions of various hypothetical situations, objects, or persons (vignettes) that have to be answered by the respondents with reference to a particular aspect the researcher is interested in (for instance, normative judgments, subjective beliefs, or intended actions, cf. Jasso 2006:344). The descriptions in a vignette are designed as an experimental setting. As such each vignette includes several dimensions (also called factors) and one of several possible levels for each individual dimension. A factorial survey about fair levels of earnings for salesmen, for instance,

[3] For example, respondents asked to assess the prospects for the local economy based their responses on the local unemployment rate and prospects for employment growth (Mason and Carlson 1994).

may include the dimensions of gender, parenthood, job experience, and level of effort at work for a fictitious vignette person. In this case, a single vignette would inform a respondent about the combination of the levels (personal attributes/characteristics) of a fictitious vignette person for these four dimensions (for example, female, 2 children, 5 years' job experience, high effort at work). The task of the respondents is to judge concrete vignette descriptions as a whole without being forced to indicate the influence of each individual vignette characteristic explicitly. An advantage of the factorial survey is that judging concrete vignettes comes much closer to judgments made in real life than answering comparably general, mostly rather abstract questions as usually used for survey research (cf. Beck and Opp 2001:304). The detailed descriptions also lead to a higher standardization (cf. also Auspurg and Hinz 2015:4, 7). This helps to minimize the problem of respondents' different subjective reference frames because in order to answer a question, respondents no longer have to retrieve relevant information from their memory (step 2 of Tourangeau et al. 2000), but to take in information from a standardized reference frame (cf. Shamon 2014). Taken together, these facts might explain why in the factorial survey a respondent's answer can be expected to be ascertained with higher reliability and higher validity[4] than is possible with more general single questions more typical of survey research (cf. Alexander and Becker 1978:93).

Although more recently factorial surveys have become increasingly popular among social scientists, only a small number of the published studies based on population or college samples (Auspurg and Hinz 2015:2) examined the effect of methodological issues of design complexity on respondents' cognitive load for different administration modes (Auspurg et al. 2009; Sauer et al. 2011; Teti et al. 2016; Auspurg and Jäckle 2017). Overall, these studies support the use of the factorial survey by showing that manifestations of cognitive overburden – such as of response inconsistencies, fade-out of vignette dimensions, fatigue effects, and vignette dimension order effects – can be avoided by choosing an appropriate level of design complexity.[5] One area where no published studies exist up to now is the question of whether the *presentation format of vignettes* has any impact on the cognitive burden and, thus, on the answer behavior of respondents (cf. also Auspurg and Hinz 2015:71). Research on this topic is important because information intake is central to vignette studies. Basically, vignettes can be presented in a *text format* (running text) or in a *tabular format* (cf. Jasso 2006:411-412).

Besides *varying information*, some factorial surveys might also consider to include *fixed background information* (only one level for a dimension). Fixed background information refer to those dimensions that are from a theoretical point of view assumed to be relevant for

[4] A basic idea of factorial surveys is to combine the high internal validity of experimental designs with the high external validity of survey research (cf. Rossi and Anderson 1982:15-16; Sniderman and Grob 1996:377-378; Atzmüller and Steiner 2010:128). The experimental setting of a factorial survey increases the internal validity (the observed variation in the outcome variable is caused by the experimental stimuli). The high external validity of survey research is based on using a random sample from the population, which allows the results to be generalized to the population as a whole. If a non-random sample is used, the factorial survey still permits general conclusions about causal mechanisms (high internal validity, cf. also Auspurg and Hinz 2015:11-12, 60-64).

[5] For instance, Sauer et al. (2011) recommend for factorial surveys administered to general population samples to use 20 or less vignettes with less than 12 dimensions.

respondents' answer behavior, but whose examination is not central to a researcher's substantial interest in a particular study. However, although the researcher is neither substantively interested in fixed background information nor able to estimate its impact on the answer behavior, fixed background information is sometimes needed to standardize the vignettes sufficiently. Research suggests that people's short-term memory is limited to store 'seven, plus or minus two' items of information at a time (Miller 1994). Therefore, presenting fixed background information in the introduction of a factorial survey, as proposed by Auspurg and Hinz (2015:19), might not be sufficient to prevent from biases in the effect sizes, particularly in case of complex factorial survey designs. Specifying additionally fixed background information in a visually separated space at each vignette has the advantage that the information is always available and, if needed, can be read anew. In this way, possible ambiguities which otherwise might exist (and which might bias the results) are removed from the vignettes. Another advantage of using fixed background information is that in this way the complexity of the factorial survey is reduced.

In order to facilitate information intake, the researcher might decide to present the varying information for text formats, for instance, by underlining or by using a tabular format. A rationale for using a text format, according to Auspurg and Hinz (2015:70), is that short stories might facilitate the understanding of the described situation. It might help respondents to put themselves in the situation or to empathize with the described person. The initial research of Auspurg and Hinz (2015:71) with university students, however, suggests that tabular vignettes, if not overtly complex, result in similar evaluations to text vignettes. Although the authors found no substantial differences for highly educated students, important differences might not become visible until people with different educational backgrounds participate in a factorial survey. From past research it is well known that the more syllables a word contains and the higher the number of words that are used in a sentence, the more difficult it becomes for people to understand the sentence (cf. Flesch 1948). This aspect is especially relevant for less well educated people with poor reading skills, whereas people with good reading skills are better able to understand semantically complex sentences (cf. Knäuper et al. 1997). By reducing the number of words and by placing the relevant information at exactly the same location on a vignette (a predefined column), table formats contribute to make information more accessible to respondents (cf. also Hainmueller et al. 2015:2396). This might help respondents to extract the potentially relevant information easier and, by giving a better overview over the varying information, may contribute to a betterunderstanding even for respondents with lower cognitive skills.[6] In a direct comparison with an unsupported text format (varying information not highlighted, for instance by underlining) the tabular format performed better than text format regarding predicting real-world behavior, i.e., the tabular

[6] An advantage of using a tabular format is, according to Auspurg and Hinz (2015; cf. also Auspurg and Jäckle 2017), that it allows to randomize the vignette dimensions in order to avoid possible order effects. For text vignettes, this might conflict with a smooth text flow. The substantive logic of the order of the vignette dimension, however, does probably in most cases also for table vignettes not allow a randomization of the dimensions. It, furthermore, might also confuse the respondents.

format showed the higher external validity (cf. Hainmueller et al. 2015). Nevertheless, the tabular format might not be practical for every research question.[7]

The aim of this study is to investigate the impact of different presentation formats (text vs. tabular) on the answer behavior as well as on the processing time of respondents. We examine the formats' impacts on the probability of the respondent refusing to participate in the factorial survey, on the probability of the respondent breaking off their participation in the factorial survey, on the probability of unanswered vignettes, on the number of faded-out vignette dimensions, on coefficient weights of vignette dimensions, on respondents' response inconsistency as well as the impact on the processing time taken by the respondent. The examined answer behaviors – refusal, break-off, and vignette nonresponse – may be understood as an expression of a respondent's strong satisficing strategy, while number of faded-out vignette dimensions, differences in coefficient weights and response inconsistency point to the potential impact of the presentation format on respondents' choice of a weak satisficing strategy for reducing the cognitive demands of an interview. Less well educated people and older people might be especially prone to such strategies. If these groups are systematically affected by a higher rate of invalid or missing answers, then the data quality will suffer from a systematic bias.

The topic of this study is important: firstly, because a longer processing time increases the costs of a survey. Secondly, having better knowledge about the impact of the different formats on the cognitive demand of the survey makes it possible to reduce the percentage of missing values, which in turn contributes to better data quality (higher statistical power, lower risk of potential bias, cf. also Shoemaker et al. 2002; Rubin 1976). Our survey is based on the substantive ideas of a factorial survey conducted by Shamon and Dülmer (2014). The participants for our study were recruited via quota sampling from a German online-access-panel. The quotas were generated by crossing age, gender, and education. The 498 respondents who participated in the online survey were randomly assigned to one of 3

[7] Tabular formats may not be useful if researchers need to specify vignette dimensions by using one or more sentences to ensure the pragmatic understanding by respondents. For instance, in a country where sudden electricity import interruptions have never lead to unplanned interruptions in the power supply (a situation that occurred for example in Ukraine), it might be insufficient to describe on the vignettes that a hypothetical country imports 10% (vs. 70%) of the consumed electricity from another country. Instead, it is advisable to operationalize this energy security issue by a sentence such as: "Country A is dependent on other countries, because it imports 10% (vs. 70%) of its electricity consumption.". One might also argue that the tabular format cannot be applied for face-to-face interviews, since only the text format allows the interviewee to understand the vignette text read aloud by the interviewer. However, for face-to-face interviews it is best practice to support information intake by using showcards (cf., for instance, the European Social Survey 2016). Without using showcards for the vignettes, one might furthermore assume that neither the table format nor the supportive text format would be applicable any longer. Although abandoning the use of vignette showcards is surely not ideal, highlighting words of text vignettes might also be done by emphasizing the pronunciation of the varying text information or for table vignettes by informing the respondents that the interviewer will read aloud different categories like income and thereafter specify the respective amount for different situations or persons. How far table formats work for developing countries with a different cultural background (for research on factorial surveys, for instance, conducted in different African countries, cf. Liebe et al. 2017) remains a question that should be addressed by future research.

experimental groups. In the following, we will give an overview of the theoretical frame of the study. Thereafter, we will present a modified cognitive scheme for the 'question-answer-process' in factorial surveys, which was introduced by Shamon (2014). Based on this theoretical framework, we will derive our hypotheses. Finally, we will introduce our research strategy, test the hypotheses, discuss the results, and draw some conclusions.

2. Theoretical Framework

2.1 Cognitive Schemes for Factorial Surveys and Answer Behavior Categorization

Participating in a survey results in costs, which arise, among other things, from cognitive efforts that are required for answering questions. Based on the anticipated costs as well as the subjectively expected net utility, a respondent will decide either to participate or not to participate in a survey (cf. Dillman 1978). People who decided to participate, however, can a) work through the cognitive steps carefully and comprehensively (*optimizing*), b) try to reduce the costs by completing the cognitive steps involved in answering questions with relatively little care (*weak form of satisficing*) or c) try to reduce the costs by omitting at least one of the required cognitive steps (*strong form of satisficing*, cf. Krosnick 1991).[8] The probability of using a satisficing strategy depends on the requirements of a task/question as well as on the skills and the motivation of the person responding to a respective question. Krosnick (1991) assumes that the more time has passed, the lower the motivation of a respondent to complete all cognitive steps carefully. Krosnick et al. (2002) showed that a respondent's ability and motivation as well as the position of the question in the survey determine strong forms of satisficing that manifest themselves in the absence of a respondent's answer to a survey item.

In factorial surveys, participants are presented with a single description (one-vignette-per-respondent factorial surveys) or a set of varying descriptions of hypothetical situations or persons (multiple-vignettes-per-respondent factorial surveys). They are asked to evaluate the presented vignette(s) with regard to a particular aspect the researcher is interested in. In this context, the question and answer process proposed by Tourangeau et al. (2000) cannot be used without modifications. Shamon (2014) proposed a modified cognitive scheme for factorial surveys. According to this scheme, instead of understanding a question, the first step in the question-answer-process is information intake. Besides the varying information, which constitutes the core of the experimental design, some vignette studies also present fixed background information for each individual vignette. From the respondent's perspective the (fixed) background information from the first vignette only has to be stored in the memory once and then retrieved every time a subsequent vignette description is evaluated. Varying

[8] Choosing the first best solution is a *weak form of satisficing* since all cognitive steps have been completed, although not with due diligence. Choosing a 'don't know'-category in order to avoid the step of judgment formation based on retrieved information is seen as a *strong form of satisficing* (Krosnick 1991). Constant answer behavior across different items of an item battery is, at least sometimes, another possible form of *strong satisficing*.

information on the other hand has to be absorbed from each individual vignette. Without this information intake, varying vignette characteristics cannot have any impact on judgment behavior. In general, whenever fixed background information is used in a factorial survey, it is recommended that respondents are informed about the fixed vignette characteristics in the introduction, presented before the factorial survey starts.

After information intake, reading and understanding the vignette question follows as a second step.[9] Like survey questions, vignette questions have to be formulated in a way that allows the respondents to unambiguously understand the semantic and pragmatic sense of the question.[10] In a third step, respondents have to form a judgment. Since respondents are free to decide which of the presented information they want to take into account and how strongly they want to weight the respective information, the information presented in the vignettes can be understood as the objective basis for judgment formation. In this sense it is possible that respondents consciously or unconsciously ignore, partially or even completely, in their subjective reference frame some of the objectively given vignette information in the judgment formation step because they perceive it as irrelevant or because they want to reduce cognitive effort in answering the vignette questions. In the subsequent fourth step, the judgment has to be communicated to the interviewer. For factorial surveys, open or closed answer formats can be used. In contrast to open answer formats that require the interviewee to form a precise answer to a judgment task while allowing him or her to precisely communicate his or her optimal judgment, closed answer formats require the selection of the best-fitting answer category out of an answer scale predefined by the researcher before the judgment can be communicated (Sudman et al. 1996).

Based on the theoretical considerations, the information in a vignette provides respondents with all the information needed to form a judgment. Hence, a respondent omitting to make a judgment reflects the application of a strong satisficing strategy rather than a valid indication that he or she has no judgment at all (cf. Krosnick et al. 2002:379). In one-vignette-per-respondent factorial surveys it is only possible ex post to distinguish between respondents who judged and those who did not judge the single vignette, either by skipping a question or by choosing a potentially offered 'don't know' option. The analysis of the answer behavior in multiple-vignettes-per-respondent factorial surveys allows for a much more nuanced analysis of the strong form of satisficing. Here we can distinguish ex post between a) participants who applied a strong form of satisficing (*refusal*) to all presented vignettes, b) participants who started with an optimizing strategy or with a weak form of satisficing and changed to a strong form of satisficing (*break-off*) during the factorial survey and c) respondents who alternately apply a weak form of satisficing or optimizing and a strong form of satisficing (*vignette nonresponse*).

[9] A vignette question is a question that follows and refers to a vignette and makes explicit, from which aspect the vignette is supposed to be judged by a respondent.

[10] Semantic understanding means that it must be clear to a respondent what the meaning of a question, a formulation, or a term is; pragmatic understanding implies that it should be clear to a respondent what the researcher or interviewer intends to find out with the question (Porst 2008:18-23).

Beside answer behaviors rooted in strong forms of satisficing, multiple-vignettes-per-respondent factorial surveys allow also for a nuanced analysis of weak forms of satisficing. Multiple-vignettes-per-respondent factorial surveys allow ex post to distinguish between respondents who reduce the cognitive costs of completing the cognitive steps involved in answering questions a) by reducing the number of vignette dimensions considered in their subjective reference frame right from the beginning of the factorial survey (*consequent dimension reduction*)[11], b) by changing their subjective reference frame from vignette to vignette (*partial dimension reduction*) and c) by paying less attention on responding consistently to the vignette dimensions presented on the multiple vignettes to them (*response inconsistency*; cf. also Auspurg et al. 2009; Sauer et al. 2011; Teti et al. 2016).[12]

An advantage of using our nuanced classification scheme is that not only does it allow the total amount of vignette nonresponse to be analyzed but also makes it possible to identify the most important sources of different forms of strong and weak satisficing behavior across different presentation formats.

2.2 The Impact of the Presentation Format on Answer Behavior

Presentation formats of factorial surveys (tabular vs. text vignettes) differ in the degree to which they pre-structure the varying information on the vignettes. By choosing an optimal presentation format, information intake by respondents will be facilitated, which contributes to valid answer behavior. A presentation format that supports the information intake of the varying information (*supportive presentation format*) should reduce a respondent's expected as well as real costs in terms of cognitive effort and time spent participating in a survey. In so doing, a supportive presentation format should help to lessen a possible decrease in respondents' motivation over the time spent doing the survey. Therefore, our general expectation is that the likelihood of choosing a strong satisficing strategy – whether intentional or not – will be reduced by using a supportive presentation format. In a similar vein, we expect that the use of a supportive presentation format would reduce the likelihood of choosing a weak satisficing strategy and, hence, contribute to preventing consequent dimension reduction, partial dimension reduction, and response inconsistencies. Moreover, we expect that a supportive presentation format allows for faster information intake such that it reduces the processing time of respondents who complete all cognitive steps, i.e., without applying a strong satisficing strategy.

In text formats, the information intake of respondents can be facilitated by highlighting the varying information through the use of underlining or bold letters for instance. In this way, the

[11] Consequent dimension reduction does only reflect weak satisficing if respondents aim at reducing cognitive effort in answering the vignette questions.

[12] It should be mentioned that one can expect a trade-off between consequent dimension reduction and response inconsistency (cf. also Auspurg et al. 2009:89) as well as between partial dimension reduction and response inconsistency. While consequent dimension reduction can be expected to decrease response inconsistency, partial dimension reduction can be expected to increase response inconsistency.

varying information is already emphasized (pre-structured) and the text format becomes supportive, while a text format without highlighted varying information is a non-supportive one. Using a tabular format for conducting a factorial survey is always a supportive format: in this case the varying information is already spotlighted (pre-structured) by presenting it in a specific column in each vignette (*ordinary tabular presentation format*). The support given by a tabular format is probably more structured than when highlighted varying information is used in text vignettes. Compared to text vignettes, where varying information is highlighted but spread throughout the vignette text, table vignettes display all relevant information, further organized in columns and rows, which probably allows respondents to get a faster and better overview of all varying information. This is generally the advantage of using tables. Based on these theoretical considerations we will now formulate specific hypotheses about text and tabular vignette formats.

H1a: Tabular vignette formats and supportive text formats are expected to outperform non-supportive text formats in terms of refusals, break-offs, vignette nonresponse total vignette nonresponse, consequent dimension reduction, partial dimension reduction and response inconsistency (*non-supportive-text-format-satisficing hypothesis*).

Since tabular vignette formats organize the varying information in a specific vignette area, the support for information intake is seen as stronger for tabular vignettes than for text vignettes where varying information is only highlighted in a block of text. Hence, we have the following expectation:

H1b: Tabular vignette formats are expected to outperform text vignettes with regard to refusals, break-offs, vignette nonresponse total vignette nonresponse, consequent dimension reduction, partial dimension reduction and response inconsistency (*tabular vs. text format-satisficing hypothesis)*.

Facilitating information intake should allow for a faster information intake and thus reduce the processing time of respondents who complete all cognitive steps. In accordance with H1a and H1b we expect the following hypotheses concerning the processing time:

H2a: Tabular vignette formats and supportive text formats are expected to outperform non-supportive text formats regarding the processing time (*non-supportive text format-time hypothesis*).

Since tabular vignette formats organize the varying information in a specific vignette area, the support for information intake is seen as stronger for tabular vignettes than for text vignettes where varying information is only highlighted in a block of text. Hence, we have the following expectation:

H2b: Tabular vignette formats are expected to be superior text vignettes with regard to processing time (*tabular vs. text format-time hypothesis)*.

The task of picking out relevant information is assumed to be more difficult and therefore more costly as well as less motivating, the lower the reading competence of a respondent. Based on these considerations, in our *presentation format-education hypothesis* we expect that comparably less well educated respondents, i.e., people with relatively low reading competence,[13] will profit most from supportive presentation formats:

H3: The less well educated a respondent is, the higher the likelihood a satisficing strategy will be chosen. The expected difference between the less well educated and the more highly educated, however, should be smaller for a supportive presentation format than for a non-supportive presentation format.

From psychological research it is well known (cf. Hartshorne and Germine 2015) that due to cognitive ageing, information intake generally takes longer for older people. Therefore, answering survey questions, ceteris paribus, should be more time-consuming and for this reason also more costly for older than for younger respondents. Based on this knowledge we will also test in our *presentation format-age hypothesis* whether older people profit more from supportive presentation formats than younger people do:

H4: Older people are more likely to choose a satisficing strategy than younger people. This difference, however, should be smaller for a supportive presentation format than for a non-supportive presentation format.

3. Operationalizations and Research Design

3.1 Vignette Study

The present study is based on a factorial survey carried out in 2012 by Shamon and Dülmer (2014). During the internet-based survey each participant had to evaluate a set of 16 vignettes, each describing a fictitious industrial sales representative, 35 years old, married to a spouse who is not gainfully employed, and with a monthly gross salary of €3,000. The five underlying dimensions (employee's occupation, employee's age, employee's marital status, spouse's occupational status, employee's gross salary) were identical in each of the vignettes (fixed background information). These five dimensions were necessary to sufficiently specify the hypothetical situations and in this way to rule out different interpretations of the situation across our respondents. The vignettes varied with regard to 6 dimensions of substantial interest: gender (male, female), parenthood (no or two children), job experience (5 or 10 years), level of effort at work (low or high), regional average salary (€2,622, €3,009, €3,450),

[13] According to the empirical results of Maehler et al. (2013) a strong relationship exists between reading competencies and formal education. Education instead of reading competence is used for this study because for an online survey it is more difficult to capture reading competencies. Furthermore, good items for measuring such skills developed by PIAAC (2016) are not published.

and regional pay inequality (range of gross salary €2,300 to €4,100 or €1,600 to €5,400).[14] Table 1 gives an overview of all 6 dimensions, their levels, and the coding of the levels.

Table 1: Vignette Dimensions, Vignette Characteristics and Coding of the Vignette Characteristics

The person H. O. (Industrial sales representative, aged 35, married, spouse does not work) has further characteristics:		
Factors or dimensions describing the fictitious vignette person	Levels (vignette characteristics) for the dimensions	Dummy coding (0: reference category)
Gender	Male	0
	Female	1
Own children	No children	0
	2 children	1
Job experience	5 years	0
	10 years	1
Effort at work	Low	0
	High	1
In the region in which H. O. works the following holds for industrial sales representatives:		
Factors or dimensions describing the region where the fictitious vignette person lives	Levels (vignette characteristics) for the dimensions	Dummy coding (0: reference category)
Average gross salary	€2,622	1
	€3,009	0
	€3,450	1
Pay inequality (range of gross salary)	Between €2,300 and €4,100	0
	Between €1,600 and €5,400	1

Note: The initial letters of the names of the fictitious vignette persons were generated randomly in order to increase the impression that different persons were described. The initials, however, do not represent a substantial dimension (which might have been the case if full names had been used instead).

The fully crossed vignette universe of our factorial survey consists of 96 ($=2\cdot2\cdot2\cdot2\cdot3\cdot2$) vignettes. Among these vignettes there were no implausible combinations. An interaction between one of the dichotomous and the trichotomous variables was expected. While Bose and Rossi (1983) expose household as well as college sample respondents in personal interviews to 110 vignettes, Rossi and Anderson (1982:41-42) fear that participants might refuse even to start answering a factorial survey when each respondent is expected to judge more than 25 vignettes in a mail survey. Beck and Opp (2001:291) assume that for a

[14] In order to avoid potential confusion about the meaning of the dimensions and levels, which results in a potential bias of the estimated effects, dimensions and levels should ideally be formulated very precisely, which allows all respondents to interpret them unambiguously (cf. also Auspurg and Hinz 2015:21). This is easy to realize in cases where natural units (for instance, number of children) or established predefined units (for instance, income in Dollar) exist. In situations, however, where natural units do not exist, even textbooks prefer, for instance, to use relatively vague levels as 'only few' and 'a lot of' for job experience or 'short' and 'long' for job tenure (cf. Auspurg and Hinz 2015:21). Since in our factorial survey there exists no natural or predefined unit for the level of work, we used instead the terms 'low' and 'high' for the reason that the participants are familiar with these labels from daily life conversations.

representative survey, the number of vignettes per respondent should not exceed the range of 10 to 20 vignettes. Methodological research on this issue supports these suggestions by recommending not to use more than 20 vignettes per respondent for factorial surveys in general population samples (Sauer et al. 2011).[15] Within the given limit for a reasonable set size (number of vignettes to be judged by a single respondent) no suitable fractional factorial design (Alexander and Becker 1978; Gunst and Mason 1993) could be found. Instead, a D-efficient design (Kuhfeld et al. 1994; cf. also Dülmer 2007; Dülmer 2016) was generated by using the computer program SAS. If all vignette variables are standardized orthogonally contrast- coded (Kuhfeld 2010:74), then D-efficiency is scaled to the range from 0 to 100 (Kuhfeld et al. 1994:547, 549). A D-efficiency of 100 will only be reached by balanced orthogonal designs. 'Balanced' means that for each vignette dimension (including interaction terms which are expected to be important and for this reason are part of the design), the chosen levels within a vignette set appear with equal frequency; 'orthogonal' means that variables of different dimensions (including important interaction terms) are uncorrelated (Kuhfeld et al. 1994; cf. also Dülmer 2016). In our case, a local maximum for D-efficiency was reached for a set size of 16 vignettes per respondent. The D-efficiency for the selected quota design is 96.6591.[16]

3.2 Experimental Settings

To test our hypotheses concerning the impact of different presentation formats for vignettes, three experimental settings were created: two for text vignettes (*Settings 1 and 2*) and one for tabular vignettes (*Settings 3*). The difference between Setting 1 and 2 is that varying information was underlined for Setting 2 in order to facilitate information intake (supportive presentation format), but not for Setting 1 (non-supportive presentation format), while in Setting 3 varying information was presented in a column (*ordinary tabular presentation format*). Hence, with respect to information intake, a rank order is expected for the three settings: *The higher the setting number, the easier information intake should be.* In all settings, an introductory text was shown prior to presenting the 16 vignettes. Table 2a shows the introductory text, while Table 2b and 2c depict an example vignette for different formats (a text and a tabular format respectively).

Table 2a: Introduction to the Factorial Survey shown in each Setting

On the following pages we will present **16 individuals** from 16 different regions to you. Every person is **35 years old** and earns **€3,000 per month gross** as an **industrial sales**

[15] This recommendation should not be misunderstood as a golden rule that applies to all factorial surveys. Rather, researchers are encouraged to reflect on the cognitive abilities of their target population, the survey mode, on the cognitive demands of their substantive question and on practical reasons (e.g., the necessity to conduct respondent specific estimations) when deciding on a reasonable set size for their research endeavor.

[16] For further details, see Shamon and Dülmer (2014).

representative. Furthermore, all individuals are happily **married** to a **partner who does not work**.

However, the 16 people differ with regard to other personal characteristics. Furthermore, the working conditions in the regions in which each person lives are also different.

For each person, we will ask you to indicate what a fair salary should be.

Table 2b: Example Text Vignette Setting 2

The person H. O. is an industrial sales representative, aged 35 years and married. His spouse does not work. Mr H. O. has 2 children. He has 10 years of work experience and is well known for the high level of effort he puts in at work.

In the region in which H. O. works, the highest gross salary among industrial sales representatives is €4,100. On average they earn €3,009. The lowest gross salary is €2,300.

Note: For Setting 1 no underlining was used.

Table 2c: Example Tabular Vignette Setting 3

The person H. O. (industrial sales representative, aged 35 years, married, spouse does not work) has the further characteristics:	
Gender	Male
Own children	2
Job experience	10 years
Effort at work	High
In the region in which H. O. works, the following holds for industrial sales representatives:	
Highest gross salary	€4,100
Average gross salary	€3,009
Lowest gross salary	€2,300

Participants of each of these three settings were asked, to fill in for each of the described fictitious vignette persons the gross salary which, in their opinion, would be a just earning. The answers had to be entered in an open text field.[17] Open answer formats have been shown to produce more missing data than closed answer formats (Schuman and Presser 1979, 1996; Urša et al. 2003) and, hence, are more suitable for nonresponse analyses. Table 3 displays the open answer format. The entry for the gross salary reflects a participant's normative judgment

[17] The open text field allowed respondents to enter numeric values between €0 and €15,000. This restriction was implemented in order to prevent typing errors. Values outside this range produced an automatic warning.

– that is a respondent's notion of how things should be (cf. Hermkens and Boerman 1989; Shamon and Dülmer 2014) – and hereafter will be called 'justified salary'.

Table 3: Open Answer Format

The described person, H. O., earns €3,000 gross a month. How much should H. O., in your opinion*, justly earn*?	
☐ Euros a month.	☐ Don't know

3.3 Data

All participants of the present internet-based study that was conducted at the end of September and the beginning of October 2014 are members of an online access panel. The target population consisted of persons aged 20 to 69 who were living in Germany in 2014. For sampling the respondents a combined quota scheme for age, gender, and education was applied. Quotas are based on the German Census of 2011. For the completion of the survey, participants were paid a small monetary incentive (€2) by the polling agency. All in all, 498 persons who were assigned randomly to one of the three experimental settings participated in the study. In order to prevent a potential bias caused by systematic order effects, the order of the 16 vignettes of the D-efficient design was randomized for each respondent (cf. Jasso 2006:343; Auspurg and Hinz 2015:72-73).

3.4 Operationalizations and Method for Analyzing the Data

The operationalization of different strong and weak satisficing strategies follows our theoretical distinction between *refusals*, *break-offs, vignette nonresponse, total vignette nonresponse, consequent dimension reduction*, *partial dimension reduction*, and *response inconsistency*.

Refusals: In factorial surveys, we observe different answer patterns that can be categorized as refusals. Besides participants who evaluate none of the vignettes (implicit refusal in the original sense[18]), some of the respondents may show no variation in their answers to different vignettes (invalid constant answer behavior)[19], while other respondents answer all 16

[18] AAPOR (2015:37) recommends to classify respondents of internet surveys who visit the survey's URL and log in with an ID and/or password, but leave all survey items blank without providing any explanation as implicit refusals. In the same vein, AAPOR (2015:29) recommends to classify respondents of mail surveys who return entirely blank questionnaires without any remarks as to why the questionnaire was returned as implicit refusals.
[19] In some factorial surveys, constant answer behavior might be a valid answer pattern: in a factorial survey about discrimination, for instance, constant answer behavior could indicate that a respondent treats all people equally (no discrimination). In our case, constant answer behavior may indicate a respondent's preference for the principle that everybody should be paid equally irrespective of personal or contextual factors. However, we could

vignettes with 'don't know'[20]. What these three answer patterns have in common is that respondents omit at least one of the four cognitive steps of our modified cognitive scheme for vignette studies from the first to the last vignette of the factorial survey, which is why these three answer patterns can be traced back to a strong form of satisficing strategy (cf. Krosnick 1991), i.e., to a respondent's non-participation in the factorial survey. Furthermore, in each of these three response patterns, respondents do not provide any explanation as to why they do not participate meaningful in the factorial survey. For this reason, all participants who evaluated none of the vignettes (either implicit refusals in the original sense or 'don't know') or showed no variability in their answers (invalid constant answer behavior) over the 16 vignettes have been coded as refusals (*refusal*=1) on a 0-1 coded dummy variable on the respondent level. Furthermore, we screened the answers of respondents whose standard deviation for the just earning was lower than €150. In cases in which we had good reasons to assume that the observed variation was caused by a simple typing error (e.g., a respondent answered to 15 vignettes €3,000 and to one vignette €300) we categorized these answers as invalid constant answer behaviors.[21,22]

Break-off: Another form of answer behavior can be traced back to a decreasing willingness to complete all four steps of the modified cognitive scheme. Such a decline in motivation could be caused by a high cognitive demand of the task so that participants become increasingly fatigued or even bored. As a result, respondents might simplify the task at a certain point by switching to a strong form of satisficing: they either completely stop answering the questions or they stop their valid answer behavior by switching to an invalid constant answer behavior. Respondents with such answer patterns were coded on a respondent-level 0-1 dummy variable as stopping the factorial survey (*break-off=1*) if at least the first vignette was answered and stopping occurred after the first or at the latest before the 16th vignette.

Vignette nonresponse: A further answer behavior consists in alternately judging a vignette and not judging a vignette. Skipping vignettes indicates a strong form of satisficing for the unjudged vignettes. For the evaluated vignettes it remains unclear whether the respective respondent applied a weak form of satisficing or whether he or she followed the required cognitive steps carefully. In such cases, all judged vignettes might be included in the analysis.

---

logically rule out such an interpretation, because none of the respondents with constant answer behavior stated a clear preference for the equality principle in the conventional survey item following the factorial survey.

[20] While a 'don't know' option conceptually allows respondents to give an adequate answer to a question when information retrieval does not result in sufficient information for judgment formation (cf. also Norman 1972), this response option also constitutes an opportunity for respondents to implement strong satisficing by saying 'don't know' (cf. Krosnick et al. 2002).This means that 'don't know' in response to a survey question is often "another way of saying 'I don't want to get involved'" (Bogart 1967:344). In our self-administered factorial survey, vignettes provide respondents with (sufficient) information to respond to a relatively unambiguous judgment task. Therefore, it can be assumed that respondents choose the 'don't know' option because they are not motivated to form a judgment.

[21] According to this operationalization, respondents who rated fewer than 16 vignettes in a constant manner are also counted as refusals.

[22] Following this procedure, we categorized the answer behavior of two respondents in Setting 1 and of one respondent in Setting 3 as invalid constant answer behavior. The results of our study presented in following section were robust towards these changes.

To analyze the impact of the presentation format and/or answer format on the probability of an invalid answer being given, a 0-1 coded dummy variable was computed. Vignettes of respondents who neither refused to answer the factorial survey nor broke off their participation, were coded as 1 on a vignette-level dummy if a vignette was answered with 'don't know' or was left unanswered (*vignette nonresponse=1*).

Total vignette nonresponse: Certainly, a distinction between refusals, breaking-offs, and vignette nonresponse identifies the main sources of invalid answer behavior. Nonetheless, what is actually most important to researchers in the end is the total loss of information. Therefore, a further vignette-level 0-1 coded dummy variable was computed for vignettes that were judged with an invalid answer ('don't know' or unanswered), in this case, however, independently of whether the respondent was one of the participants who refused or stopped answering the factorial survey (*total vignette nonresponse=1*).

For analyzing refusals and break-offs we estimated a respective logistic regression at the respondent level. The same was done at the vignette level for vignette nonresponse and total vignette nonresponse, respectively.[23] For analyzing the weak satisficing strategies and the total processing time, we restricted our sample to respondents who never used a strong satisficing strategy when they answered the factorial survey (*non-strong-satisficing*), because it is plausible to assume that the different answer behaviors encompassing a strong form of satisficing are differently correlated with the settings as well as with weak satisficing and processing time. Hence, the restriction to non-strong-satisficing participants is expected to increase the homogeneity of the analysis sample and, thus, the meaningfulness of our results regarding the following indicators:

Consequent dimension reduction: When respondents fade out a certain vignette dimension from the first to the last vignette, the respective vignette dimension has no impact on their answers (i.e., the effect size of the ignored vignette dimension is virtually zero). In order to determine the existence of consequent dimension reduction, we regressed for each respondent the answers to the 16 vignettes on the vignette dimensions using OLS and saved the resulting respondent specific weights of the vignette dimensions in a data file.[24] Subsequently, we coded a respondent-level dummy variable for each vignette dimension as 1 if the respective vignette dimension was faded out by the respondent (*consequent dimension reduction=1)* and analyzed it in a logistic regression as a dependent variable. As mentioned above, consequent dimension reduction might take place due to justified reasons (information is unimportant to the respondent) or as an attempt to reduce cognitive efforts. By comparing the outcome of *consequent dimension reduction* across different formats, these dummy variables allow analyzing at least in relative terms whether or not a format is more prone to fade-outs as a consequence of a strong satisficing strategy regarding the number of

[23] We also estimated a survival model in the event of a break-off to assess whether the presentation format affects the timing of the decision to break off the factorial survey. However, the break-offs in our sample were too few to estimate a survival model that accounts for all necessary interaction effects.

[24] For the estimation model we used the coding scheme presented in Table 1.

persons who follow the strategy (*person related scope*). To examine the intensity with which respondents apply this strategy, we additionally counted for each respondent the total number of coefficients with an effect size of zero (*intensity related scope*) and used it as dependent variable in an OLS-regression. A higher number of zero-sized coefficients suggest a higher intensity with which the strategy was performed.

Partial dimension reduction: When respondents consider all vignette dimensions in their subjective reference frame at the first vignette(s) of the factorial survey and start fading out a vignette dimension at the subsequent vignettes, then the effect size of the respective vignette dimension will be larger than zero in absolute terms, but attenuated compared to a situation where respondents optimize on all presented vignettes. As it is difficult to distinguish between small non-zero effects as a consequence of an optimizing strategy and attenuated non-zero effects as a consequence of a partial dimension reduction-strategy, we refrained from defining a cut off value that distinguishes between the weak satisficing and the optimizing strategy. Instead, we estimated for each vignette dimension an OLS-regression with respondents' coefficients for the respective vignette dimension as dependent variable.

Response inconsistency: For measuring response inconsistency, studies in the context of factorial surveys (cf. e.g., Sauer et al. 2011; Teti et al. 2016) focused on the unexplained variance of estimation models regarding respondents' (valid) answers to vignettes. We followed the operationalization applied by Sauer et al. (2011:94). On the basis of non-strong-satisficing respondents, we estimated a random intercept multilevel regression with respondents' answers to the 16 vignettes as dependent variable and the six vignette dimensions as predictors. In a subsequent step, we determined the squared residuals for each valid answer as an indicator for response inconsistency and used it as a dependent variable in a random intercept multilevel model for hierarchical data. Higher squared residuals reflect higher unexplained variances and, hence, higher response inconsistencies.

Total processing time: Para data were used in order to analyze the impact of the presentation format on the *total processing time* needed by respondents to judge the vignettes. Since one vignette was presented per page of the questionnaire, we were able to use the para data on the time spent per page to compute the total time a respondent needed to answer all 16 vignettes. To explain respondents' total processing time, this variable was used as dependent variable in an OLS-regression.[25]

In each of our models we included the dummy variables for each respective experimental setting. We also controlled for respondents' gender (0=male, 1=female), age groups (age groups 20 to 29 years, 30 to 39 years, 40 to 49 years, 50 to 59 years, 60 years and older, the first age category serves as reference for the other 0-1 dummy coded age groups), and for the highest level of education. The coding for education is based on a slightly modified ISCED-97

[25] Respondents of an internet survey can interrupt the survey whenever they want. Therefore, the total time needed to finish the interview can become quite long. In order to correct for such cases it was decided to exclude respondents who needed more than three times of the interquartile range to complete the interview. This criterion was fulfilled by 9 participants.

classification, which was developed for comparison purposes within the framework of the European Values Study 2008/2009 (cf. Dülmer et al. 2008). By selecting ISCED 1 and 2 as reference, we used four 0-1 coded dummy variables (ISCED 3, ISCED 5, ISCED 6, ISCED 7 and 8) for capturing the impact of education.[26]

Subsequent to the estimation of each model, we computed the average adjusted predictions for each experimental setting on the basis of the estimated model. For the calculation of average adjusted predictions, several predictions are calculated for each respondent. For each prediction and each respondent, the value of the variable of interest (e.g., the setting variable) is varied according to the number of its levels, while all other independent model variables are left as they are (Williams 2012). That is, in the case of a factor variable with three levels, we obtain three predictions for each respondent. The average(s) over each of these three predictions constitute the average adjusted prediction. In comparison to estimations of representative values of the other independent model variables, this procedure has the advantage that the resulting predictions are not only restricted to specific values but are based on empirical values (Williams 2012). In comparison to estimations of mean values of the other independent model variables, this procedure also has the advantage that the use of meaningless values can be ruled out (e.g., 0.42 for a dummy variable for female, if 42 percent of the respondents in the analysis sample are female) (Williams 2012). In a further step, we compared pairwise the resulting average adjusted probabilities and processing times, respectively, by using the Sidak-Test. The Sidak-Test is a test for pairwise multiple comparisons that is based on a t-statistic and accounts for the multiple comparisons in the significance level (Abdi 2007).

4. Empirical Results

4.1 Strong Satisficing Behavior – Descriptive Results

Respondents were randomly assigned to one of our three experimental settings. Column 2 of Table 4 shows the distribution of the 498 participants across the resulting samples. All three settings include exactly 166 respondents. According to Table 4, 74 respondents (column 3) evaluated no vignette at all and 17 persons (column 4) showed no variation in their answer behavior, so that in sum 91 participants were counted as refusals. A further 29 participants stopped answering the factorial survey between the first and the last vignette (break-off). Hence, by far the most common answer pattern for all refusals, including breaks-offs, was that of evaluating no vignette at all (74:120=61.7 percent). Setting 1, where non-supportive text vignettes were used, performs worst (in total 31.3 percent refusals including break-offs), while Setting 3, where tabular vignette formats were used, performs best (in total 17.5 percent refusals including break-offs).

[26] None of our respondents belonged to ISCED 4.

Table 4: Sample Sizes and Answer Behavior (Respondent Level)

Experimental Setting	Sample Size (n)	No Vignette Evaluated (Implicit Refusal, Don't Know)	Invalid Constant Answer Behavior	Refusals	Break-off	Total
		(1)	(2)	(1)+(2)	(3)	(1)+(2)+(3)
1	166	28	8	36 (21.7%)	16 (9.6%)	52 (31.3%)
2	166	28	5	33 (19.9%)	6 (3.6%)	39 (23.5%)
3	166	18	4	22 (13.3%)	7 (4.2%)	29 (17.5%)
Sum	498	74	17	91 (18.3%)	29 (5.8%)	120 (24.1%)

Notes: The reported row percentages are computed on the basis of a respective sample size;
Setting 1: text vignettes, no underlining;
Setting 2: text vignette, varying information underlined;
Setting 3: tabular vignettes, no separate columns f or varying information.

While the results in Table 4 refer to the respondent level, similar analyses have also been carried out for the vignette level. The gross sample size of vignettes that were assigned to the three experimental settings was 7,968 (cf. Table 5). However, the gross sample of vignettes was reduced by 1,456 vignettes (18.3 percent of the gross sample of vignettes, cf. Table 5) due to the refusals (n=91, cf. Table 4) and by 298 vignettes (3.7 percent of the gross sample of vignettes, cf. Table 5) due to the break-offs (n=29, cf. Table 4). Moreover, vignette nonresponse decreased the number of validly judged vignettes by 278 vignettes (3.5 percent). So, total vignette nonresponse decreased the percentage of validly judged vignettes in total by 2,032 vignettes (25.5 percent). Under both conditions (vignette nonresponse and total vignette nonresponse) and in line with H1b, experimental Setting 3 (tabular vignette format) performs best, whereas Setting 2 (supportive text format) performs worst with respect to vignette nonresponse. The percentage of unusable vignettes, however, is also quite low for Setting 2 (4.3 percent). With respect to the total vignette nonresponse, Setting 1 again performs worst (29.7 percent total loss of vignettes), which so far corroborates H1a. All in all, the rank order among the first three settings for total vignette nonresponse is Setting 3, 2, 1 and, hence, in line with our expectation that tabular vignette formats outperform text formats (H1b).

Table 5: Sample Sizes and Answer Behavior (Vignette Level)

	Gross Sample Size (n=498)			Net Sample Size after Excluding Refusals and Break-offs (n=378)			
	Gross Sample of Vignettes	Refusals	Break-off	'Don't Know'	No Answer	Vignette Non-response	Total Vignette Nonresponse
E.S.	(m)	(1)+(2)	(3)	(4)	(5)	(4)+(5)	(1)+(2)+(3)+ (4)+(5)
1	2,656	576 (21.7%)	146 (5.5%)	54	14	68 (2.6%)	790 (29.7%)

2	2,656	528 (19.9%)	69 (2.6%)	78	37	115 (4.3%)	712 (26.8%)
3	2,656	352 (13.3%)	83 (3.1%)	77	18	95 (3.6%)	530 (20.0%)
Sum	7,968	1,456 (18.3%)	298 (3.7%)	209	69	278 (3.5%)	2,032 (25.5%)

Notes: E.S.: Experimental Setting; Gross person sample refers to all respondents, net person sample to the gross person sample excluding refusals (n=91) and break-offs (n=29); the gross sample of vignettes is calculated by multiplying the group size of an experimental setting by the set size (16 vignettes per respondent); the reported row percentages are based on the gross sample of vignettes assigned to each experimental setting; in contrast to 'vignette nonresponse', 'total vignette nonresponse' includes participants who refused or broke off the factorial survey (last column of Table 4).

4.2 Strong Satisficing Behavior – Impact of Experimental Settings

Based on a logistic regression model in which we controlled for age, education, and gender of our respondents, for each separate setting we predicted the average adjusted probabilities (cf. Williams 2012) for refusals, break-offs, vignette nonresponse, and total vignette nonresponse. Analyses for refusals and break-offs are estimated at the respondent level, analyses for vignette nonresponse and total vignette nonresponse at the vignette level. Table 6 presents the results, rank ordered by performance, separately for each answer behavior.

Table 6: Average Adjusted Predictions (AAP) for Different Answer Pattern Rooted in Strong Satisficing

	Refusals		Break-off		Vignette Non-response		Total Vignette Nonresponse	
	i=91 (18.3 %)		i=29 (5.8 %)		i=278 (3.5 %)		i=2,032 (25.5 %)	
Rank	E.S.	AAP	E.S.	AAP	E.S.	AAP	E.S.	AAP
1	3	.133**	2	.033**	1	.025**	3	.201**
2	2	.198**	3	.046**	3	.035**	2	.266**
3	1	.216**	1	.097**	2	.045**	1	.297**
	Respondent level n=498		Respondent level n=498		Vignette level m=7,968		Vignette level m=7,968	

Notes: i: number of incidences; E.S.: Experimental Setting; ** p<0.01, * p<0.05; significance levels refer to differences from zero and are tested one-tailed.

The main source of nonresponses at the respondent level is refusals (18.3 percent). Break-offs and vignette nonresponse are comparably very rare (5.8 percent and 3.5 percent, respectively). Since refusals make up the larger part of the total vignette nonresponse and both categories (refusals and total nonresponse) show the highest differences between different settings, they will be the main focus of our interpretation. However, the final decision concerning the best performing setting has to be based solely on the total vignette nonresponse, which acts as a summary measure that identifies the vignettes which can be used for the substantial analyses.

In line with our expectation (H1b), Setting 3 (tabulate format) produces on average the lowest refusal rate. Here, the average for the age, education, and gender adjusted probability of a

refusal is 13.3 percent. This probability is 8.3 percent points lower than the average adjusted probability for the worst performing experimental setting, which in this case is Setting 1 (AAP=21.6 percent). For refusals and total vignette nonresponse, Setting 3 is consistently followed by Setting 2 (Rank 2) and Setting 1 (Rank 3). So far, for the main sources of nonresponse these results are in accordance with our expectation H1a (non-supportive text vignettes perform worst) and H1b (tabular vignettes outperform text vignettes). The results for break-offs and vignette nonresponse are somewhat different: text vignettes with underlined varying information (Setting 2) outperform tabular vignettes when it comes to break-offs, whereas even non-supportive text vignettes (Setting 1) show a lower average adjusted predicted probability for vignette nonresponse than tabular vignettes do. Compared to the main sources of nonresponse (refusals, total vignette nonresponse), for these two answer patterns (break off, vignette nonresponse) the differences in the average adjusted predicted probability are relatively small and for this reason of minor importance.

In a further step we used a Sidak multiple comparison test to see whether the average adjusted predictions differed significantly between the settings (cf. Table 7). In Setting 3 (ordinary tabular vignettes), for instance, the average adjusted probability for refusals is 8.3 percentage points smaller than in Setting 1 (non-supportive text vignettes). The difference, however, is not significant. The same also applies to the other two observed differences in the average adjusted predictions for refusals. With respect to the break-offs, only the difference between Setting 2 (supportive text vignette) and Setting 1 (non-supportive text vignette) turns out to be significant (AAP difference=-.064, $p<0.05$), corroborating the *non-supportive-text-format-satisficing hypothesis* (H1a).

Table 7: Results of Sidak-Test-Based Multiple Pairwise Comparisons of Average Adjusted Predictions (AAP) for Different Nonresponse Pattern of Different Experimental Settings

Refusals			Break-off		
vs.	1	2	vs.	1	2
2	-.018		2	-.064*	
3	-.083	-.065	3	-.051	0.01
Vignette Nonresponse			**Total Vignette Nonresponse**		
vs.	1	2	vs.	1	2
2	.020**		2	-.030*	
3	.010*	-.011	3	-.095**	-.065**

Notes: ** p<0.01, * p<0.05; directed hypotheses tested with a one-tailed test (only in cases where the expected pattern fitted the observed one); a negative sign indicates that the setting in a row outperforms the setting in a column

We now turn from the analyses on the respondent level to the vignette level. The results for vignette nonresponse clearly contradict our non-supportive-text-format-satisficing hypothesis (H1a) and our tabular vs. text format-satisficing hypothesis (H1b): Setting 1 (non-supportive text format-) turns out to significantly outperform Setting 2 and Setting 3, while

Setting 3 does not perform significantly better than Setting 2. However, regarding total vignette nonresponse the assumptions stated in H1a and H1b, respectively, are clearly confirmed: the tabulate format shows a significantly lower predicted average adjusted probability for total vignette nonresponse than Setting 1 (non-supportive text format) and Setting 2 (supportive text format), while at the same time the supportive text format (Setting 2) performs significantly better than the non-supportive text format (Setting 1). Hence, of all settings, Setting 3 (ordinary tabulate format) turns out to be the best performing format, significantly outperforming all other settings. A final comparison of the total vignette nonresponse with the refusals, which make up the most part of the total vignette nonresponse, shows that the observed patterns for both are similar. Hence, one of the reasons why for refusals even relatively high differences in the average adjusted predictions turned out to be insignificant is that the significance tests for the refusals are based on the number of respondents, whereas the significance tests for the total vignette nonresponse is based on the much higher number of vignettes.

4.3 Strong Satisficing Behavior – Impact of Education and Age on Strong Satisficing Behavior across Different Settings

To test our *presentation format-education hypothesis* (H3) and *our presentation format-age hypothesis* (H4) in our models for refusals, break-offs, vignette-nonresponse, and total vignette nonresponse we included interaction effects between the settings on the one hand, and both education and age on the other hand. In the following, we will restrict our presentation to the models where significant differences between different settings were found, as captured by interaction terms.

Table 8 shows the average adjusted predictions for total vignette nonresponse. The results are based on a logistic estimation in which we controlled for age, education, and gender and where interaction effects between age and setting, and between education and setting, were included. The AAP's of the interaction model differ only slightly from the AAP's of the logistic regression that only controlled for the main effects of age, education, and gender (c.f. Table 6, column 5).

Table 8: Average Adjusted Predictions (AAP) for Total Vignette Nonresponse Based on Interaction-Models

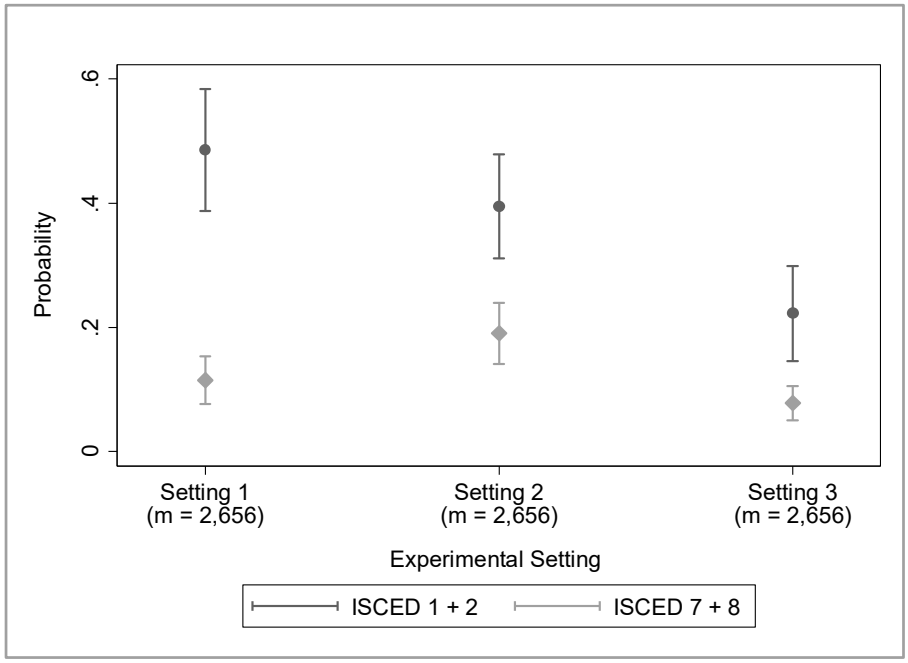
		Total Vignette Nonresponse (i=2,032)
Rank	E.S.	AAP
1	3	.203**
2	2	.264**
3	1	.298**
		Vignette level m=7,968

Notes: E.S.: Experimental Setting, ** p<0.01, * p<0.05; hypotheses tested with a one-tailed test.

For our three different settings Figure 1 shows the total vignette nonresponse probabilities for the group of less well educated persons (ISCED 1 and 2) compared to the group of highly educated persons (ISCED 7 and 8). In each of the Settings 1, 2, and 3 the nonresponse probability is significantly higher for the less well educated group compared to the highly educated group (cf. Sidak-Test, Table 9, column 2). Regarding our *presentation format-education hypothesis* (H3) we find that the differences between both educational groups decrease from Setting 1 to Setting 3. This result corroborates H3.

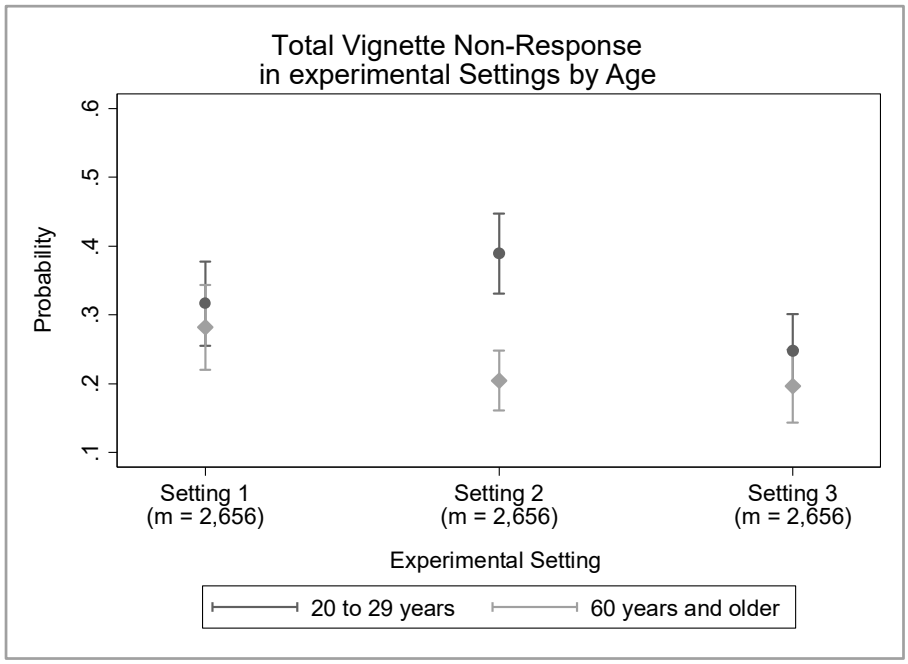
Figure 2 shows the total vignette nonresponse probabilities for the group of young persons (20 to 29 years) compared to the group of older persons (60 years and older). In Setting 1, 2, and 3 the older group shows lower total vignette nonresponse rates than the younger group. However, the Sidak-Test (Table 9, column 3) reveals that the difference in total vignette nonresponse rates between the two age groups is significant only for Setting 2. Hence, we do not find evidence for our presentation format-age hypothesis (H4).

Figure 1: The Impact of Education on Total Vignette Nonresponse across Different Settings



Notes: 37 persons with ISCED 1 + 2 and 103 persons with ISCED 7 + 8 were randomly distributed across the three settings; predictions based on logistic model (m = 7,968); dependent variable: 1=unit non-response; controlled for sex and age; confidence interval for level 95 %, adjusted for multiple comparisons.

Figure 2: The Impact of Age on Total Vignette Nonresponse across Different Settings



Notes: 91 persons aged 20 to 29 years and 87 persons aged 60 years and older were randomly distributed across the three settings; predictions based on logistic model (m = 7,968); dependent variable: 1 = unit non-response; controlled for sex and education; confidence interval for level 95 %, adjusted for multiple comparisons.

Table 9: Results of the Sidak-Test by Setting for Total Vignette Nonresponse by Education and Age

Experimental Setting	Education	Age
	ISCED 7 and 8 vs. ISCED 1 and 2	60 years and older vs. 20 to 29 years
1	-.408**	-.034
2	-.244**	-.185**
3	-.130**	-.051

Notes: ** p<0.01, * p<0.05; hypotheses for education tested with a one-tailed test.

4.4 Weak Satisficing Behavior and Processing Time – Impact of Experimental Settings

To increase the meaningfulness of our results (cf. Section 3.4), the analysis of weak satisficing behavior and processing time is based on the 322 non-strong-satisficing respondents who answered all 16 vignettes in a formally valid manner.

4.4.1 Consequent Dimension Reduction

Table 10 (cf. column 2) shows that 84 respondents (85.7 percent) in Setting 1 considered all vignette dimensions in their answers (Setting 2: 88.5 percent, Setting 3: 90.8 percent), while 14 respondents (14.3 percent) applied a consequent dimension reduction strategy by consequently fading-out at least one of the dimensions presented on the vignettes (Setting 2:

11.5 percent, Setting 3: 9.2 percent). In Setting 1, 38 (5.5 percent) of the 686 (=98 x 7) estimated coefficients were equal to zero, while 21 of the 728 (=104 x 7) coefficients in Setting 2 and 21 of the 840 (=120 x 7) coefficients in Setting 3, respectively were equal to zero. Hence, in line with H1a and H1b, Setting 1 performs worst regarding the person related scope of the consequent dimension reduction strategy as well as in terms of its intensity. Corroborating H1b, Setting 3 (tabular format) outperforms Setting 2 (supportive text format) in terms of the person related scope of the consequent dimension reduction strategy as well as in terms of the intensity of this strategy. Taken together, these results suggest that for the vast majority of respondents who answered all 16 vignettes, the subjective reference frame corresponded to the objective reference frame which we designed to examine the impact factors of ideas on just earnings.

Table 10: Consequent Dimension Reduction

		Person Related Scope n=322			Intensity Related Scope w=2254	
		A	B		C	D
Setting 1 (n=98)	n	84	14	w	648	38
	%	85.71	14.29	%	94.46	5.54
Setting 2 (n=104)	n	92	12	w	707	21
	%	88.46	11.54	%	97.12	2.88
Setting 3 (n=120)	n	109	11	w	819	21
	%	90.83	9.17	%	97.50	2.50
Total	n	285	37	w	2174	80
	%	88.51	11.49	%	96.45	3.55

Notes: n: number of non-strong-satisficing respondents; w: number of respondent specific coefficients; A: Respondents who did not consequently fade-out any of the vignette dimensions; B: Respondents who consequently faded-out at least one vignette dimension; C – Coefficients of size not equaling zero; D – Coefficients of size zero.

Table 11 shows the average adjusted probability for applying a consequent dimension reduction strategy (cf. Table 11, person related scope) and the average adjusted number of fade-outs per respondent (cf. Table 11, intensity). The results are based, respectively, on a logistic regression and an OLS-regression, in which we took age, education, and gender into account. As expected in H1b, Setting 3 (tabular format) performs best regarding both issues of the consequent dimension reduction strategy; Setting 2 performs better than Setting 1, corroborating H1a. The average adjusted probability to apply this weak form of satisficing amounts to 13.9 percent in Setting 1 and 9.2 percent in Setting 3; at the same time respondents in Setting 1 on average faded out more dimensions (.377) than respondents in Setting 3 (.183). However, the difference is not significant in statistical terms, as the results of the Sidak-Test show in Table 12. None of the examined presentation formats is particularly more prone to consequent dimension reduction.

4.4.2 Partial Dimension Reduction

To account for respondents whose subjective reference frame was not constant across all 16 vignettes, we compared the average coefficients of the vignette dimensions across the settings. Table 11 shows for each vignette dimension its average adjusted prediction in the respective setting. These average adjusted predictions reflect the setting specific adjusted isolated effects of a dimension on respondents' judgments. For instance, in Setting 1 respondents assigned in accordance with existing stereotypes on average a lower salary (by €16.99) to fictitious vignette persons when these were females rather than males, while respondents in Setting 3 contrary to existing stereotypes assigned to fictitious female vignette persons on average a €41.57 higher salary than to fictitious male vignette persons (this value is very close to €43.50, i.e., the average of Setting 2).[27] However, the vignette dimension gender was in both settings not significantly (α=0.05) different from zero, meaning that male and female fictitious vignette persons were assigned the same salary.

According to H1a and H1b we would expect that for each vignette dimension the average adjusted predictions would be largest, in absolute terms, in Setting 3 and smallest in Setting 1. However, the rank orders of the average adjusted predictions (in absolute terms) presented in Table 11 do not systematically match our expectation; as is the case, for example, with the vignette dimension effort at work (cf. Table 11, vignette dimension effort at work). Even more, the results of the Sidak-Test (cf. Table 12) show that only one comparison among the 21 comparisons for partial dimension reduction yielded in a significant difference between the effect sizes. Respondents in Setting 2 assigned on average €122.90 less to a fictitious vignette person than respondents in Setting 1, if the pay inequality in the fictitious region was described as being high. This single significant difference at least partially undermines our expectation (H1a).

4.4.3 Response Inconsistency

In order to retrieve the squared residuals of this estimation model we estimated a random-intercept multilevel model with respondents' answers to the 16 vignettes as dependent variable and the vignette dimensions as independent variables of the model. The squared residuals serve as indicator for response inconsistency (higher positive values mean higher

[27] Conscious or unconscious stereotypes in our societies, *ceteris paribus,* privilege males over females regarding their income, a result that unconsciously also applies to highly educated females (Moss-Racusin et al. 2012, cf. also Ross 2014:XV). Even though factorials surveys are less prone to *social desirability bias* than directly asked single-item questions (Auspurg et al. 2015:143-144, cf. also Auspurg and Hinz 2015:4, 11), Beyer and Liebe (2015) could show that factorial surveys are also susceptible to social desirability bias in case of sensitive topics where perceived public norms of "political correctness" exist (e.g., discrimination of minorities, income discrimination of females). Auspurg and Hinz (2015) expect that supported presentation formats are more affected by social desirability bias than unsupported formats, since for table formats the respondents' attention is "more clearly bound to the manipulation of the researcher" (p. 71) whereas for supported text formats "respondents can more easily identify the highlighted dimensions" (p. 72). In our study, the observed differences for the dimension 'gender' tend to support the direction of a social desirability bias among the supportive formats as assumed by Auspurg and Hinz (2015) (cf. Table 11). If tested one-tailed, however, only the difference between the two text formats (Setting 2 vs. Setting 1) would become significant (43.50-(-16.99)=60.49, p=.041).

inconsistency). They were used as dependent variable in a separate random intercept multilevel model in which we examined the impact of the presentation format on response inconsistency, while controlling for respondents' education, age, processing time, and the vignette's position. The average adjusted predictions presented in Table 11 (cf. response inconsistency) were calculated on the basis of this model. Against our expectation (H1b), Setting 3 (tabular vignette) performs worse than the settings in which text-based vignettes were used. In line with H1a, supportive text vignettes had a lower response inconsistency than non-supportive text vignettes.[28] However, we do not find any significant differences across the three experimental settings using the Sidak-Test that accounts for multiple comparisons (cf. Table 12).

4.4.4 Processing Time

The last column of Table 11 gives an overview of the average adjusted processing time. In accordance with our tabular vs. text format-time hypothesis (H2b) – Setting 3 performs best: respondents needed 334.5 seconds on average to judge all 16 vignettes, 48.6 seconds less than in Setting 2, which performed worst. Against our non-supportive text format-time hypothesis (H2a), the supportive text vignette format (Setting 2) performs with an adjusted processing time of 383.1 seconds worse than the non-supportive text format hypothesis (Setting 1). However, the difference between both text formats is negligible (6 seconds).[29]

According to H2a, the processing time for a supportive presentation format should be shorter than the processing time for a non-supportive presentation format. Empirically, this expectation does not hold for the comparison between Setting 1 (non-supportive text vignettes) and Setting 2 (supportive text vignettes, AAP difference=5.99, $p>0.05$, cf. Table 12). Hypothesis H2b, however, is at least partially confirmed empirically; the processing time for tabular vignette format (Setting 3) is significantly lower than the processing time for the supportive text format (Setting 2) but not significantly lower than the non-supportive text format (Setting 3).

[28] The response inconsistency in Setting 3 is not significantly lower ($\alpha=0.05$) than in Setting 1.

[29] We replicated the analysis regarding the processing time on the basis of 364 respondents who answered eight or more vignettes (whereas 14 outliers were disregarded from the analysis) and on the basis of all 392 respondents who answered at least one vignette (whereas 14 outliers were disregarded from the analysis). These changes did not affect the pattern in the processing times reported in Table 11.

Table 11: Average Adjusted Predictions (AAP) for Different Answer Pattern Rooted in Weak Satisficing

	Consequent Dimension Reduction				Partial Dimension Reduction											
	Person Related Scope n=37 (11.5 %)		Intensity Related Scope i=80 (.25)		Vignette Dimension Gender		Vignette Dimension Parenthood		Vignette Dimension Job Experience		Vignette Dimension Effort at Work		Vignette Dimension Pay Inequality		Vignette Dimension Low Average Salary		Vignette Dimension High Average Salary	
Rank	E.S.	AAP	E.S.	AAP	E.S.	AAP	E.S.	AAP	E.S.	AAP	E.S.	AAP	E.S.	AAP	E.S.	AAP	E.S.	AAP
1	3	.092**	3	.183**	2	43.50	3	229.90**	3	128.76**	1	850.68**	1	139.13**	1	-65.02**	1	122.72**
2	2	.118**	2	.203**	3	41.57	1	171.59**	2	98.69**	2	774.18**	3	73.57**	2	-37.23*	3	119.10**
3	1	.139**	1	.377**	1	-16.99	2	127.78**	1	81.56**	3	728.60**	2	16.23	3	-31.89	2	105.17**
	Respondent level n=322				Respondent level n=322													

Notes: E.S.: Experimental Setting; i: number of incidences; ** p<0.01, * p<0.05; significance levels refer to differences from zero and are tested one-tailed, except for partial dimension direction where no directed hypotheses (positive or negative relationship) were formulated.

Table 11 continued

	Response Inconsistency		Processing Time for all 16 Vignettes (in Seconds)	
Rank	E.S.	AAP	E.S.	AAP
1	3	453163**	3	334.5**
2	1	417526**	1	377.1**
3	2	347868**	2	383.1**
	Respondent level n=322 [a)]			

Notes: E.S.: Experimental Setting; ** p<0.01, * p<0.05; significance levels refer to differences from zero and are tested one-tailed;
[a)] With respect to the processing time 9 outliers were identified and therefore excluded from analyses (cf. FN 24).

Table 12: Results of Sidak-Test-Based Multiple Pairwise Comparisons of Vignette Dimensions' Average Effect Sizes in Different Experimental Settings

Consequent Dimension Reduction (Person Related Scope)		
vs.	1	2
2	-.021	
3	-.047	-.026

Consequent Dimension Reduction (Intensity Related Scope)		
vs.	1	2
2	-.175	
3	-.194	-.020

Partial Dimension Reduction (Gender)		
vs.	1	2
2	60.49	
3	58.57	-1.92

Partial Dimension Reduction (Parenthood)		
vs.	1	2
2	-43.81	
3	58.31	102.11

Partial Dimension Reduction (Job Experience)		
vs.	1	2
2	17.13	
3	47.20	30.07

Partial Dimension Reduction (Effort at Work)		
vs.	1	2
2	-76.50	
3	-122.08	-45.58

Partial Dimension Reduction (Pay Inequality)		
vs.	1	2
2	-122.90**	57.34
3	-65.56	

Partial Dimension Reduction (Low Average Salary)		
vs.	1	2
2	27.79	5.33
3	33.12	

Partial Dimension Reduction (High Average Salary)		
vs.	1	2
2	-17.55	-13.94
3	-3.61	

Response Inconsistency		
vs.	1	2
2	-69,658	
3	35,637	105,295

Processing Time for Judging all 16 Vignettes		
vs.	1	2
2	5.99	
3	-42.61	-48.60*

Notes: ** p<0.01, * p<0.05; hypotheses tested one-tailed, expect for partial dimension reduction where no directed hypotheses were formulated; a negative sign indicates that respondents of the setting in a row assigned on average a smaller amount to the fictitious person than respondents of the setting in a column.

5. Discussion and Conclusions

Former methodological studies strengthened the confidence in the factorial survey by examining issues of design complexity in terms of vignette dimension number, vignette number, and order effects. In this study we addressed another design related issue by examining the impact of the presentation format of vignettes (tabular vs. text) on the response behavior in factorial surveys. For our internet-based study, we recruited 498 individuals living in Germany according to a quota plan with crossed quota on age, sex, and education. Our respondents were randomly assigned to one of three experimental settings. In each of the experimental settings, respondents were presented with 16 vignettes. The presentation format varied across the three experimental settings (non-supportive and supportive text vignette format, tabular vignette format). Respondents' answers to the

vignette judgment task with open answer format were analyzed regarding different types of strong satisficing strategies (refusal, break-off, vignette nonresponse) as well as different types of weak satisficing strategies (consequent dimension reduction, partial dimension reduction, response inconsistency).

For the respondent level, refusal rates varied in our factorial survey between 13.3 percent in Setting 3 (tabular vignette format) and 21.7 percent in Setting 1 (non-supportive text vignette format). We used an open answer format and did not force our respondents to answer each judgment task by explicitly offering a 'don't know' option. The open answer format has been shown to result in higher missing value rates than closed answer formats (Schuman and Presser 1979, 1996; Urša et al. 2003) and offering a 'don't know' option, which make it easier for respondents to strongly satisfice by simply answering 'don't know' (Krosnick et al. 2002; Krosnick et al. 1997; Krosnick 1991). Even though refusal rates might seem to be comparably high in their absolute level, they still allow examining them in relative terms by comparing the presentation formats regarding the tendency to refuse participation in the factorial survey. Although the refusal rate was 8.3 percent points lower for tabular vignettes than for non-supportive text vignettes, none of the presentation formats significantly increased the probability of refusing to participate in the factorial survey. Breaking off the factorial survey occurred systematically more often with non-supportive text format vignettes (Setting 1) than with supportive text format vignettes (Setting 2), whereas break-off rates in Setting 3 (tabular format) were not significantly different from break-off rates in the text format vignettes (Setting 1 and 2). For future studies it might be good to have a higher sample size on the respondent level, which would allow also to detect comparably small effects. Apart from that, although different nonresponse patterns shed light on different sources of nonresponse, using the total vignette nonresponse as a summary measure for all vignettes that are not available for substantial analyses is more important for the researcher. By using total vignette nonresponse as the central criterion for selecting the best performing presentation format, our results clearly show that Setting 3 (tabular presentation format) performed best regarding strong satisficing.

Beside these findings, which on average hold for all respondents, we examined whether our outcomes differ with regard to a respondent's education and age, respectively. Concerning total vignette nonresponse, we found that individuals with the highest level of education perform significantly better than persons with the lowest level of education. This applies to all presentation formats. The smallest difference between these two groups was found in Setting 3, which indicates that the lowest educated could profit most from using a tabular format. Regarding age we found that respondents aged 60 and older did show lower instead of higher total vignette nonresponse rates than respondents aged between 20 and 29 years. The difference, however, only was significant for the supportive text format. The former result might be explained by higher perseverance of older respondents, which compensates for their slower information intake. Taken together, systematic

information loss and processing time was minimized by Setting 3 (table presentation format).

While table formats performed best regarding strong satisficing, our study reveals that response behavior grounded in weak satisficing was not triggered by any of the presentation formats. That is, we found no statistical evidence for differences across the three different presentation formats concerning consequent dimension reduction, partial dimension reduction, and response inconsistency. In contrast to answer behaviors rooted in strong satisficing, answer behaviors rooted in weak satisficing are less easy to detect for researchers. Respecting the processing time we found that Setting 3 (tabular presentation format) performed significantly better than supportive text format vignettes (Settings 2), while there were no systematic difference between Setting 3 and Setting 1 (non-supportive text format). Furthermore, analysis of consequent dimension reduction showed that in the process of judgment formation the vast majority of respondents, who rated all 16 vignettes, did not fade out in their subjective reference frame any of the vignette dimensions of the objective reference frame that we designed for this study. This result might help to explain why respondents' answers in factorial surveys can be expected to be ascertained with higher reliability and higher validity than with single survey items in typical survey research. Finally, in line with earlier studies, our results point in the direction that for sensitive topics where perceived public norms of "political correctness" exist, supportive formats are more prone to social desirability bias than non-supportive formats. In order to reduce such potential biases, it might be recommended to present only one vignette per respondent in a non-supportive format.

References

AAPOR (American Association for Public Opinion Research). 2015. Standard Definitions –Final Dispositions of Case Codes and Outcome Rates for Surveys.

Abdi, Hervé. 2007. "The Bonferonni and Šidák Corrections for Multiple Comparisons." Pp. 103-107 in *Encyclopedia of Measurement and Statistics*. Edited by Neil J. Salkind. Thousand Oaks, CA: Sage.

Alexander, Cheryl S. and Henry Jay Becker. 1978. The Use of Vignettes in Survey Research. *Public Opinion Quarterly* 42(1):93-104.

Atzmüller, Christiane and Peter M. Steiner. 2010. Experimental Vignette Studies in Survey Research. *Methodology – European Journal of Research Methods for the Behavioral and Social Sciences* 6(3):128-138.

Auspurg, Katrin and Thomas Hinz. 2015. *Factorial Survey Experiments.* Series: Quantitative Applications in the Social Sciences 175. Thousand Oaks, CA: Sage.

Auspurg, Katrin, Thomas Hinz, and Stefan Liebig. 2009. Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey. *Methoden – Daten – Analysen* 3(1):59-96.

Auspurg, Katrin, Thomas Hinz, Stefan Liebig, and Carsten Sauer. 2015. "The Factorial Survey as a Method for Measuring Sensive Issues." Pp. 137-149 in *Improving Survey Methods. Lessons from Recent Research*. Edited by Uwe Engels, Ben Jann, Peter Lynn, Annette Scherpenzeel, and Patrick Sturgis. New York: Routledge.

Auspurg, Katrin and Annette Jäckle. 2017. First Equals Most Important? Order Effects in Vignette-Based Measurement. *Sociological Methods & Research* 46(3):490-539.

Beck, Michael and Karl-Dieter Opp. 2001. Der faktorielle Survey und die Messung von Normen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 53(2):283-306.

Beyer, Heiko and Ulf Liebe. 2015. Three Experimental Approaches to Measure the Social Context Dependence of Prejudice Communication and Discriminatroy Behavior. *Social Science Research* 49:343-355.

Bogart, Leo. 1967. No Opinion, Don't Know, and Maybe No Answer. *The Public Opinion Quarterly,* 31(3):331-345.

Bose, Christine E. and Peter. H. Rossi. 1983. Gender and Jobs: Prestige Standings of Occupations as Affected by Gender. *American Sociological Review* 48(3):316-330.

Converse, Philip. 1964. "The Nature of Belief Systems in Mass Publics." Pp. 206-261 in *Ideology and Discontent.* Edited by David Apter. New York: Free Press.

Dillman, Don A. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley-Interscience.

Dülmer, Hermann. 2007. Experimental Plans in Factorial Surveys. *Sociological Methods & Research* 35(3):382-409.

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

Dülmer, Hermann, Wolfgang Jagodzinski, and Pascal Siegers. 2008. Classification Scheme for the Modified Education Variable of the EVS 1999/2000, mimeo.

European Social Survey (ESS). 2016. United Kingdom. Data and Documents (https://www.europeansocialsurvey.org/data/country.html?c=united_kingdom).

Flesch, Rudolph. 1948. A New Readability Yardstick. *Journal of Applied Psychology* 32(3):221-233.

Gunst, Richard F. and Robert L. Mason. 1993. *How to Construct Fractional Factorial Experiments.* The Basic References in Quality Control: Statistical Techniques 14. Milwaukee, WI: ASQC Quality Press.

Hartshorne, Joshua K. and Laura T. Germine. 2015. When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span. *Psychological Science* 26(4):433-443.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015: Validating Vignette and Conjoint Survey Experiments against Real-World Behavior. *Proceedings of the National Academy of Sciences of the United States of America* 112(8):2395-2400.

Hermkens, Piet L. J. and Frank A. Boerman. 1989. Consensus With Respect to the Fairness of Incomes: Differences Between Social Groups. *Social Justice Research* 3(3):201-215.

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

Knäuper, Baerbel, Robert F. Belli, Daniel H. Hill, and A. Regula Herzog. 1997. Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality. *Journal of Official Statistics* 13(2):181-199.

Krosnick, John A. 1991. Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology* 5(3):213-236.

Krosnick, John A., Allyson L. Holbrook, Matthew K. Berent, Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp., Robert Cameron Mitchell, Stanley Presser, Paul A. Ruud, V. Kerry Smith, Wendy R. Moody, Melanie C. Green, Michael Conaway. 2002. The Impact of "No Opinion" Response Options on Data Quality. *Public Opinion Quarterly* 66(3):371-403.

Krosnick, John A. and Leandre R. Fabrigar. 1997. "Designing Rating Scales for Effective Measurement in Surveys." Pp. 141-164 in *Survey Measurement and Process Quality.* Edited by In Lars E. Lyberg, Paul P. Biemer, Martin Collins, Edith De Leeuw, Cathryn. Dippo, Norbert Schwarz, and Dennis Trewin.

Kuhfeld, Warren F. 2010. "Experimental Design: Efficiency, Coding, and Choice Design." Pp. 53-241 in *Marketing Research Methods in SAS. Experimental Design, Choice, Conjoint, and Graphical Techniques*. Edited by Warren F. Kuhfeld, Retrieved October 5, 2013 (http://support.sas.com/resources/papers/tnote/tnote_marketresearch.html).

Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt. 1994. Efficient Experimental Design with Marketing Research Applications. *Journal of Marketing Research* 31:545-557. A revised version appeared in SAS Technical Papers – Marketing Research, 243-264. Retrieved August 28, 2014 (http://support.sas.com/techsup/technote/mr2010d.pdf).

Liebe, Ulf, Ismaïl M. Moumouni, Christine Bigler, Chantal Ingabire, and Sabin Bieri. 2017. Using Factorial Survey Experiments to Measure Attitudes, Social Norms, and Fainess Concerns in Developing Countries. *Sociological Methods & Research*, online first:1-32.

Maehler, Déborah B., Natascha Massing, Susanne Helmschrott, Beate Rammstedt, Ursula M. Staudinger, and Chistof Wolf. 2013. "Grundlegende Kompetenzen in verschiedenen Bevölkerungsgruppen." Pp. 77-126 in *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012.* Edited by Beate Rammstedt. Münster: Waxmann.

Mason, Robert and John E. Carlson. 1994. Contrast Effects and Subtraction in Part-Whole Questions. *Public Opinion Quarterly 58*(4):569-578.

Miller, George A. 1994. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review* 101(2):343-352.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. Science Faculty's Subtle Gender Biases Favor Male Students. *Proceedings of the National Academy of Sciences of the United States of America* 109 (41): 16474–16479.

Norman, Donald A. 1972. Memory, Knowledge, and the Answering of Questions. Paper Presented at the *Loyola Symposium on Cognitive Psychology*, Chicago, 1972.

PIAAC (Programme for the International Assessment of Adult Competencies). 2016. *Programme for the International Assessment of Adult Competencies*. Retrieved April 9, 2017 (http://www.oecd.org/skills/piaac).

Porst, Rolf. 2008. *Fragebogen: Ein Arbeitsbuch* (Vol. 1). Wiesbaden: VS-Verlag.

Ross, Howard J. 2014: *Everyday Bias: Identifying and Navigating Unconscious Judgments in Our Daily Lives.* Lanham Maryland: Rowman & Littlefield.

Rossi, Peter H. and Andy B. Anderson. 1982. “The Factorial Survey Approach: An Introduction.” Pp. 15-67 in *Measuring Social Judgments: The Factorial Survey Approach.* Edited by Peter H. Rossi and Steven L. Nock. Thousand Oaks, CA: Sage Publications.

Rubin, Donald B. 1976. Inference and Missing Data. *Biometrika* 63(3):581-592.

Saris, Willem. E. and Irmtraud N. Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionaires for Survey Research*. Hoboken, NJ: Wiley-Interscience.

Sauer, Carsten, Katrin Auspurg, Thomas Hinz, and Stefan Liebig. 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.

Schuman, Howard and Stanley Presser. 1979. The Open and Closed Question. *American Sociological Review* 44(5):692-712.

Schuman, Howard and Stanley Presser. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks: Sage.

Shamon, Hawal. 2014. *Der Einfluss des sozialen Kontexts auf Gerechtigkeitsurteile über das Einkommen – eine Untersuchung des kausalen Effekts auf Grundlage von internetbasierten experimentellen Umfragedesigns*. Druckschrift, Universität zu Köln, Köln.

Shamon, Hawal and Hermann Dülmer. 2014. Raising the Question on 'Who Should Get What?' Again: On the Importance of Ideal and Existential Standards. *Social Justice Research* 27(3):340-368.

Shoemaker, Pamela J., Martin Eichholz, and Elizabeth A. Skewes. 2002. Item Nonresponse: Distinguishing between Don't Know and Refuse. *International Journal of Public Opinion Research* 14(2):193-201.

Sniderman, Paul M. and Douglas B. Grob. 1996. Innovations in Experimental Design in Attitudes Surveys. *Annual Review of Sociology* 22:377-399.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking about Answers*. San Francisco: Jossey-Bass.

Teti, Andrea, Christiane Gross, Nina Knoll, and Stefan Blüher. 2016. Feasibility of the Factorial Survey Method in Aging Research: Consistency Effects Among Older Respondents. *Research on Aging 38*(7):715-741.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Urša, Reja, Katja Lozar Manfreda, Valentina Hlebec, and Vasja Vehovar. 2003. Open-ended vs. Close-ended Questions in Web Questionnaires. Retrieved September 12, 2018 (http://mrvar.fdv.uni-lj.si/pub/mz/mz19/reja.pdf).

Van der Veld, Williams M. and Willem E. Saris. 2003. “Seperation of Error, Method Effects, Instability and Attitude Strength.” Pp. 37-63 in *The Issue of Belief: Essays in the Intersection of Nonattitudes and Attitude Change.* Edited by Willem E. Saris and Paul M. Sniderman. Princeton: University Press.

Williams, Richard. 2012. Using the Margins Command to Estimate and Interpret Adjusted Predictions and Marginal Effects. *The Stata Journal* 12(2):308-331.