Survey Research Methods (2019)
Vol. 13, No. 1, pp. 103-121
doi:10.18148/srm/2019.v13i1.7243

© *European Survey Research Association*
ISSN 1864-3361
http://www.surveymethods.org

Hiding Sensitive Topics by Design? An Experiment on the Reduction of Social Desirability Bias in Factorial Surveys

Sandra Walzenbach
University of Konstanz, Germany
and Institute for Social and Economic Research (ISER)
University of Essex, UK

Factorial survey designs have gained increasing popularity within the social sciences. Compared to single-item questions, the method allows the researcher to model more realistic, multidimensional decision scenarios. Furthermore, it has been argued that assessing sensitive dimensions in factorial surveys can help to overcome social desirability bias.

One rarely used implementation mode is the between subject design, in which the sensitive dimension varies only between respondents. This method is assumed to attract less attention than a design based on the usual within subject implementation, where respondents see variations on the sensitive dimension among their vignettes. In order to empirically evaluate the between design and its potential to reduce social desirability bias, we conducted an experiment within a general population online survey. Using a split-half design, the sensitive dimension in the vignette texts was either varied within or between subjects. More precisely, the factorial survey module under study assessed respondents' judgements on just fees for early childcare. Among other dimensions, the vignette texts included the child's religious denomination (Christian, Muslim, none) as one possible attribute on which discrimination can be based. The split-half approach allows us to compare the widely used within subject design to the alternative between approach. Furthermore, data on respondent characteristics is used to obtain insights about differential design effects for different education groups (differential social desirability bias) and respondents from different religious backgrounds (ingroup favouritism). While results concerning a differential social desirability bias were inconclusive, we found evidence for ingroup favouritism from respondents without a religious denomination in the between condition. In general, our findings suggest that the between subject design is a suitable method for reducing social desirability bias in factorial surveys.

Keywords: factorial survey; vignette study; sensitive topics; social desirability bias; justice norms

1 Introduction

Although multifactorial surveys have become increasingly popular in the social sciences over recent years and the method has been claimed to be a particularly suitable approach to survey sensitive topics, surprisingly little is known about the design features that are most favourable in regard to reducing social desirability bias and enhancing data quality in vignette designs. It has been argued – mostly theoretically – that asking questions indirectly in the form of fictitious scenarios is a particularly unobtrusive way to cover sensitive topics in surveys. It thus seems promising to refine the method in order to further the pursuit of more truthful respondent behaviour.

One approach that is discussed in this context is the so-called between subject design. While respondents are usually given multiple vignettes, in which all dimensions of interest vary from scenario to scenario, it has been claimed that keeping the sensitive dimension constant within one respondent's vignette set should reduce social desirability bias. According to theoretical arguments, the latter approach attracts less attention and gives less of an incentive to distort socially undesirable answers than the usual within subject implementation. Between designs, however, come with a trade-off: the potential reduction in social desirability bias goes hand in hand with larger standard errors. Which design should researchers choose under these circumstances? Should they accept a loss in statistical power in order to reduce social desirability bias? Can between subject design help to reveal correlations that would otherwise be disguised by distorted

Contact information: Sandra Walzenbach, Universität Konstanz, Arbeitsbereich für empirische Sozialforschung, Postfach 40, Universitätsstr. 10, D-78457 Konstanz (E-mail: sandra.walzenbach@essex.ac.uk)

answers?

This paper contributes to the scientific discussion about reducing social desirability bias in factorial surveys by presenting results from an experimental test of different design features, which provides valuable empirical evidence for a debate that at the moment is mainly theoretical.

The experiment was implemented in a general population survey, for which data collection took place in the southwest of Germany in 2014 (Arbeitsgruppe Hinz, 2014). With regard to content, the vignette module was designed to investigate contributive justice norms in the context of just fees for early childcare. As a sensitive dimension, religious denomination (with the categories Christian, Muslim and none) was included in the descriptions of parents looking for a public childcare institution. Using a split-half design, respondents were randomly attributed to one of two experimental conditions: They either answered five vignettes, in which the sensitive dimension was randomly varied (within subject design) or held constant throughout the vignette set (between subject design). Due to the experimental inclusion of important context variables in the vignette texts, we can estimate an unbiased religion effect, net of other characteristics that would be correlated with religious affiliation in the real world. Based on the assumption that more discriminatory evaluations constitute more truthful answers, we examine which design features are preferable for eliciting socially undesirable responses.

This paper is structured as follows. To begin with, we provide a short general introduction about social desirability bias and multifactorial survey designs. We then give an overview of the little research that has been conducted on the topic before and present our own study design, as well as its results. Most crucially, we examine empirically which design is better suited and thus superior when it comes to reducing social desirability bias. Additional analyses show that the observed differences between the experimental conditions are caused by an underlying mechanism of ingroup favouritism that we observe for respondents without any religious denomination. Interestingly, we only find this effect in the between subject condition, while it would have remained undiscovered in the within subject design. Apart from these promising findings on the potential of between subject designs, we also present results on a theoretically assumed differential social desirability bias dependent on respondent's educational background. We conclude with a discussion, in which the reliability of the presented results and directions for future research are addressed.

2 Theoretical Considerations

2.1 Social Desirability Bias

Social desirability bias is one of the potential threats to data quality in surveys. Especially when sensitive questions (e.g. about sexual practices or delinquent behaviour) are assessed, respondents might conceal their true opinion in order to maintain a positive self-perception or to obtain social approval in an interview situation. Particularly if they fear that honesty might have negative consequences, respondents will tend to distort their answers (Esser, 1986; Tourangeau, Rips, & Rasinski, 2010; Tourangeau & Smith, 1996).

To avoid social desirability bias, standard recommendations from textbooks on survey methodology include the utilisation of neutral question wordings (see e.g. Lensvelt-Mulders, 2008, 468 et seq) and, if possible, the implementation of self-administered questionnaires instead of personal or telephone interviews (for a meta-analysis of 52 studies published between 1961 and 1990 see De Leeuw, 1992; for tabular overviews Tourangeau, Conrad, & Couper, 2013, Chapter 7; Tourangeau & Smith, 1996, 278 et seq). In addition, researchers are sometimes reminded that perceivable interviewer characteristics, such as ethnicity or gender, as well as the sponsor of the survey, provide cues about socially acceptable answers and might foster interviewer or sponsorship bias (for an overview of interviewer effects see Groves et al., 2009, Chapter 9; interesting case studies include Houle et al., 2016; Liu & Stainback, 2013; Schnell & Kreuter, 2005; Turner, Sturgis, Martin, & Skinner, 2015; for a study on sponsorship see Corstange, 2014; Tourangeau, Presser, & Sun, 2014).

Another popular but somewhat controversial approach to reducing social desirability are randomised response techniques (RRTs), as well as related methods that are sometimes referred to as non-randomised response techniques (for an overview of existing specifications see Fox, 2015). These questions techniques were developed to enhance privacy protection and share the common feature that they deliberately add noise to the collected data (e.g. by means of a random device like dice or coins). As a consequence, an individual answer is only probabilistically linked to the sensitive behaviour and can no longer be revealing.

However, some methodological research on the topic has identified severe problems (like low statistical efficiency, high variance in results despite equal experimental conditions, misinterpretation of the rather complex procedure or intentional misreporting) related to these methods (Edgell, Himmelfarb, & Duchan, 1982; Höglinger & Diekmann, 2017; Holbrook & Krosnick, 2010; Umesh & Peterson, 1991; Walzenbach & Hinz, 2014). Moreover, the procedure comes with enlarged standard errors and is designed to estimate prevalence rates on the aggregate level, although researchers ideally need data on the individual level to validate the answers and quantify the bias (Höglinger & Jann, 2016).

More recently, factorial surveys have also been discussed as an appropriate alternative tool for reducing social desirability bias in surveys (Auspurg & Hinz, 2015, Chapter 2; Mutz, 2011). Going back to Peter Rossi, who originally used factorial surveys to study normative judgements (Rossi & Andersen, 1982), the method has become increasingly popular in the social sciences over the last years. It meanwhile has been applied to a wide range of areas, including punishment preferences for deviant behavior, measurements of social status, normative perceptions of fair earnings, and attitudes towards immigration (e.g. Atzmüller & Steiner, 2010; Auspurg, Hinz, & Sauer, 2017; Hainmueller, Hangartner, & Yamamoto, 2015; Jasso & Webster, 1999; for an overview see Wallander, 2009).

Auspurg, Hinz, Sauer, and Liebig (2015) have provided promising initial empirical evidence on the implementation of sensitive topics in factorial survey experiments: Compared to a direct question format, the vignette module yielded less socially desirable answers concerning a just gender wage gap. Similar results have previously been found in articles by Armacost, Hosseini, Morris, and Rehbein (1991) and Burstin, Doughtie, and Raphaeli (1980), who also compared direct questioning formats to scenario questions. However, the authors varied none or only one dimension of their vignettes, meaning that their designs are only roughly related to factorial surveys.

2.2 Social Desirability Bias in Factorial Surveys

In factorial surveys, multiple dimensions are experimentally varied in a scenario description. Respondents are asked to perform the quite complex task of evaluating all of them simultaneously and translating their opinion into one judgement about the vignette as a whole (for an introduction see Auspurg & Hinz, 2015; Wallander, 2009). Given a thorough design plan, in which the vignette dimensions and their interactions are uncorrelated (orthogonal) and occur in equal frequency (are balanced), statistical analysis produces unbiased and independent estimates for all varied dimensions (Kuhfeld, 1997).

As an example, the factorial survey we are going to analyse in this article deals with fair contributions to early childcare.[1]

Respondents were asked to evaluate on an 11-point Likert scale how just a certain monthly childcare fee would be. Every vignette contained a description of a couple that used the services of a public institution to look after their first child, who was stated to be two years old. The couples differed in their employment situation, their family background and their social integration in the local community (see Figure 1 for an example).

The vignette dimension of primary concern for our research question is the family's religious denomination (Christian, Muslim, none). Although equal treatment, irrespective of religious beliefs, is assured by the German constitution (Basic Law §3 cl. 3)[2], adherence to Islam is one of the attributes, on which discrimination is most typically based in Germany and other European countries. We therefore expect that social desirability bias should play a particularly important role when respondents evaluate this sensitive dimension.

Compared to a fictitious single-item question about the effect that religious denomination should have on childcare fees, the factorial survey provides several advantages.

First of all, it provides contextual information. This is crucial because respondents are likely to use statistical discrimination as a strategy to compensate for a lack of contextual information (Liebig, Sauer, & Friedhoff, 2015, 320 et seq). As an example, a respondent could reason that, if Muslims residing in Germany are compared to the Christian majority, their religious denomination on average is correlated with a higher number of children and a more unfavourable employment and income situation.[3] Such considerations could lead

[1] Fairness perceptions on welfare state services and particularly early childcare are highly relevant for policy makers, who e.g. have to decide to what extent childcare services should be funded not only by general tax money but also by parents' financial contributions. In Germany, citizens are used to a highly government-financed educational sector. At the same time, childcare policies are devolved to the local municipalities, resulting in huge differences between contribution schemes within the country. Although there is some comparative scientific work on attitudes towards childcare as a governmental responsibility across different countries (see e.g. Guo & Gilbert, 2014) as well as a more vivid debate on how self-interest and normative values shape attitudes to welfare states and redistribution in general (e.g. Alesina & Angeletos, 2005; Owens & Pedulla, 2014; Svallfors, 2012), studies on individual preferences and fairness perceptions in the area of early childcare are virtually non-existent. Despite its importance, we decided to neglect an in-depth discussion of the topic, because the research question of this paper is a design-related one.

[2] Basic Law: Grundgesetz (GG) für die Bundesrepublik Deutschland vom 23. Mai 1949 (BGBl. S. 1), zuletzt geändert durch Artikel 1 des Gesetzes vom 23.12.2014 (BGBl. I S. 2438). See for English version https://www.bundestag.de/blob/284870/ce0d03414872b427e57fccb703634dcd/basic_law-data.pdf

[3] It is likely that the respondents' images of typical Muslim immigrants will be mainly shaped by the Turkish migrant workers actively recruited for blue-collar jobs in the 1960s: that is, the ethnic group that constitutes the vast majority of immigrants in Germany. Those Turkish migrant workers usually came from very poor educational backgrounds. In combination with more traditional gender roles and marriage norms (Haug, 2002), this aspect might explain the comparatively high number of Turkish households that uniquely rely on male employment in low status jobs (for the survey "Muslim Life in Germany": Haug, Müssig, & Stichs, 2009; for results from the Socio-Economic Panel: Tucci, 2016). In addition, Turkish women tend to have more children than German women, although the trend is diminishing for second generation immigrants (Naderi, 2015; Schmid & Kohls, 2011). Furthermore, it should be mentioned that data collection took place at the end of 2014, when the refugee crisis intensified and immigration was about to become a controversial discussion topic in the media. This political background might also have led respondents to associate the Muslim families in our vignettes with Syrian asylum-seekers, as far as other specifications

The child's mother is **working part time**, the father is **working full time**. The parents and the child are **living together** in a household. **This household's overall monthly net income** (including income from rent or other sources) is **2800 Euro**. The child's **grandparents are not available** to help with childcare. The family has **always lived in Konstanz** and **does not belong to any religious community**.
The fee for the day-care facility is **100 Euro** per month.

In your opinion, is this monthly childcar fee fair, or is it unfairly high or low?

unfairly										unfairly
low					fair					high
-5	-4	-3	-2	-1	0	1	2	3	4	5
□	□	□	□	□	□	□	□	□	□	□

Figure 1. Example of a vignette on just childcare fees, varied dimensions highlighted in bold letters

to the application of justice principles other than equality. More precisely, redistributive mechanisms based on current need or past contributions to, for example, the social welfare or tax system (equity) could play a role (for an overview of common justice principles see Deutsch, 1975; Forsyth, 2010, pp. 388–389). In unidimensional question formats, the researcher could only guess which assumptions the answer was based on. In the vignette study at hand, however, the number of children is held constant throughout the vignettes by an introductory text and employment status, income and migration background are among the experimentally varied dimensions. Religious denomination does not correlate with a particular neediness or specific (lower or higher) contributions to the tax system.

Our vignette design hence accounts for typical lines of argument that would allow respondents to use statistical discrimination as a strategy to judge. We are confident that this approach allows us to estimate an unbiased effect of religious affiliation net of other characteristics.

A second advantage of vignettes over single-item questions is their multidimensionality, which makes single dimensions less obtrusive. This is a particularly useful feature when eliciting sensitive information from respondents. The rather complex decision task undertaken by respondents requires a simultaneous evaluation of multiple factors, which makes it unlikely that respondents will concentrate on evaluating each dimension's sensitivity. This should be particularly true if respondents are elderly, have a low educational background and/or are not used to the question format. In other words, we expect that the potential of factorial surveys to reduce social desirability bias to some extent will differ between certain respondent subgroups. We will come back to this idea later, when we test empirically if age and educational background matters.

In practice, respondents will commonly answer a set of several vignettes, because it reduces survey costs to obtain more than one evaluation from each participant. Given these implementation practices, researchers have to decide if the sensitive dimension in a vignette varies throughout one respondent's vignette sequence (within subject design) or is kept constant in all the vignettes one respondent sees, but varies between respondents instead (between subject design). The latter approach has been proposed as an alternative that should further improve the reduction of social desirability bias in factorial surveys. This is based on the assumption that constant dimensions attract less attention than varying ones.

Technically, the two approaches also differ in the amount of traceability of sensitive answers on an individual level. In within subject designs, it is still technically feasible to extract information about individual discriminatory responses. The particular effects of all vignette dimensions can be estimated in respondent-specific regression models as long as the number of vignettes per respondent is high enough. In between subject designs, on the contrary, the researcher cannot determine if a particular respondent applied discriminatory judgement rules, different from the equality principle. This is true as long as one respondent's individual vignette set only comprises one category of the sensitive dimension and vignette evaluations cannot be compared to evaluations about a reference group. However, between subject designs come with lower statistical precision than within subject designs. Somewhat similar to RRTs, the additional privacy and reduction in social desirability bias comes with a loss in statistical power.

There is very little methodological research on within and between subject designs, meaning that it is unclear which

within the vignette allowed this.

implementation is superior. Empirical evidence is needed to find out whether the loss of statistical power pays off in terms of a reduction in social desirability bias. In this case, between subject designs could help to discover effects that would otherwise be disguised by socially desirable answers.

2.3 Previous Research

To the best of our knowledge, there is only one study that compares within and between subject design in a factorial survey: in their vignettes on the fairness of earnings, Auspurg et al. (2015) examine which variables should affect income from the respondents' point of view. In a split-half experiment, the sensitive dimension, namely the vignette person's sex, is either varied within or between respondents. In contrast to what would theoretically be expected, the between design does not reduce social desirability bias. Instead, only the respondents' judgements in the within condition indicate that being a woman should have a negative effect on earnings. However, the difference between both experimental conditions is not statistically significant ($p = 0.15$).

In addition, there are some studies (Burstin et al., 1980; Pager & Freese, 2004; Schuman & Bobo, 1988; Sniderman & Piazza, 2004; Steiner, Atzmüller, & Su, 2016) that examine the effect of sex, ethnicity or religion using scenario designs that consciously avoid varying the sensitive dimension within respondents. However, none of these studies can be directly compared to our approach: none of them included a within condition to actually test the underlying assumption that between designs are less prone to social desirability bias than within designs. Burstin et al. (1980) are the only ones to add a direct question condition as a reference group (which performs worse than the scenario format). Moreover, only the designs proposed by Pager and Freese (2004) and Steiner et al. (2016) are real multifactorial vignette design, whereas the other authors only vary one – namely the sensitive – dimension in their scenarios.

Another interesting finding of the vignette study by Auspurg et al. (2015) is that, when asked directly, the higher educated in particular stated that gender should not influence earnings (89% compared to 78% for the lower educated), while these differences disappear entirely in the vignette module. The authors conclude that the higher educated tend to conceal socially undesirable attitudes more than the lower educated and that factorial surveys are also an adequate method for reducing social desirability bias for groups that are particularly prone to concealing their true opinions.

In respect to the question if educational background and social desirability bias are correlated, two crucial – and somewhat contradicting – arguments have been put forward:

- On the one hand, the higher and the lower educated might indeed differ in their opinions. In this case, more liberal answers from higher educated respondents would represent honest answers. This mechanism was, for example, empirically confirmed in an RRT-study by Ostapczuk, Musch, and Moshagen (2009), in which the authors identified truly diverging opinions as the primary cause for different levels of racism reported by respondents with varying educational backgrounds.
- On the other hand, liberal answers could also be driven by a higher susceptibility to social desirability bias. Higher educated respondents might be more aware of the sensitivity of certain items or topics. Moreover, we assume that they are better prepared to do several things at the same time, such as answering a cognitively demanding question task and simultaneously considering issues of sensitivity. In line with previous findings by Auspurg et al. (2015), and to some extent also Ostapczuk et al. (2009), we argue that the higher educated are more susceptible to social desirability bias and should show a generally higher tendency to conceal socially undesirable opinions.

3 The Factorial Survey Experiment

3.1 Experimental Set-up and Survey Implementation

In our factorial survey module on the contributive fairness of childcare fees, eight dimensions with different numbers of levels were varied within the vignettes (see Table 1). It is our primary concern to test if the between design reduces social desirability bias, when respondents evaluate our most sensitive dimension – namely, the religious affiliation of the described vignette families.

To test which design feature is most effective in terms of reducing social desirability bias in such estimates, we collected vignette data in two different experimental conditions. Applying a split-half experiment, we randomly assigned respondents to a set of five vignettes in which the sensitive dimension either varied (within subject design) or in which it was kept at one specific level (between subject design). In this latter condition, respondents evaluated either only Christians, or only Muslims, or only families without any religious denomination.

All in all, the so-called vignette universe consisted of 20,736 possible combinations of dimensions for our vignette module. Since we could not implement all of these in a survey, and as we still wanted to ensure unbiased estimations for the main effects and all relevant interaction effects of the vignette dimensions, we reduced the number of vignettes to a d-efficient sample of 350 vignettes (d-efficiency: 90.2). In contrast to random samples, d-efficient samples are optimized on the basis of certain design principles (namely balance and independence) and therefore do not leave unbiased estimates to chance (for details see Dülmer, 2007; Kuhfeld, 1997; Kuhfeld, Randall, & Garratt, 1994).[4]

[4] Except for two interactions that we did not consider important (mother's employment status * income, father's employment status * income), all other main effects and interactions between the

Table 1
Varied dimensions in the vignette module on just childcare fees

Childcare fee
€ 0 / € 100 / € 200 / € 300 / € 400 / € 500

Civil status
single mother / parents living together

Mother's employment status
seeking work / housewife / working part time / working full time

Father's employment status
seeking work / househusband / working part time / working full time

Grandparents' support
grandparents can help with childcare / grandparents are not available to help with childcare

Net household income
€ 750 / € 1000 / € 1250 / € 1500 / € 2000 / € 5000
(based on the equivalent household disposable income the values were multiplied by 1.4 in the case of singlemother households: € 1050, € 1400, € 1750, € 2100, € 2800, € 7000)

Local connectedness
have always been living in Konstanz / moved from another German city / moved from abroad

Religious affiliation
does not belong to any religious community / belongs to a Christian community / belongs to a Muslim community

The 350 drawn vignettes were split into 70 different vignette decks, containing five vignettes each. To ensure that our estimates are not confounded with respondent characteristics, vignette decks were randomly assigned to respondents as soon as they started the questionnaire. At the same time, a random order of the vignettes within the individual decks accounted for potential order effects (as far as this is possible; see Su & Steiner, 2018, for a detailed discussion). Vignettes were presented sequentially, but respondents could go back in the questionnaire to edit previous responses if they wished to. With respect to the split-half experiment, respondents answered the same vignettes in both experimental conditions. Only the sensitive dimension changed between the groups.

By implementing our vignette module in a general population survey, we profit from the combined advantages of an experimental design plan that makes causal conclusions more credible and a survey with a heterogeneous respondent sample that ensures high external validity. The vignette module was part of the seventh wave of a general population panel survey that took place in 2014 in a small town in the southwest of Germany. The survey is a project involving cooperation between the town council and the University of Konstanz and has been conducted annually since 2008. The sample is drawn randomly by means of the ocal population register, meaning that all registered citizens of Konstanz that are at least 18 years old are eligible. Apart from the respondents already registered for the panel, additional refresher samples are drawn regularly to counteract the effects of panel mortality. Selected respondents receive a postal letter that invites them to sign up for the panel (for more details on the survey see Hinz, Mozer, & Walzenbach, 2015).

Data collection mainly took place online, with paper questionnaires available upon request for respondents without access to the internet. However, our vignette module was only implemented in the online version of the questionnaire. Our sample therefore consists of 1255 online participants who at least evaluated one vignette. Excluding those respondents whose evaluations did not differ at all between vignettes leaves us with 5878 vignette evaluations, answered by 1201 respondents (588 in the within subject design, 613 in the between subject design). This means that every vignette has been evaluated between 13 and 20 times. The average rating was 0.66, which is slightly above the mid-point of the ordinal scale and indicates that respondents generally perceived the overall range of the presented childcare fees as realistic.

Figure 2 shows the responses to all vignettes dependent on the experimental condition that respondents were assigned to. The overall distribution of the vignette evaluations looks rather similar in the within and the between design. In both groups, respondents used the whole range of the scale and there are no major differences in floor or ceiling effects that would limit the comparability across experimental conditions.

vignette dimensions can be estimated in our case.

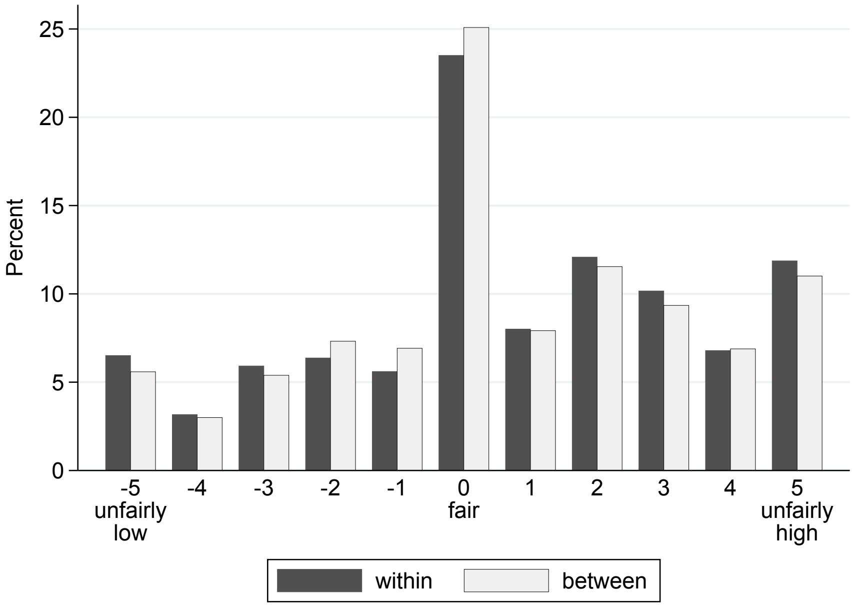

Figure 2. Distribution of responses dependent on experimental condition

Slight differences occur in the number of vignettes evaluated as “fair” (23.5 versus 25.1%) as well as in the willingness to select one of the extreme categories of the scale (6.5 versus 5.6% and 11.9 versus 11.0%). These tendencies result in a higher variance in the within condition (8.05 versus 7.55; $p = 0.044$ according to Levene’s robust variance test). Apart from that, differences are minor: a median and a Wilcoxon rank-sum test do not suggest any significant group differences ($p = 0.09$ and $p = 0.32$).

3.2 Hypotheses

Our first crucial hypothesis aims to answer the question regarding which vignette design should be chosen to elicit honest answers to sensitive questions. As argued above, we assume that constant information is less salient to the respondent than a varying dimension. Moreover, a vignette evaluation should only become sensitive if the individual respondent’s evaluations provide a reference point to judge if a certain answer is discriminatory, which is not the case in the between condition:

1) If the between subject design reduces social desirability bias, the effect of the sensitive dimension (namely religious denomination) should be stronger in the between condition than in the within condition.

Secondly, we will examine if differences in educational background lead to different levels of socially desirable answers: that is, differential social desirability bias, as proposed by Auspurg et al. (2015). As discussed above, more liberal answers from higher educated respondents could in principle also reflect true differences in opinion. For our concrete case, this means that we might find less discriminatory answers among the higher educated but the education effect should not differ across experimental conditions. However, if the higher educated are more likely to conceal socially undesirable opinions, this tendency should be more pronounced in the within than in the between design, given that the latter successfully reduces bias. In other words, we should find an interaction between educational background and experimental condition:

2a) If the higher educated conceal true opinions, the design effect should be bigger for the higher educated (differential SDB).

A similar argument can be made for elderly respondents. Respondents from different age groups are likely to differ in their opinions but also in their susceptibility to social desirability bias. While elderly people should respond rather truthfully, irrespective of experimental condition, younger people will deal more easily with the complex question format and might still have enough cognitive capacity to conceal socially undesirable opinions. Design decision should therefore particularly matter for younger respondents:

2b) The design effect should be bigger for younger respondents.

In a third step, we will examine to what extent the unequal evaluations of vignette families’ religion depend on the respondents’ own affiliation. According to the social psychological concept that humans favour ingroups and discriminate against outgroups (Tajfel, Billig, Bundy, & Flament, 1971), we expect respondents to benefit vignette persons with similar characteristics to themselves. Put concretely, religious respondents should indulge other religious people but give harsh judgements when evaluating vignette families without religious denomination. The opposite should be true for respondents without any religious affiliation. If the between subject condition reduces social desirability bias, this underlying mechanism of affectivity-based favouritism should be stronger in the between condition, while respondents should tend to suppress such unjustifiable judgement rules in the within subject design:

3) In line with the concept of ingroup favouritism, we expect an interaction effect between the vignette family's and the respondent's religious denomination that should be more pronounced in the between condition.

4 Results

Since vignette judgements were made on an 11-point-scale, we test our hypothesis with linear regression models. Beside the vignette person's religious denomination, which is the crucial sensitive variable in our vignette module, we also included all the other vignette dimensions (namely childcare fee, family status, employment constellation variables, grandparents' support, household income, local connectedness) in all the analyses to be presented. In addition, we control for certain respondent characteristics: sex, age, educational background and religious affiliation.[5] Following the recommendation of an anonymous reviewer, we furthermore included dummy variables for the 70 different sets of vignettes to account for set-specific context effects (see Su & Steiner, 2018, for a detailed discussion) for a detailed discussion).[6] Since individual respondents evaluated several vignettes, cluster robust standard errors were applied to account for the nested data structure (Hox, Kreft, & Hermkens, 1991). We chose to reduce our empirical results in the following paragraph to concise graphical presentations of the relevant sensitive vignette dimension. The full regression tables can be found in the appendix.

4.1 Comparison of Within and Between Subject Design

Most crucially, it was hypothesised that in comparison to the within subject design, the between subject design should reduce social desirability bias and bring about more honest (that is, discriminatory) answers. Figure 3 shows the vignette evaluations for the two experimental conditions and the three religious denominations of the described families. No religious affiliation was used as the reference category in the corresponding regression model. Negative point estimates indicate that the respondents considered the childcare fees to be unfairly low. 95% confidence intervals are shown in grey. The additional markers towards the ends of the lines represent 90% confidence intervals. Despite comparable numbers of cases, we generally find larger confidence intervals in the between subject condition than in the within design. These reflect the design-dependent differences in statistical precision that were already mentioned earlier.

All the presented effects are negative, meaning that respondents generally want religious vignette families to pay higher fees for childcare in a public institution than families without a religious denomination. This tendency was stronger in the between condition and particularly for Muslim vignette families.

In the within subject design, the vignette ratings for Christians and Muslims did not statistically differ from the evaluations of vignette families without religious affiliation, while there were small but significant differences between vignette families with and without religious background in the between subject design ($p = 0.03$ for Christian families and $p = 0.004$ for Muslims). Although the design effect was biggest for Muslim vignette families, the interaction between the experimental design and Muslim denomination was only significant on a 10% level in a pooled model ($p = 0.078$).

Put together, these results indicate that respondents gave slightly less socially desirable answers if the sensitive dimension was varied only between respondents. As hypothesised, the between design thus seems to have favourable effects on the reduction of social desirability bias, although the results marginally failed to reach the conventional significance level.

4.2 Social Desirability Bias, Educational Background and Age

Our second hypothesis suggests different response behaviours according to age (see Figure 5) and the highest educational degree obtained (see Figure 4). In other words, the between condition should be particularly helpful to reduce social desirability bias in higher educated and younger respondent groups, while the lower educated and the elderly would always answer rather honestly, irrespective of the experimental condition they have been assigned to.

Figure 4 shows the vignette judgements elicited for Muslim families (compared to those without any religious affiliation) separately for three different education groups: those

[5] Assuming that randomisation worked ideally, it would be sufficient to run regressions without any respondent characteristics to get valid estimates for the vignette dimensions. However, additional analyses showed that respondent age and highest educational degree did not distribute equally to all experimental conditions. Although the correlations between these variables and the assigned experimental (sub)group are not very strong (highest Cramer's V value: 0.053), a chi^2 test suggests significant differences. As a consequence, all the presented regression models include sociodemographic information as control variables.

[6] Some of those set indicators were significant in our regression models. However, in a multilevel model with random effects for vignette sets and respondents only 0.4% of the residual variance was attributed to vignette sets. Accordingly, vignette judgments do not seem to vary much between sets.

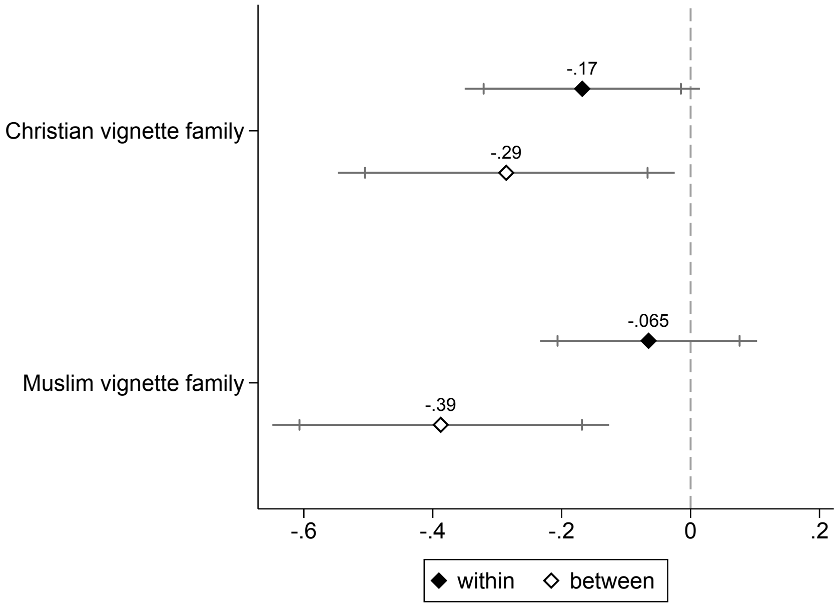

Figure 3. Comparison of within and between subject design

having degrees lower than a university entrance qualification, those with higher secondary education allowing university entrance and those with tertiary education together with university students.[7] Figure 5 does the same for the age groups 18 to 30, 31 to 59 and 60 or older.

Our crucial finding is that the between condition yields lower coefficients, and thus probably more valid responses, for all subgroups. We do not find support for a systematic interaction between education and design (hypothesis 2a), nor can we see consistent differences in opinions dependent on educational background. If anything, the between subject design seems to be slightly more helpful in regard to reducing social desirability bias in younger respondents (hypothesis 2b). However, the difference between the estimated coefficients for the two design conditions is far from reaching statistical significance.

All in all, we do not find evidence for our hypotheses that the between design's potential to reduce social desirability bias depends on respondent characteristics. Rather, the between subject design seems to be slightly superior for the majority of respondents, irrespective of the respondents' age and educational background.

4.3 Ingroup/Outgroup Differences

In the third hypothesis, it was argued that the unequal evaluations of the vignette families' religious backgrounds that were reported in Figure 3 would occur as a result of ingroup favouritism. To test this assumption, the vignette evaluations were examined separately for respondents without (see left column) and with (see right column) a religious denomination in Figure 6.[8] Again, no religious affiliation is used as a reference group and negative point estimates indicate that respondents wanted religious vignette families to pay higher childcare fees.

As in Figure 3, none of the coefficients for the within condition differs significantly from zero. This means that respondents in the within design evaluate all religious denominations equally, which is in line with our assumption that respondents should answer more socially desirable in this experimental condition.

At the same time, we do find some more discriminatory answers in the between design, but interestingly only for respondents without a religious denomination, who want religious vignette families to pay significantly higher childcare fees. This is true for Christians (design effect: $p = 0.078$) as well as for Muslim families (design effect: $p = 0.04$), which indicates that respondents do not discriminate against a specific religion but against religion in general. The significant, negative coefficients presented in Figure 3 are thus solely driven by those respondents without religious affiliation, whereas religious respondents treat all vignette families

[7] Figure 4 shows an easily understandable but somewhat simplified summary of our results that necessarily leaves out some heterogeneity within the displayed education groups. In the survey, educational background was measured in more detail, partly producing categories with extremely small numbers of cases for lower educational backgrounds.

[8] Dividing respondents' religious denominations further into Christian, Muslims and others would not have made sense due to rather small numbers of cases in the two latter groups.

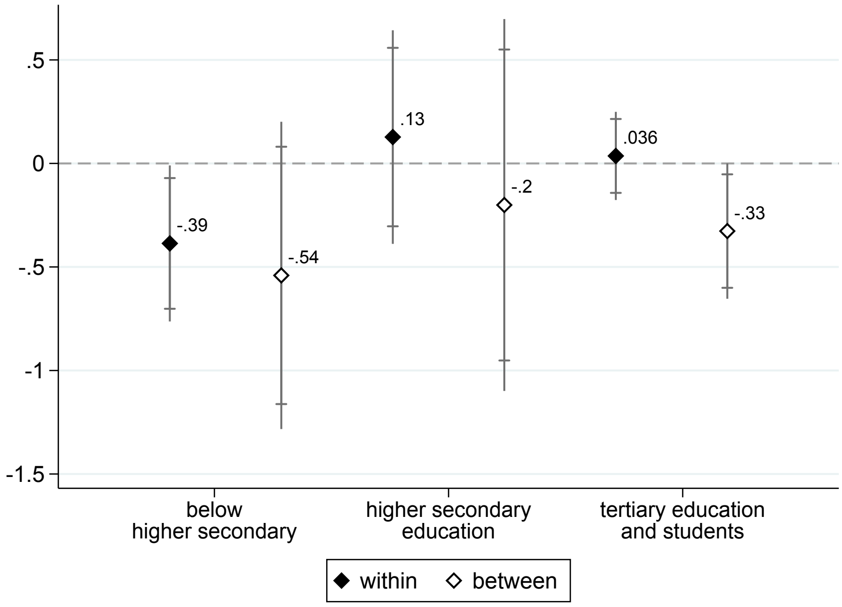

Figure 4. Judgements of the vignette dimension "Muslim" dependent on educational background

equally, irrespective of their religious beliefs.

In regard to our hypothesis, this means that we only find ingroup favouritism in the subsample of respondents in the between condition that do not belong to any religious community. The interactions between the respective respondent characteristic and the vignette dimensions are both statistically significant at a 5% level (for vignette descriptions of Christian families: p = 0.005, for Muslim families: p = 0.03).

However, we refrain from concluding that irreligious people are in general more discriminatory against outgroups. Instead, this finding might have to do with the fact that the vignette introduction text explicitly talked about public childcare institutions. Although religious childcare facilities are in large part financed by public authorities in Germany, they tend to include religious practices in their everyday routines, which arguably makes them less attractive for irreligious parents (Frerk, 2002, 121 et seq; Müller, 2013, Chapters 3 and 5). It could be this peculiarity that leads respondents to the conclusion that Christian childcare institutions are meant for religious families, while, as a consequence, public ones should preferably take care of those citizens without religious affiliations.

5 Summary and Discussion

Despite the increasing popularity of factorial surveys in the social sciences over recent years, little is known about favourable design choices to enhance data quality when asking sensitive questions. This article has provided an experimental approach to reducing social desirability bias in factorial surveys. For this purpose, we implemented a split-half design in a vignette module on just fees for early childcare, which was run as part of a general population survey in the southwest of Germany. The most sensitive vignette dimension, namely the religious affiliation, was varied either within or between respondents.[9] The crucial question was whether the loss in statistical power associated with between subject designs would pay off in terms of a reduction in social desirability bias. In a worst case scenario, we could have failed to discover an effect in the between subject condition merely due to a lack of statistical power.

Assuming that less socially desirable answers reflect more truthful response behaviour, however, our results suggest otherwise. The between subject design indeed reduced social desirability bias and thus is a suitable method for enhancing data quality when sensitive topics are assessed. While respondents in the within subject design treated the described vignette families equally, irrespective of their religious affiliation, respondents in the between subject design wanted both Christian and Muslim families to pay significantly higher

[9] In the course of the review process, we ran extensive robustness checks on further potential differences between the experimental conditions apart from this intended variation in design. These checks addressed the use of the response scale, context effects and confounding structures. Although discussing them in detail was beyond the scope of this paper, we sometimes referred to them throughout the paper to strengthen our line of argument. A summary of all these checks is available upon request.

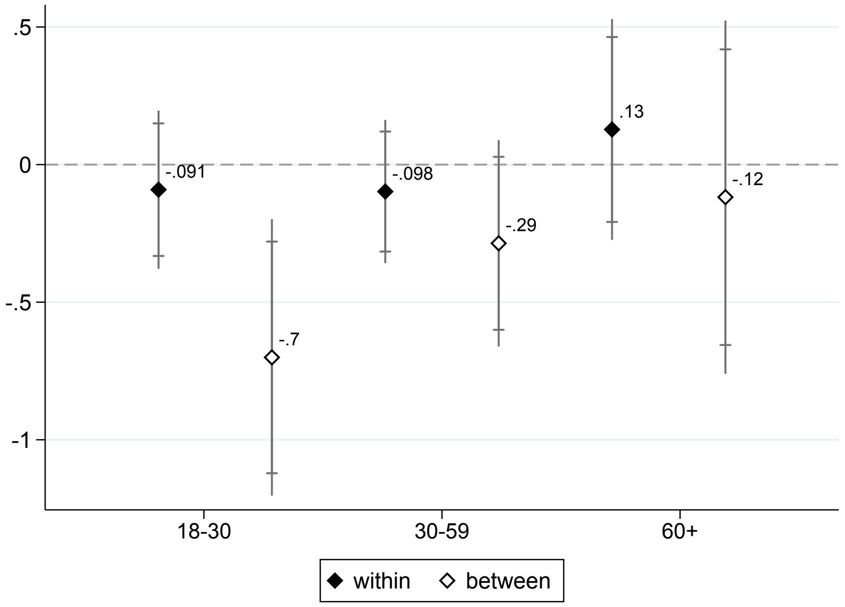

Figure 5. Judgements of the vignette dimension "Muslim" dependent on respondents' age

fees for public childcare institutions than families without any religious denomination. Further analyses showed that those differences were driven by ingroup favouritism of respondents without any religious affiliation, who benefited vignette families with similar characteristics to themselves but discriminated against the perceived religious outgroup. Interestingly, this mechanism could only be uncovered in the between subject condition: we would have missed it in the within design, where respondents showed a stronger tendency to distort socially undesirable opinions.

As discussed above, we do not attribute these findings to more discriminatory attitudes of irreligious people in general. Instead, we suspect that these results might reflect structural peculiarities of the German childcare system. Further research could also show to what extent Christian and Muslim respondents demand higher childcare fees for undenominational children in religious institutions. Since our vignette module concentrated on public institutions facilities, we cannot draw any conclusions regarding this aspect.

With respect to differential social desirability bias as a function of the respondent's age and education level, we found somewhat inconclusive results. The between subject design was not particularly useful for eliciting sensitive information from any of the examined subgroups. Rather, the between subject design seemed to be slightly superior for the majority of respondents, irrespective of the respondents' age and educational background.

In our data, the benefits of the between design outweighed its disadvantages in terms of statistical power. However, this reduction in social desirability bias is paid for by more imprecise point estimates (as indicated by the bigger confidence intervals in all our graphs). We therefore recommend practitioners implement within subject designs whenever the questions' sensitivity level allows it. If social desirability bias is a threat to data quality, however, between subject designs can be a valuable tool with which to foster truthful answers. In cases of uncertainty, pre-tests should be used as a method to assess the sensitivity level of certain questions or vignette dimensions, so as to adapt design choices accordingly.

It is noteworthy that our main result is contradictory to the only other study conducted so far that has compared within and between subject designs to reduce social desirability bias in factorial survey modules. For their vignettes on the fairness of earnings, Auspurg et al. (2015) report that, in contrast to their expectations, only respondents in the within condition declared that women should earn significantly less than men. Although the difference between the experimental conditions was not statistically significant ($p = 0.14$), those results clearly point in a different direction to ours. It is difficult to explain these differences, although some considerations can be mentioned. For one thing, not only did the topic of the vignette module differ from our vignette module, the sensitive dimension in question also differed. We cannot exclude the possibility that respondents associated the two vignette modules with unequal levels of sensitivity, e.g. because the gender wage gap is, in contrast to childcare fees depending on religion, a well-known reality that respondents may have become used to.

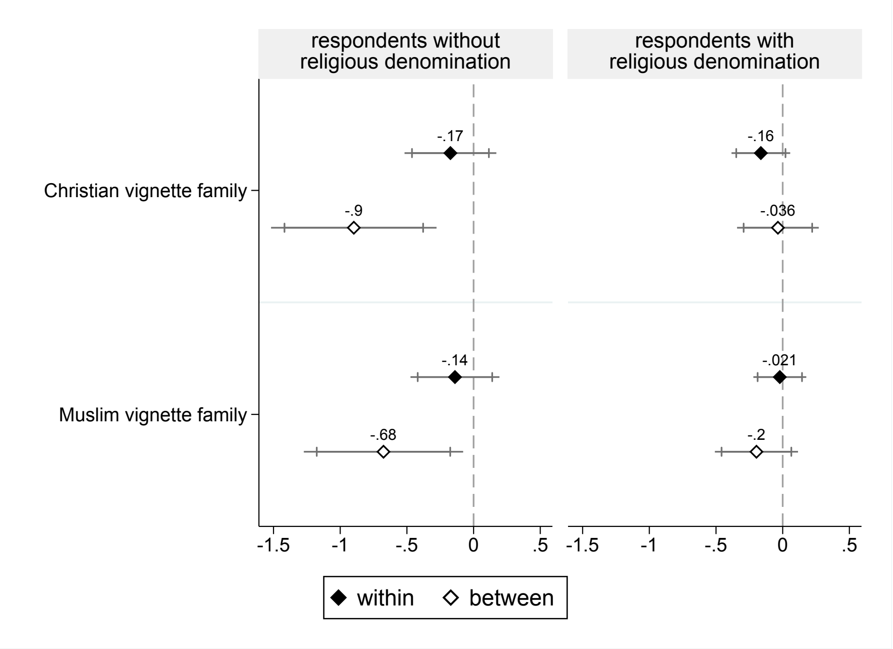

Figure 6. Ingroup/outgroup differences dependent on respondents' religious denomination

In addition, discrepancies in statistical power could play a role. Between subject designs always go hand in hand with a loss of power, compared to within subject designs, meaning that the main effect for the sensitive dimension is estimated with higher statistical power in the within compared to the between condition. This design feature that makes significant effects in the within condition more likely (keeping the effect and sample size equal) could theoretically offer a design-based explanation for the findings presented by Auspurg et al. (2015). However, this is not true for our study as statistical power should instead undermine the significant effects we consistently found in the between condition.

Moreover, although both studies were conducted with a similarly heterogeneous general population sample, there are some remarkable differences in the number of vignettes per respondent (5 vs 10), in the sample size used (1139 vs 437, if the actual numbers of valid cases in the crucial regression models are considered), as well as in the number of levels the sensitive vignette dimension has (2 vs 3) – characteristics that have effects on statistical power (Auspurg & Hinz, 2015, Chapter 3.5.2; Snijders, 2005). For the case at hand, an exact power analysis could be carried out for the two-level interaction effect between sensitive dimension and vignette design, by means of simulations. Although not many additional insights are to be expected from post-hoc power analysis for our particular case (O'Keefe, 2007), a simulation-based approach with a more comprehensive scope on a variety of design features would clearly be very valuable, to enable other researchers to take better design decisions when planning vignette modules.

Last but not least, it is an inherent feature of inferential statistics that hypotheses can sometimes be erroneously dismissed (type I error) or retained (type II error). Further research on the topic is needed to confirm and strengthen the scarce empirical evidence that exists so far in order to conclusively answer the question to what extent and under which circumstances between designs help to reduce social desirability bias in factorial surveys.

In terms of our experimental design, we considered it crucial that our vignettes included information on important context factors to account for statistical discrimination as an evaluation strategy in the context of just childcare fees. However, this approach came with the disadvantage of a very comprehensive vignette universe, meaning that we could not use all possible combinations in the vignette universe, but had to rely on a d-efficient design and random assignment of vignettes to respondents. Future studies might make different design choices and include less vignette dimensions but a design that evaluates all possible vignette combinations to complete the picture.

6 Acknowledgements

I am grateful to Katrin Auspurg and Thomas Hinz for their support in implementing this factorial survey experiment as well as their valuable feedback on an earlier draft of this paper. Also I would like to thank the anonymous reviewers for their knowledgeable and helpful comments. All remaining errors are of course my own. Earlier versions of this paper were presented at European Survey Research Association (ESRA) 2015 in Reykjavik and at VIU Rational Choice Conference 2016 in Venice.

References

Alesina, A. & Angeletos, G.-M. (2005). Fairness and redistribution. *The American Economic Review*, *95*(4), 960–980.

Arbeitsgruppe Hinz. (2014). Konstanzer Bürgerbefragung Welle 7 Data. Retrieved from https://www.buergerbefragung.uni-konstanz.de/

Armacost, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An empirical comparison of direct questioning, scenario, and randomized response methods for obtaining sensitive business information. *Decision Sciences*, *22*(5), 1073–1090. doi:10.1111/j.1540-5915.1991.tb01907.x

Atzmüller, C. & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(3), 128–38.

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

Auspurg, K., Hinz, T., & Sauer, C. (2017). Why should women get less? evidence on the gender pay gap from multifactorial survey experiments. *82*(1), 179–210.

Auspurg, K., Hinz, T., Sauer, C., & Liebig, S. (2015). The factorial survey as a method for measuring sensitive issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving survey methods: Lessons from recent research.* (pp. 137–149). New York: Routledge.

Burstin, K., Doughtie, E. B., & Raphaeli, A. (1980). Contrastive vignette technique: An indirect methodology designed to address reactive social attitude measurement. *Journal of Applied Social Psychology*, *10*(2), 147–165. doi:10.1111/j.1559-1816.1980.tb00699.x

Corstange, D. (2014). Foreign-sponsorship effects in developing-world surveys: Evidence from a field experiment in lebanon. *Public Opinion Quarterly*, *78*(2), 474–484. doi:10.1093/poq/nfu024

De Leeuw, E. D. (1992). *Data quality in mail, telephone, and face to face surveys.* Amsterdam: TT-Publikaties.

Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues*, *31*(3), 137–149.

Dülmer, H. (2007). Experimental plans in factorial surveys random or quota design? *Sociological Methods & Research*, *35*(3), 382–409. doi:10.1177/0049124106292367

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of forced responses in a randomized response model. *Sociological Methods & Research*, *11*(1), 89–100. doi:10.1177/0049124182011001005

Esser, H. (1986). Können Befragte lügen? Zum Konzept des 'wahren' Wertes im Rahmen der handlungstheoretischen Interpretation des Befragtenverhaltens. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie: KZfSS*, *38*(2), 314–336. Retrieved from https://ub-madoc.bib.uni-mannheim.de/18625/

Forsyth, D. R. (2010). *Group dynamics.* Please provide address: Wadsworth.

Fox, J. A. (2015). *Randomized response and related methods: Surveying sensitive data.* Los Angeles: Sage.

Frerk, C. (2002). *Finanzen und Vermögen der Kirchen in Deutschland.* Alibri Verlag.

Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology.* Hoboken, NJ: Wiley.

Guo, J. & Gilbert, N. (2014). Public attitudes toward government responsibility for child care: The impact of individual characteristics and welfare regimes. *Children and Youth Services Review*, *44*, 82–89.

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

Haug, S. (2002). Familie, soziales Kapital und soziale Integration. Zur Erklärung ethnischer Unterschiede in Partnerwahl und generativem Verhalten bei jungen Erwachsenen deutscher, italienischer und türkischer Abstammung. *Zeitschrift Für Bevölkerungswissenschaft*, *27*(4), 393–425.

Haug, S., Müssig, S., & Stichs, A. (2009). Muslim life in germany. a study conducted on behalf of the German Conference on Islam. Research report 6. Bundesamt für Migration und Flüchtlinge. Retrieved from http://www.bamf.de/SharedDocs/Anlagen/EN/Publikationen/Forschungsberichte/fb06-muslimisches-leben.html

Hinz, T., Mozer, K., & Walzenbach, S. (2015). Kommune und Bürger im Dialog und Lebenszufriedenheit: Ergebnisse der Konstanzer Bürgerbefragung 2014 - 7. Welle. Statistik Bericht 2/2015, Stadt Konstanz. Retrieved from http://nbn-resolving.de/urn:nbn:de:bsz:352-0-314583

Höglinger, M. & Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: False positives undermine the crosswise-model RRT. *Political Analysis*, *25*(1), 131–137. doi:10.1017/pan.2016.5

Höglinger, M. & Jann, B. (2016). More is not always better: An experimental individual level validation of the randomized response technique and the crosswise model. University of Bern Social Sciences Working Paper

No.18. Retrieved from https://ideas.repec.org/p/bss/wpaper/18.html

Holbrook, A. L. & Krosnick, J. A. (2010). Measuring voter turnout by using the randomized response technique evidence calling into question the method's validity. *Public Opinion Quarterly*, *74*(2), 328–343. doi:10.1093/poq/nfq012

Houle, B., Angotti, N., Clark, S. J., Williams, J., Gomez-Olive, F. X., Menken, J., & . . . Tollman, S. M. (2016). Let's talk about sex, maybe: Interviewers, respondents, and sexual behavior reporting in rural south africa. *Field Methods*, *28*(2), 112–132. doi:10.1177/1525822X15595343

Hox, J. J., Kreft, I. G. G., & Hermkens, P. L. J. (1991). The analysis of factorial surveys. *Sociological Methods & Research*, *19*(4), 493–510. doi:10.1177/0049124191019004003

Jasso, G. & Webster, M. J. (1999). Assessing the gender gap in just earnings and its underlying mechanisms. *Social Psychology Quarterly*, *62*(4), 367–380.

Kuhfeld, W. F. (1997). Efficient experimental designs using computerized searches. Sawtooth Software Research Paper Series. Retrieved from https://www.sawtoothsoftware.com/support/technical-papers/design-of-conjoint-experiments/efficient-experimental-designs-using-computerized-searches-1997

Kuhfeld, W. F., Randall, T. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research*, *31*(4), 545–557. Retrieved from http://www.jstor.org/stable/3151882

Lensvelt-Mulders, G. (2008). Surveying sensitive topics. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 461–478). Lawrence Erlbaum Ass.

Liebig, S., Sauer, C., & Friedhoff, S. (2015). Empirische Gerechtigkeitsforschung mit dem faktoriellen Survey. *Soziale Welt. Sonderheft: Experimente in den Sozialwissenschaften*, *22*, 316–334.

Liu, M. & Stainback, K. (2013). Interviewer gender effects on survey responses to marriage-related questions. *Public Opinion Quarterly*, *77*(2), 606–618. doi:10.1093/poq/nft019

Müller, E. (2013). *Gott hat hohe Nebenkosten: wer wirklich für die Kirchen zahlt.* Kiepenheuer & Witsch.

Mutz, D. C. (2011). *Population-based survey experiments.* Princeton, NJ: Princeton University Press.

Naderi, R. (2015). Kinderzahl und Migrationshintergrund. Ein Vergleich zwischen Frauen türkischer Herkunft mit oder ohne eigene Wanderungserfahrung sowie Frauen ohne Migrationshintergrund in Westdeutschland. *ZfF – Zeitschrift für Familienforschung / Journal of Family Research*, *27*(3), 322–342. doi:10.3224/zff.v27i3.21277

O'Keefe, D. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures*, *1*(4), 291–299. doi:10.1080/19312450701641375

Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology*, *39*(6), 920–931. doi:10.1002/ejsp.588

Owens, L. & Pedulla, D. (2014). Material welfare and changing political preferences: The case of support for redistributive social policies. *Social Forces*, *92*(3), 1087–1113.

Pager, D. & Freese, J. (2004). *Who deserves a helping hand? attitudes about government assistance for the unemployed by race, incarceration status, and worker history.* Proposal Submitted to Time-Sharing Experiments in the Social Sciences (TESS). Retrieved from http://www.ssc.wisc.edu/%5C~%7B%7Djfreese/soc750/pagerfreese%5C_tess2.pdf

Rossi, P. & Andersen, A. (1982). The factorial survey approach: An introduction. In R. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments: The factorial survey approach.* (pp. 15–67). Beverly Hills: Sage.

Schmid, S. & Kohls, M. (2011). Generatives Verhalten und Migration. Eine Bestandsaufnahme des generativen Verhaltens von Migrantinnen in Deutschland. Forschungsbericht 10. Bundesamt für Migration und Flüchtlinge. Retrieved from http://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/Forschungsberichte/fb10-generativesverhaltenundmigration.pdf?__blob=publicationFile

Schnell, R. & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, *21*(3). Retrieved from http://search.proquest.com/docview/1266792195/abstract/F112104D4A75457FPQ/2

Schuman, H. & Bobo, L. (1988). Survey-based experiments on white racial attitudes toward residential integration. *American Journal of Sociology*, *94*(2), 273–299. Retrieved from http://www.jstor.org/stable/2780776

Sniderman, P. M. & Piazza, T. (2004). *Black pride and black prejudice.* Princeton, N.J.: Princeton University Press.

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. S. Howell (Eds.), *Encyclopedia of statistics in behavioral science vol. 3* (pp. 1570–1573). Chicester: Wiley. doi:10.1002/0470013192.bsa492

Steiner, P. M., Atzmüller, C., & Su, D. (2016). Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap. *Journal of Methods and Measurement in the Social Sciences*, *7*(2), 52–90.

Su, D. & Steiner, P. M. (2018). An evaluation of experimental designs for constructing vignette sets in factorial surveys. *Sociological Methods & Research. online first*. doi:10.1177/0049124117746427

Svallfors, S. (2012). *Contested welfare states: Welfare attitudes in europe and beyond.* Stanford: Stanford University Press.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. doi:10.1002/ejsp.2420010202

Tourangeau, R., Conrad, F. G., & Couper, M. (2013). *The science of web surveys.* Oxford University Press.

Tourangeau, R., Presser, S., & Sun, H. (2014). The impact of partisan sponsorship on political surveys. *Public Opinion Quarterly*, *78*(2), 510–522. doi:10.1093/poq/nfu020

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2010). *The psychology of survey response.* Cambridge: Cambridge University Press.

Tourangeau, R. & Smith, T. W. (1996). Asking sensitive questions. the impact of data collection mode, question format and question context. *Public Opinion Quarterly*, *60*(2), 275–304. doi:10.1086/297751

Tucci, I. (2016). Lebenssituation von Migranten und deren Nachkommen. In Statistisches Bundesamt & Wissenschaftszentrum Berlin für Sozialforschung (Eds.), *Datenreport 2016. Ein Sozialbericht für die Bundesrepublik Deutschland* (pp. 236–244). Bonn: Bundeszentrale für politische Bildung.

Turner, M., Sturgis, P., Martin, D., & Skinner, C. (2015). Can interviewer personality, attitudes and experience explain the design effect in face-to-face surveys? In *Improving survey methods. lessons from recent research* (pp. 122–136). New York: Routledge.

Umesh, U. N. & Peterson, R. A. (1991). A critical evaluation of the randomized response method. applications, validations and research agenda. *Sociological Methods & Research*, *20*(1), 104–138. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/0049124191020001004

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, *38*(3), 505–520. doi:10.1016/j.ssresearch.2009.03.004

Walzenbach, S. & Hinz, T. (2014). *Pouring water into wine. the advantages of the crosswise model asking sensitive questions revisited.* Presented at the VIU Rational Choice Conference, Venice.

Appendix
Tables

Appendix tables start on next page.

Table A1
Regression models for Figure 3

	Within	Between
sex	−0.20^{+}	0.25*
	(−0.11)	(−0.11)
age: 31-59	0.10	0.19
	(−0.14)	(−0.13)
age: 60^{+}	−0.04	0.31^{+}
	(−0.17)	(−0.16)
educ: higher second.	−0.43*	0.41*
	(−0.19)	(−0.19)
educ: tertiary	−0.18	0.25^{+}
	(−0.13)	(−0.14)
religious affiliation	−0.11	−0.39**
	(−0.13)	(−0.13)
childcare fee	0.95**	0.90**
	(−0.02)	(−0.02)
family status	−0.56**	−0.57**
	(−0.07)	(−0.07)
m_housewife	−0.05	−0.13
	(−0.10)	(−0.09)
m_parttime	−0.13	−0.09
	(−0.10)	(−0.10)
m_fulltime	−0.18^{+}	0.07
	(−0.10)	(−0.10)
f_househusband	−0.24*	−0.08
	(−0.10)	(−0.10)
f_parttime	−0.19^{+}	0.03
	(−0.10)	(−0.09)
f_fulltime	−0.45**	−0.25*
	(−0.10)	(−0.10)
grandparents	0.26**	0.04
	(−0.07)	(−0.06)
income	−0.55**	−0.60**
	(−0.03)	(−0.03)
from Germany	−0.08	0.07
	(−0.09)	(−0.09)
from abroad	0.01	0.17^{+}
	(−0.09)	(−0.09)
rel_christian	−0.17^{+}	−0.29*
	(−0.09)	(−0.13)
rel_muslim	−0.07	−0.39**
	(−0.09)	(−0.13)
_cons	6.04**	6.57**
	(−0.56)	(−0.36)
R^2	0.5	0.45
N (vignettes)	2730	2865
N (respondents)	556	583

Cluster robust standard errors in parentheses. All models additionally control for set dummies.
$^{+}$ $p < 0.10$ * $p < 0.05$ ** $p < 0.01$

Table A2
Regression models for Figure 4

	Lower education		Uni entrance		Tertiary education	
	Within	Between	Within	Between	Within	Between
sex	−0.45*	−0.02	−0.39^{+}	0.37	−0.02	0.29*
	(−0.20)	(−0.23)	(−0.23)	(−0.34)	(−0.15)	(−0.14)
age: 31-59	0.23	−0.44	0.16	0.91	0.08	0.24
	(−0.38)	(−0.32)	(−0.80)	(−0.59)	(−0.15)	(−0.15)
age: 60^{+}	0.03	−0.23	−0.89	0.54	0.08	0.31^{+}
	(−0.37)	(−0.32)	(−0.88)	(−0.74)	(−0.23)	(−0.19)
educ: higher second.	-	-	-	-	-	-
	-	-	-	-	-	-
educ: tertiary	-	-	-	-	-	-
	-	-	-	-	-	-
religious affiliation	−0.10	0.17	−0.35	−0.69	−0.01	−0.29^{+}
	(−0.28)	(−0.27)	(−0.34)	(−0.42)	(−0.16)	(−0.15)
childcare fee	0.91**	0.88**	0.96**	0.93**	0.96**	0.90**
	(−0.05)	(−0.06)	(−0.08)	(−0.07)	(−0.03)	(−0.03)
family status	−0.53**	−0.50**	−0.62**	−0.58**	−0.57**	−0.58**
	(−0.16)	(−0.16)	(−0.22)	(−0.20)	(−0.09)	(−0.09)
m_housewife	0.17	0.05	−0.74*	−0.27	−0.05	−0.17
	(−0.20)	(−0.23)	(−0.35)	(−0.25)	(−0.13)	(−0.12)
m_parttime	−0.08	0.15	−0.25	−0.19	−0.16	−0.19
	(−0.22)	(−0.22)	(−0.33)	(−0.29)	(−0.13)	(−0.12)
m_fulltime	−0.34^{+}	0.24	−0.23	−0.28	−0.10	0.09
	(−0.19)	(−0.22)	(−0.33)	(−0.25)	(−0.14)	(−0.12)
f_househusband	−0.39^{+}	−0.32	−0.32	−0.19	−0.20	0.03
	(−0.20)	(−0.22)	(−0.33)	(−0.29)	(−0.13)	(−0.12)
f_parttime	−0.50*	0.02	−0.14	0.05	−0.08	0.02
	(−0.21)	(−0.22)	(−0.27)	(−0.28)	(−0.13)	(−0.12)
f_fulltime	−0.80**	−0.42^{+}	−0.69*	−0.58^{+}	−0.28*	−0.14
	(−0.22)	(−0.23)	(−0.29)	(−0.33)	(−0.13)	(−0.13)
grandparents	0.45**	−0.02	0.47*	−0.23	0.15	0.12
	(−0.15)	(−0.14)	(−0.20)	(−0.23)	(−0.09)	(−0.08)
income	−0.55**	−0.64**	−0.53**	−0.52**	−0.57**	−0.59**
	(−0.05)	(−0.05)	(−0.08)	(−0.07)	(−0.03)	(−0.03)
from Germany	−0.21	0.03	−0.23	0.16	0.00	0.07
	(−0.16)	(−0.21)	(−0.24)	(−0.25)	(−0.11)	(−0.11)
from abroad	−0.19	0.43*	−0.18	0.08	0.15	0.09
	(−0.19)	(−0.19)	(−0.26)	(−0.27)	(−0.12)	(−0.11)
rel_christian	−0.13	−0.64^{+}	−0.43	0.02	−0.15	−0.29^{+}
	(−0.18)	(−0.34)	(−0.28)	(−0.45)	(−0.12)	(−0.16)
rel_muslim	−0.39*	−0.54	0.13	−0.20	0.04	−0.33^{+}
	(−0.19)	(−0.38)	(−0.26)	(−0.45)	(−0.11)	(−0.17)
_cons	5.70**	6.88**	7.00**	6.86**	5.43**	6.76**
	(−0.66)	(−0.59)	(−1.32)	(−0.87)	(−0.57)	(−0.47)
R^2	0.5	0.47	0.56	0.48	0.45	0.45
N (vignettes)	684	650	313	391	1733	1824
N (respondents)	140	133	65	79	351	371

Cluster robust standard errors in parentheses. All models additionally control for set dummies.
$^{+}$ $p < 0.10$ * $p < 0.05$ ** $p < 0.01$

Table A3
Regression models for Figure 5

	18–30 year		31–59 years		60+ years	
	Within	Between	Within	Between	Within	Between
sex	−0.19	0.25	−0.06	0.18	−0.20	0.61^{+}
	(−0.22)	(−0.21)	(−0.16)	(−0.16)	(−0.25)	(−0.37)
age: 31-59	-	-	-	-	-	-
	-	-	-	-	-	-
age: 60^{+}	-	-	-	-	-	-
	-	-	-	-	-	-
educ: higher second.	−0.93	−0.35	−0.37	0.33	−0.61*	1.63**
	(−0.57)	(−0.46)	(−0.25)	(−0.24)	(−0.29)	(−0.45)
educ: tertiary	−0.55	−0.42	−0.25	0.55**	0.06	−0.07
	(−0.45)	(−0.30)	(−0.17)	(−0.20)	(−0.29)	(−0.24)
religious affiliation	−0.28	−0.03	0.00	−0.29	−0.13	−0.23
	(−0.32)	(−0.27)	(−0.17)	(−0.18)	(−0.33)	(−0.26)
childcare fee	0.89**	0.81**	1.03**	0.94**	0.85**	0.96**
	(−0.04)	(−0.04)	(−0.04)	(−0.03)	(−0.06)	(−0.06)
family status	−0.67**	−0.58**	−0.47**	−0.58**	−0.69**	−0.50**
	(−0.12)	(−0.12)	(−0.11)	(−0.10)	(−0.17)	(−0.19)
m_housewife	−0.16	−0.10	0.10	−0.13	−0.30	−0.24
	(−0.17)	(−0.15)	(−0.16)	(−0.15)	(−0.24)	(−0.25)
m_parttime	−0.12	−0.10	−0.11	0.06	−0.21	−0.57*
	(−0.18)	(−0.17)	(−0.15)	(−0.15)	(−0.27)	(−0.25)
m_fulltime	−0.31^{+}	0.05	−0.04	0.06	−0.41	0.10
	(−0.18)	(−0.17)	(−0.16)	(−0.14)	(−0.25)	(−0.25)
f_househusband	−0.26	0.05	−0.22	−0.08	−0.23	−0.31
	(−0.17)	(−0.17)	(−0.15)	(−0.14)	(−0.21)	(−0.24)
f_parttime	−0.05	0.01	−0.15	0.03	−0.37	0.10
	(−0.19)	(−0.15)	(−0.15)	(−0.14)	(−0.23)	(−0.25)
f_fulltime	−0.21	−0.16	−0.42**	−0.20	−0.83**	−0.62**
	(−0.16)	(−0.19)	(−0.15)	(−0.15)	(−0.27)	(−0.23)
grandparents	0.06	0.00	0.42**	0.05	0.17	0.08
	(−0.12)	(−0.12)	(−0.11)	(−0.09)	(−0.17)	(−0.16)
income	−0.50**	−0.51**	−0.62**	−0.65**	−0.50**	−0.60**
	(−0.04)	(−0.05)	(−0.04)	(−0.03)	(−0.06)	(−0.06)
from Germany	−0.10	0.29^{+}	−0.14	−0.14	0.11	0.18
	(−0.15)	(−0.16)	(−0.13)	(−0.12)	(−0.20)	(−0.25)
from abroad	0.02	0.11	0.01	0.10	−0.04	0.42^{+}
	(−0.15)	(−0.15)	(−0.14)	(−0.13)	(−0.21)	(−0.22)
rel_christian	−0.05	−0.43^{+}	−0.36**	−0.48*	0.11	0.12
	(−0.16)	(−0.26)	(−0.14)	(−0.19)	(−0.22)	(−0.29)
rel_muslim	−0.09	−0.70**	−0.10	−0.29	0.13	−0.12
	(−0.15)	(−0.25)	(−0.13)	(−0.19)	(−0.20)	(−0.32)
_cons	7.48**	6.74**	5.20**	7.10**	5.75**	6.72**
	(−0.72)	(−0.54)	(−0.92)	(−0.45)	(−0.63)	(−0.62)
R^2	0.52	0.46	0.56	0.55	0.54	0.63
N (vignettes)	908	985	1264	1392	558	488
N (respondents)	184	199	258	282	114	102

Cluster robust standard errors in parentheses. All models additionally control for set dummies.
$^{+}$ $p < 0.10$ * $p < 0.05$ ** $p < 0.01$

Table A4
Regression models for Figure 6

	No affiliation		Religious affiliation	
	Within	Between	Within	Between
sex	−0.23	0.00	−0.15	0.27*
	(0.25)	(−0.26)	(−0.13)	(−0.12)
age: 31-59	−0.03	0.18	0.2	0.17
	(−0.34)	(−0.33)	(−0.16)	(−0.14)
age: 60+	−0.19	0.61	0.07	0.27
	(−0.47)	(−0.41)	(−0.19)	(−0.19)
educ: higher second.	−0.59	1.08*	−0.43^{+}	0.28
	(−0.41)	(−0.54)	(−0.22)	(−0.21)
educ: tertiary	−0.69*	0.18	0.04	0.27^{+}
	(−0.28)	(−0.30)	(−0.15)	(−0.15)
religious affiliation	-	-	-	-
	-	-	-	-
childcare fee	0.94**	0.90**	0.96**	0.90**
	(−0.05)	(−0.04)	(−0.03)	(−0.03)
family status	−0.80**	−0.59**	−0.45**	−0.55**
	(−0.13)	(−0.13)	(−0.09)	(−0.08)
m_housewife	−0.23	−0.12	0.02	−0.15
	(−0.18)	(−0.20)	(−0.12)	(−0.11)
m_parttime	−0.19	−0.14	−0.12	−0.08
	(−0.20)	(−0.20)	(−0.12)	(−0.12)
m_fulltime	−0.38^{+}	0.26	−0.09	−0.01
	(−0.20)	(−0.21)	(−0.12)	(−0.11)
f_househusband	0.02	0.04	−0.35**	−0.14
	(−0.19)	(−0.19)	(−0.12)	(−0.11)
f_parttime	−0.09	0.09	−0.24*	−0.02
	(−0.20)	(−0.19)	(−0.12)	(−0.11)
f_fulltime	−0.02	−0.17	−0.64**	−0.29*
	(−0.21)	(−0.22)	(−0.12)	(−0.12)
grandparents	0.21	0.00	0.29**	0.05
	(−0.14)	(−0.13)	(−0.09)	(−0.08)
income	−0.57**	−0.73**	−0.54**	−0.55**
(-0.05)	(−0.04)	(−0.03)	(−0.03)	
from Germany	0.11	0.23	−0.14	0.01
	(−0.18)	(−0.16)	(−0.10)	(−0.11)
from abroad	0.18	0.37^{+}	−0.06	0.1
	(−0.18)	(−0.19)	(−0.10)	(−0.10)
rel_christian	−0.17	−0.90**	−0.16	−0.04
	(−0.17)	(−0.31)	(−0.11)	(−0.16)
rel_muslim	−0.14	−0.68*	−0.02	−0.20
	(−0.17)	(−0.30)	(−0.10)	(−0.16)
_cons	7.27**	7.14**	5.25**	5.91**
	(−0.91)	(−0.78)	(−0.72)	(−0.36)
R^2	0.53	0.55	0.52	0.49
N (vignettes)	830	811	1900	2054
N (respondents)	169	165	387	418

Cluster robust standard errors in parentheses. All models additionally control for set dummies.
$^{+}$ $p < 0.10$ * $p < 0.05$ ** $p < 0.01$