Scaling Level of Responses, Heaping and Censoring in Factorial Surveys: Expectations and Evidence in View of a Simple Cognitive Model

Volker Lang

(frame paper of the cumulative dissertation:

Lang, V. 2020. Response behavior in factorial survey experiments: Challenges and innovative solutions. Tübingen: University of Tübingen.

Link: https://rds-tue.ibs-bw.de/opac/RDSIndexrecord/1725134705)

This online publication contains the frame paper and two appendixes including the abstract (appendix 2) and the concluding remarks (appendix 3) of the cumulative dissertation mentioned above. A print version of this dissertation is available using the link shown above. In comparison to the print version internal references to chapters of the dissertation have been removed from this online publication of the frame paper, abstract and concluding remarks to make the document self-contained readable.

Abstract

The conceptual literature on factorial survey experiments assumes that ratings are continuous and interval scaled, and that response behavior in factorial surveys can be adequately described by an additive model. Alternatively, I hypothesize that response behavior in factorial surveys is guided by simple cognitive heuristics, and the structures of these heuristics lead to ratings which are not interval scaled and heaped at salient values of response scales. In this frame paper I introduce these two different conceptualizations of response behavior in factorial surveys and summarize findings to assess my hypothesis. In line with my expectations the studies in my dissertation show that non-interval scaled, heaped and censored ratings are common in factorial surveys. My results also show that respondents likely evaluate vignettes in a stepwise manner, and that they start off their evaluations with a focus on salient aspects of the experiments. Furthermore, I find that methods of analysis which do take non-interval scaled ratings and a stepwise evaluation process into account lead to more efficient parameter estimates.

1. Introduction

In recent decades, factorial survey experiments (FSEs, Jasso and Rossi 1977; Rossi and Anderson 1982) have become an increasingly widespread and successful method for measuring and analyzing attitudes, judgments, beliefs, opinions, preferences, behavioral intentions and decisions (Wallander 2009). An FSE is a type of survey experiment consisting of—typical textual—scenarios (called vignette scenarios) combining several treatments (called dimensions) with varying doses (called levels). The vignettes are used as stimuli which are singly evaluated by respondents (Auspurg and Hinz 2015). For a short introduction to FSEs see section 2.

Category rating instruments are commonly implemented in FSEs to record responses and variants of linear regression models are used to analyze the ratings (Hox et al. 1991). According to the review of Wallander (2009), 61 % of the FSEs conducted between 1982 and 2006 implemented category rating instruments (46 % used rating instruments with ordered and 15 % unordered categories) and 71 % of the studies conducted linear regressions or analyses of variance to examine the experimental data. Early papers on FSEs are relatively open about the types of response instruments and methods of analysis used. “The numerical form of .. judgment made may .. vary, from magnitude estimation methods .. to two or three point rating scales” (Rossi and Anderson 1982, p.31) and “one need not be completely fixed on the OLS [ordinary least squares regression, VL] formulation” (Rossi and Anderson 1982, p.33). Currently the use of close-ended ordered category rating instruments has become a textbook convention for FSEs and linear regression models (with clustered standard errors at the respondent level) are the analytic method of choice for these ratings. “For most applications, we recommend the use of .. rating scales with approximately 11 response categories” (Auspurg and Hinz 2015, p.69) and “the standard approach .. [to analyze

FSEs, VL] is to employ a multivariate linear regression model .. with the assumption that the measurement of outcomes is on a metric scale" (Auspurg and Hinz 2015, p.85).

The assumption of a metric outcome implies that the ratings are interval scaled. Therefore, respondents in a FSE have to cognitively project the discrete pieces of information contained in the vignettes on a single dimension in such a way that not only the ranking of their ratings, but also the distance between their ratings—the numerical difference between their evaluations of vignettes on a response scale—is informative. Given the findings of research on human cognitive heuristics (e.g., Gigerenzer and Todd 1999) the assumption that respondents cognitively process—i.e., perceive, evaluate and aggregate—information in such a sophisticated manner is strong. In how far the ratings in FSEs match the interval scaling assumption is a gap in research which is addressed in my dissertation (Lang 2018). Nevertheless, prior laboratory experiments on the scaling level of ratings for less complex tasks than FSEs—e.g., to assess the length of lines—have shown a relevant amount of not interval scaled responses (Orth 1982).

In addition to that, the heaping of ratings on salient points of the response scale like the midpoint or endpoints is common in FSEs (e.g., Sauer et al. 2009) and sometimes such heaping of evaluations leads to censored rating distributions (e.g., Auspurg and Gundert 2015). The rating distributions do not match the assumption of a metric outcome implied by using a linear additive model for such FSEs. Furthermore, the heaps of ratings are indicative that response behavior in these FSEs is not in line with the interval scaling assumption. Methods to analyze rating distributions with heaping and censoring are developed and applied in my dissertation (Groß and Lang 2018, Lang and Groß 2020, 2020a). In the literature on FSEs so far heaping and censoring have been treated as design problems which can be addressed by offering more flexible response instruments (e.g., magnitude estimation instruments, Jasso 2006)

or providing experimental treatments—i.e., vignette dimensions—with additional variation (Auspurg and Hinz 2015). The conceptual reasons for heaping are not largely discussed.

By contrast, I argue that non-interval scaled ratings, heaping and censoring in FSEs can be understood as consequences of response behavior which is guided by simple and cognitive plausible heuristics (Gigerenzer 2008) instead of a linear additive model representing a "social calculus" (Rossi and Anderson 1982). For example, if the vignette evaluation processes of respondents follow a fast-and-frugal classification heuristic (Martignon et al. 2003), this model of response behavior predicts that ratings in FSEs are—at least partly—not interval scaled and heaped at salient points of response scales (see section 3). Against this background, my dissertation addresses the question, as to how far the conceptual focus on (linear) additive models in combination with close-ended rating instruments is problematic for research using FSEs. The studies gathered in my dissertation provide evidence which suggests that response behavior in FSEs is driven by simple heuristics and develop methods of analyses to deal with the consequences of such response behavior in FSEs.

In the following, I will subsequently provide a short introduction to FSEs. Afterwards, I will introduce two conceptual models of response behavior in FSEs and derive related hypotheses about the scaling level, heaping and censoring of vignette evaluations. I will then continue with an overview of results regarding these hypotheses based on the findings of the publications in my dissertation as well as previous related studies. In the final section I discuss implications of the results for research using FSEs.

2. What is a factorial survey experiment?

A FSE is a type of research design which addresses two problems faced by social scientists interested in factors influencing people's attitudes, preferences or decisions:

first, the multicollinearity of explanatory factors in real world situations, and second, the collection of information on representative samples of respondents.

Regarding the multicollinearity problem one can consider the classical example of a researcher, who wants to know why clients buy certain types of cars (Louvriere et al. 2000). In data on the actual sales of cars, factors like motorization and interior quality are often strongly correlated with each other as well as with the price, which makes it hard to disentangle their relative relevance for the choices of consumers. Similarly, in survey or administrative data on employments, factors like an employee's gender, number of children and hours worked are often strongly correlated. This makes it difficult to assess the relative importance of such factors for justice attitudes on earnings (Auspurg and Hinz 2015). Such multicollinearity of explanatory factors is frequent in observational data used by social scientists (Rosi and Anderson 1982).

A person inherits **60.000.000€** from **her or his parents** and
the related inheritance tax amounts to **18.000.000€**.

The inheritance consits of **30.000.000€ operating capital,**
i.e., half of it is invested in a firm.

The firm has **50 employees** and
is situated in an **economically stable region**.

The heir has **worked in the firm for three years** and
has **agreed to preserve the jobs within the firm for at least 5 years**.

How just do you deem the raised inheritance tax amount?
11-category rating scale: -5: unjust, much too low; 0: just; 5: unjust, much too large

Figure 1. Translation of a vignette on attitudes towards inheritance taxation
Source: Groß and Lang (2018), Abbildung 1
Note: Dimensions are printed in bold font for didactical reasons.

In FSEs textual descriptions of situations, offers or objects to evaluate are presented to respondents to address this problem. These descriptions are called vignette scenarios or vignettes. Figure 1 shows an example vignette of the FSE on attitudes towards

inheritance taxation conducted as part of my dissertation (Groß and Lang 2018). In the vignettes the explanatory factors are controlled by the researcher and thus, can be varied independent of each other. Thus these factors are experimental treatments called dimensions, and different categories of the dimensions which represent varying doses of the treatments are called levels. Since vignettes commonly consist of several dimensions, a FSE is a type of multifactorial experiment. In Figure 1 the dimensions which vary between the vignettes are highlighted in bold font.

To remove multicollinearity between explanatory factors, vignette dimensions and their levels are either randomly combined (Jasso 2006) or deliberately composed to maximize their orthogonality and balance (Kuhfeld 2010). The later method is called "D-efficient sampling" and minimizes the correlation of dimensions more effectively compared to randomization (Dülmer 2007). Importantly, D-efficient sampling allows for the exclusion of unrealistic vignette scenarios—i.e., for restrictions on the possible combination of levels and even dimensions—while maintaining a minimal level of correlation between the dimensions. In consequence, contrary to critical appraisals regarding the realism of vignette scenarios (Faia 1980), FSEs do not have to comprise unrealistic vignettes to solve the multicollinearity problem (Auspurg et al. 2009).

In addition, embedding each treatment of a FSE in the multifactorial framework of a vignette scenario circumvents having to ask respondents directly about factors of interest. This indirect format is especially useful for research on sensitive topics where questioning respondents directly might lead to social desirable response behavior (for details on this aspect of FSEs see Auspurg et al. 2015, Liebig 2001, Liebig et al. 2015).

Regarding sampling, experiments in the social sciences are often conducted with volunteer or other convenience samples which carry the risk of introducing sample selectivity. Since it is not possible to assess heterogeneity in the treatment effects for

estimates based on convenience samples, those studies must assume that the treatment effects are homogenous within the population of interest (Imbens and Rubin 2015). FSEs—like other types of survey experiments—are able to address this problem by implementing the vignettes as part of a survey conducted with a probability sample. Such representative samples of a population allow the assessment of heterogeneities in treatment effects, by either estimating separate models for sub-groups (Jasso 2006) or by including interactions of vignette and respondent characteristics in the analyses.

A further defining feature of FSEs is that each vignette is singly evaluated by respondents. By contrast, experimental designs in which several scenarios are evaluated in comparison are either called "conjoint experiments" if the respondents have to rank a set of scenarios or "(discrete) choice experiments" if the respondents have to choose from a set of scenarios (Louvriere et al. 2000, Auspurg and Hinz 2015a). The fact that vignettes are singly evaluated does not imply that each respondent evaluates only one vignette—which can be done (e.g., Jann 2003). Vignettes are composed in a set called "vignette deck" and these decks are then randomly assigned to respondents, instead of randomly assigning the single vignettes (Auspurg and Hinz 2015). Furthermore, vignettes or decks that are randomly assigned to respondents are essential for FSEs since the randomization ensures independence between characteristics of respondents and vignette dimensions (i.e., experimental treatments).

Mostly, the vignette evaluations of respondents in FSEs are measured using close-ended category rating instruments (Wallander 2009). Sometimes measurements are recorded with more detailed response scales like slider, magnitude estimation or number matching instruments. While the assumption that the ratings are interval scaled is not necessarily implied by implementing such response instruments, category ratings are expected to be interval scaled. This is the case as it is supposed that respondents use

the rating scales to quantify their perceptions of differences between stimuli (Orth and Wegener 1983). Pertaining to FSEs, these stimulus differences are represented by different vignettes. Furthermore, assuming interval scaled ratings is part of the conceptual fundamentals of FSEs (Rossi and Anderson 1982) and it is required by the methods of analysis commonly used for FSEs (Hox et al. 1991). For a more comprehensive introduction to FSEs see Auspurg and Hinz (2015).

3. Interval Scaled Responses and Human Cognition: a Demanding Relationship

In this section I will examine the plausibility of assuming interval scaled ratings in FSEs. Therefore, I will first introduce the standard conceptual model of response behavior in FSEs and an alternative model based on simple heuristics, which is more in line with the current literature on human cognitive processing. Afterwards, I will discuss the implications of this alternative model of response behavior in context with the scaling, heaping and censoring of ratings in FSEs, and formulate related hypotheses.

3.1. Basic Models of Response Behavior in Factorial Surveys

Formally, response behavior in FSEs is commonly conceptualized in terms of a (linear) additive model (Rossi and Anderson 1982; Auspurg and Hinz 2015):

$$y^v = \alpha + X^v\beta + \varepsilon^v \quad (1)$$

y^v is the assessment of a vignette scenario—i.e., the vignette rating, α is a fixed intercept parameter, X^v is a matrix of variables representing the experimental factors—i.e., the dimensions of the vignettes, β is a vector of the respective fixed parameters and ε^v is an error component representing deviations from the average ratings given by α and β.[1]

[1] Here, I focus on the vignette dimensions as explanatory factors and leave the distinction between vignette variables, which are controlled by the experimenter, and the respondent level variables, which are only observed, as well as the differentiation between respondent and vignette level error aside since these differences are not essential for my argument.

In their introductory paper on FSEs, Rossi and Anderson (1982) motivate the use of an additive equation like (1) as a model of a "social calculus". "Under the assumption that only individual deviations from the social calculus are involved, we can view the expression for a judgment as being comprising two parts, one part a function of the [experimental, VL] characteristics and the socially agreed-upon weights attached to the characteristics and one part representing individual deviations from that consensus. .. We will define the "social components" of judgments .. by pooling the judgments made by individuals in a population" (Rossi and Anderson 1982, p.21). Importantly, the experimental design accounts for the multicollinearity of factors common for observable data and real world situations, and thus, FSEs enable the reconstruction of the "social components" of evaluations (see section 2).[2]

This conceptualization of response behavior in FSEs being guided by an model as illustrated in (1) appeals to most social scientists—and especially sociologists—for two reasons: the first being that the premise of socially structured human behavior and evaluations which is constitutive for their research; the second being that the idea of a "social calculus" circumvents the inconvenient assumption that individual cognitive processes match a model similar to (1). As long as the mismatch between (1) and individual cognitive processes is idiosyncratic, estimates of the "social components" are not biased. However, that such a mismatch is idiosyncratic is a strong assumption given the evidence of systematic deviations from an additive model in human decisions and evaluations (e.g., Gigerenzer 2008, Tversky and Kahnemann 1974, Simon 1955).

The more recent discussion of the conceptual foundations of FSEs by Jasso (2006) mentions the possibility, that the cognitive heuristics of respondents could differ

[2] Regarding an individual cognitive model similar to an equation like (1) Rossi and Anderson (1982) only state: "There is some evidence that persons use an additive principle in arriving at the overall judgment of complex objects (Anderson, 1974)" (Rossi and Anderson 1982, p.21).

from an additive model representing the social aggregation of evaluations. More specifically, she remarks first, that cognitive models "may be more faithfully represented by a multi equation system, say, or a tree structure" (Jasso 2006, p.335f) and second, that these models could be comparably "shorter", i.e., less complex than a model like (1), "if (some) individuals pay attention to a restricted set of stimuli (Kahneman and Tversky 1979; Miller 1956)" (Jasso 2006, p.336). Nevertheless, she does not discuss possible consequences of such deviations from an additive model for analyzing human evaluations and response behavior in FSEs.

Furthermore, the fit of response behavior with an additive model like (1) may depend on the importance of the evaluation situation. "Very trivial choices are made "automatically"; even the less trivial .. may be settled .. by .. decisions which have become habits. .. The choices that surface as deliberate weighing and evaluation of alternatives tend to have more important consequences for the future .. and .. may involve explicit weighing of the positive and negative aspects of each alternative" (Rossi and Anderson 1982, p.18). Thus, it is supposed that evaluation modes differ between so called "high cost" or "high stake" situations in which the cognitive processing of humans is hypothesized to more closely match an additive model like (1), and so called "low cost" or "low stake" situations, in which humans more often stick to simple heuristics and norms (Esser 1993, Kroneberg 2005). Many FSEs focus on topics which in real life are best characterized as high cost situations like the acceptance of job offers (Abraham et al. 2013), the selection of job applicants by employers (de Wolf and van der Felden 2001), fair earnings (Jasso and Webster 1999), residential preferences (Emerson et al. 2001) or attitudes towards the death penalty (Boots et al. 2003). In such situations an evaluation behavior more in line with a model as illustrated in (1) is expected.

Nevertheless, vignette scenarios are standardized and simplified representations of these real life situations, and while they are intended to stimulate evaluation processes similar to the represented real life situations vignette evaluations are clearly less consequential. Importantly, this difference between real life situations and vignette scenarios does not imply that the evaluation rules found by vignette studies are a priori false (Faia 1980). If and in how far they differ from real live situations is an area of ongoing research (Petzold and Wolbring 2018, 2019). However, even if the real life situations represented by vignettes are "high cost" the vignette evaluation situation is better characterized as "low cost"—like almost all survey questions. Thus, a response behavior which is influenced by habits and norms and guided by simple heuristics needs to be expected. In addition, the relevance of the differentiation between high and low cost situations for understanding human evaluation behavior is controversial, as a growing number of scientists argue that the use of simple heuristics is not restricted to specific situations but part of the human condition (Gigerenzer and Todd 1999).

Whether human evaluations always follow simple cognitive heuristics or follow them only in low cost situations like survey response, it is unlikely that such heuristics involve the quantification of perceived stimulus differences. As a consequence, the ratings in FSEs are most likely not interval scaled, and the response behavior is not in line with an additive model like (1). Instead, respondents who follow simple cognitive heuristics classify stimuli into categories matching their perceptions of differences (Gigerenzer 2008). Ratings based on such an evaluation process represent a classification not a quantification of stimulus differences. Based on the framing of many evaluation tasks in FSEs, this classification would probably be ordered, but the numerical differences between the ratings would not be meaningful. Related studies of the scaling level of ratings in other applications than FSEs show variation between

individuals and experimental designs (Wegener 1982). These findings indicate that there is an interval and a non-interval scaled evaluation mode in empirical applications.

To give a specific example, when faced with the task to evaluate how just they consider the earnings in a vignette scenario, a simple heuristic describing respondents cognitive processing could be a fast-and-frugal classification tree (Martignon et al. 2003) like the one shown in Figure 2 instead of an additive model like (1). [3] Using this classification tree respondents consider earnings just, if they are no larger than 5,000 € per month or, if they are larger than 5,000 € per month and the person described in the vignette is a physician. They only consider the vignette in greater detail if these conditions are not met. Moreover, an additional classification tree that uses, for example, less than 1,500 € per month and hairdresser as alternative cues could further restrict the range of vignette scenarios respondents assess more detailed.

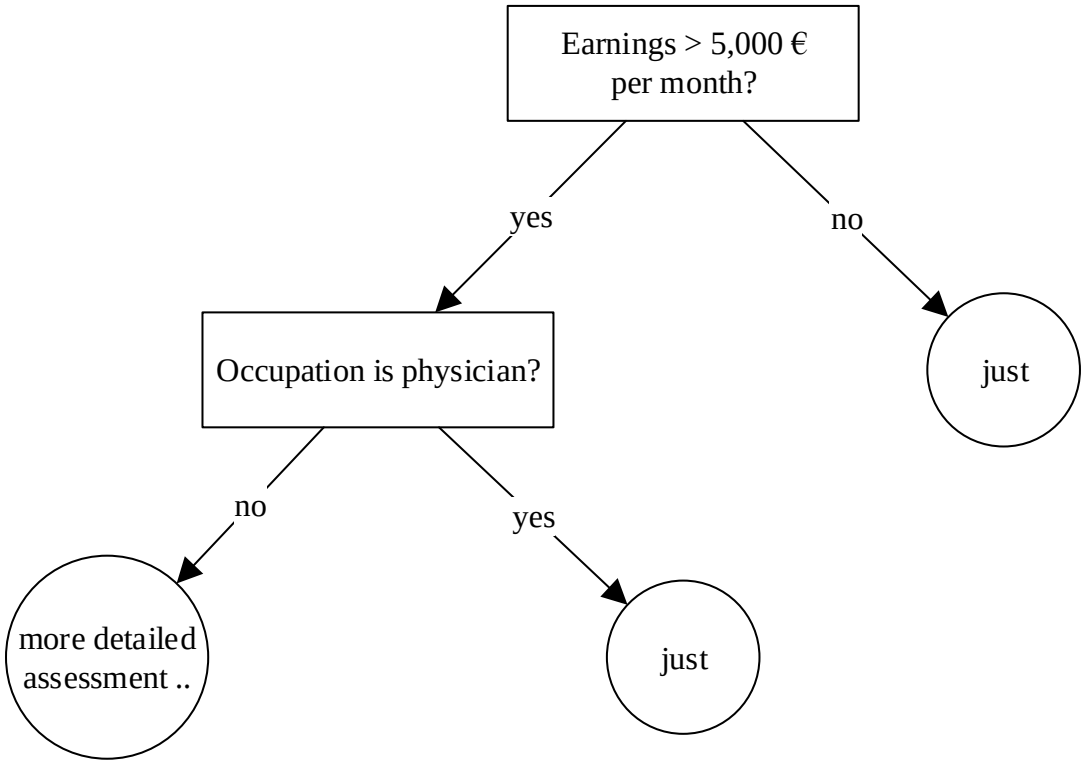

Figure 2. A fast-and-frugal classification tree to evaluate vignette earnings

[3] A classification tree is called fast-and-frugal if it consist of x cues and x+1 exits or endpoints. After the last cue there are two exits and there is one exit after each previous cue (Martignon et al. 2003). In Figure 2 the cues are levels of the vignette dimensions earnings and occupation.

Figure 2 only presents one plausible scenario, and the cues as well as the heuristic itself could differ. However, the example illustrates the general point that plausible cognitive heuristics most likely do not involve the quantification of stimulus differences. Given this alternative model of response behavior, I will subsequently discuss related expectations about the scaling, heaping, and censoring of ratings in FSEs.

3.2. Interval Scaled Responses, Heaping and Censoring in Factorial Surveys

In the first two steps of the simple heuristic described in Figure 2 respondents only produce a classification into just and unjust scenarios.[4] Using additional cues respondents likely differentiate between "unjustly too high" and "unjustly too low" scenarios which results in an ordered classification of scenarios into three categories: "unjustly too low", "just", and "unjustly too high". Since a quantification of the amount of perceived injustice is not part of the heuristic respondents use, the distance, i.e., the numerical difference on a rating scale between the vignette scenarios they consider "just", "unjustly too low" or "unjustly too high" is not informative. As a consequence, the ratings classified into these three categories are not interval scaled. After the second evaluation step, it depends on the formulation of the following further detailed evaluation steps of the heuristic in Figure 2 if the ratings within the categories of vignettes classified as unjust are interval scaled. In as much as these further evaluation steps involve the quantification of differences in perceived injustice, some of distances between ratings within this category could be informative. Therefore, the scaling level of ratings is contingent on the arrangement of the heuristics guiding response behavior.

[4] For reasons of clarity, I sometimes refer to justness ratings used in many FSEs on earnings justice attitudes and the related terms just and unjust in deriving my hypotheses. However, the described principles are applicable to FSEs on other topics. For example, in case of a FSE on students' internship preferences where participants have to evaluate the attractiveness of internships (Lang 2018) one would use the terms "indifferent" and "(not) attractive" instead of "just" and "unjust" as well as a different heuristic, but the arguments would be the same.

Nevertheless, as long as respondents follow simple classification heuristics without a quantification of perceived stimulus differences, their ratings will not match the interval scaling assumption. Based on these considerations I formulate my first hypothesis:

> *H1:* If respondents follow a simple classification heuristic, their evaluations are not interval scaled.

Furthermore, in a classification heuristic such as is described in Figure 2, the just ratings are not dealt with in later steps of the evaluation. Depending on the cues used, these just ratings can amount to a relevant share of all evaluations. As a consequence, given such a heuristic the midpoints of response scales are clear candidates for heaping. In addition, if the later evaluation steps of the heuristic also involve no quantification of perceived differences in injustice, the resulting ratings are also only roughly classified. Given a differentiated response instruments such a heuristic would lead to additional heaping of ratings at salient points of the response scales, e.g., the endpoints or other prominent values. These heaps would represent the coarse cognitive categories. This expectation regarding the use of classification heuristics is also in line with the observation that survey respondents have a tendency to use the whole range of any response scale (Tourangeau et al. 2000). This leads to my second hypothesis:

> *H2:* If respondents follow a simple classification heuristic, the resulting distribution of evaluations is heaped at salient points of the response scale (e.g., the midpoint, the endpoints or other prominent values).

Insofar the use of simple heuristics like the one describe in Figure 1 is part of the human condition (Gigerenzer and Todd 1999)—or at least part of the human condition in low cost situations—a high prevalence of respondents following simple heuristics in FSEs or survey response in general can be expected. In combination with H1 and H2, this expectation would imply a high prevalence of non-interval scaled and heaped ratings.

In addition, response behavior following a simple heuristics representing a norm could explain the censored rating distributions found in some FSEs, especially those FSEs focusing on controversial topics. If a norm defines an unconditional opinion towards a topic and a relevant share of respondents follows a heuristic based on this norm, then such an unconditional "social consensus" among a subpopulation could lead to a heaping of ratings on or close to an endpoint or extreme value of the response scale. For example, in Germany a considerable share of the population strictly opposes inheritance taxes (Schrenker and Wegener 2007). Insofar these persons adhere to the norm that inheritance taxation is extremely unjust, they will always strictly oppose such taxes. A heuristic guiding response behavior based on such a norm would result in a censored distribution of vignette ratings in related FSEs.

Moreover, the degree to which the use of simple heuristics is "hardwired" in humans—at least in low cost situations—also restricts the possibilities to influence the scaling level of responses, as well as the amount of heaping and censoring through experimental design. More specifically, close-ended rating instruments carry the risk of restricting how respondents can express the distances between ratings correctly (Jasso 2006, Rossi and Anderson 1982). Ratings heaped at the endpoints of scales are indicative for this censoring of response behavior (Tourangeau et al. 2000). However, if a heuristic only uses a limited number of categories to classify evaluations, offering a response instrument with more categories or a more continuous response scale, e.g., a slider or magnitude estimation instrument, will not lead to more interval scaled ratings or less heaping. In consequence, to which extend the design of response instruments can affect the scaling and heaping of ratings depends on how commonly respondents follow simple cognitive heuristics. Given these considerations I derive my third hypothesis:

H3: If respondents follow a simple classification heuristic, implementing fine grained response instruments (e.g., sliders or magnitude estimation instruments) will not increase the prevalence of interval scaled ratings and decrease heaping.

A mechanism like the one described by H3 could also explain why close-ended rating instruments with a limited number of categories in practice tend to work better compared to magnitude estimation instruments (Auspurg and Hinz 2015, Markovsky and Eriksson 2012) although the latter are conceptually more suitable for metric measurement (Jasso 2006).

The implications of following simple heuristics for the effects of person related factors—like motivation, general cognitive abilities or topic specific knowledge—on the scaling and heaping of rating are less clear. A growing number of studies show that simple heuristics can beat more complex algorithms not only with respect to efficiency in terms of (cognitive) resources, but also regarding the effectiveness to solve evaluation and decision problems (Gigerenzer 2008). Given these findings, it is not plausible to postulate a positive association between the complexity of the cognitive models respondents follow and their abilities and motivation. A major function of simple heuristics is to avoid cognitive load (Gigerenzer and Todd 1999) and respondents with more cognitive skills, motivation or familiarity with a topic probably have a higher threshold to feel cognitive strain. In addition, the heuristics used by such respondents may be more efficient in avoiding cognitive load in the first place. Therefore, they could be more effective in solving the evaluation tasks in FSEs by applying cues which better discriminate between vignette scenarios.

3.3. Consequences for Parameter Estimates in Factorial Surveys

In the following, I discuss the consequences of not interval scaled ratings, heaping and censoring for estimates of parameters in FSEs. I argue, that if response behavior in a

FSE is guided by heuristics as described in Figure 2 and if an additive model like (1) is used to analyze such a FSE, the estimates of the "social components" of evaluations, i.e., the estimates of vignette effects are less efficient and potentially also biased.

First, regarding the efficiency of the estimates, I postulate—in line with previous findings for other applications than FSEs (Wegener 1982)—that there is an interval scaled and a non-interval scaled response mode in FSEs. For the later type of response behavior the distances between ratings are not informative (see sub-section 3.2.). In consequence, a linear additive model which is used to analyze these ratings is misspecified. Therefore, the parameter estimates of such a model will be less efficient, i.e., the standard errors will be larger and the z-values smaller. Furthermore, if an ordinal model is used to analyze the ratings instead, such a model is also misspecified, and hence, less efficient, as it ignores the information contained in the distances between ratings in the group of interval scaled responses. Given that the ratings in a FSE are partly interval and partly non-interval scaled, I expect that the solution yielding the most efficient estimates, is to measure the scaling level of ratings with a test and then to partition the sample into two groups, one group of interval scaled and one group of non-interval scaled ratings. Afterwards, a model should be implemented which treats the ratings in each of these two groups adequately based on their scaling level (Lang 2018). This leads to my fourth hypothesis:

> *H4:* If the response behavior in a FSE matches the interval scaling assumption for some, but not for all ratings, a model which processes interval and non-interval scaled ratings adequately will provide the most efficient parameter estimates.

A derivative of this hypothesis is to expect more statistical significant parameters in such models compared to standard methods of analysis for FSEs. Similar to non-interval scaled ratings, heaping and censoring attenuate the efficiency of parameter estimates in

FSEs. In such cases, methods that model the process leading to heaping and censoring (Lang and Groß 2020a) will provide more efficient estimates in comparison to linear additive models which are designed for continuous rating distributions without heaps.

Furthermore, assuming respondents follow a simple classification heuristic expectations regarding bias in parameter estimates based on standard additive models like (1) can be derived. Studies on evaluation heuristics show that subjects tend to focus on salient aspects because they recognize these aspects faster and a related evaluation rule which prescribes a focus on features with more recognition potential is called fluency heuristic (Schooler and Hertwig 2005). For FSEs, it is also plausible to expect that respondents tend to focus on salient aspects insofar response behavior is guided by a fluency heuristic. The more salient aspects of a FSE are the vignette dimensions which show more variation across scenarios (Ausprug et al. 2009). As an example, consider again the fast-and-frugal classification tree presented in Figure 2 as it is a type of fluency heuristic. The amount of earnings and the occupation persons work in are commonly considered important determinants for related justice attitudes. Therefore, earnings and occupation are plausible cues used by respondents and often the salient aspects of FSEs on earning justice attitudes, i.e., the dimensions with the most variation.

We can expect to find stronger effects of earnings and occupation on ratings in the earlier evaluation steps which focus on these two cues, inasmuch as response behavior in this FSE is guided by a fluency heuristic. To identify these comparably stronger effects a model which differentiates between the evaluation steps of a heuristic is needed. In such a case, the parameter estimates of an additive model as in (1) will be biased, since such a model supposes that there are no different evaluation steps. Instead a model such as (1) assumes that the information of all vignette dimensions is

simultaneously considered and weighted against each other (see sub-section 3.1.). Based on these considerations I formulate my final hypothesis:

> *H5:* If respondents follow a simple classification heuristic which begins with focusing on salient dimensions of the vignettes, the effects of these dimensions are larger at earlier evaluation steps in a model which differentiates these steps.

4. Results

4.1. Findings on the Scaling of Responses, Heaping and Censoring in FSEs

First, I present results regarding the scaling level of ratings in FSEs. Lang (2018) is the first study testing the scaling level of responses in a FSE. For each respondent, the test is based on comparing a ranking of distances between ratings of single vignettes with a ranking for ratings of distances between pairs of vignettes. This test setup is a type of conjoint measurement which is called paired difference rating design (for further details see Lang 2018). I developed this design based on previous studies testing the scaling level of ratings for concepts with one dimension (e.g., the length of lines). The test results in two measures for the scaling level of ratings: first, the Pearson's correlation r and, second, the rank correlation Kendall's τ_b. Conceptually, both correlations range from -1 to 1. A value of one indicates perfectly interval scaled ratings and values between zero and one point to a mixture of ordinal and interval scaled ratings. Values closer to one suggest more interval and values closer to zero indicate more ordinal scaled ratings. By contrast, values smaller than zero mean that the ratings are not ordinal, but rather nominal scaled.

Figure 2 shows the cumulative distribution functions of r and τ_b over the 225 tertiary students for which the scaling level of their vignette ratings was tested in Lang (2018). The mean of r is 0.45 and the mean of τ_b is 0.34, whilst the correlation between r and τ_b is 0.83. The means of r and τ_b are substantially smaller than one, which clearly

indicates violations of the interval scaling assumption. The ratings of none of the respondents are perfectly interval scaled. The maximum value of r and τ_b attained by one respondent is 0.96. For 11 % of the sample r (24 cases) and τ_b (25 cases) are smaller than zero—demonstrating that these ratings are only nominal scaled.

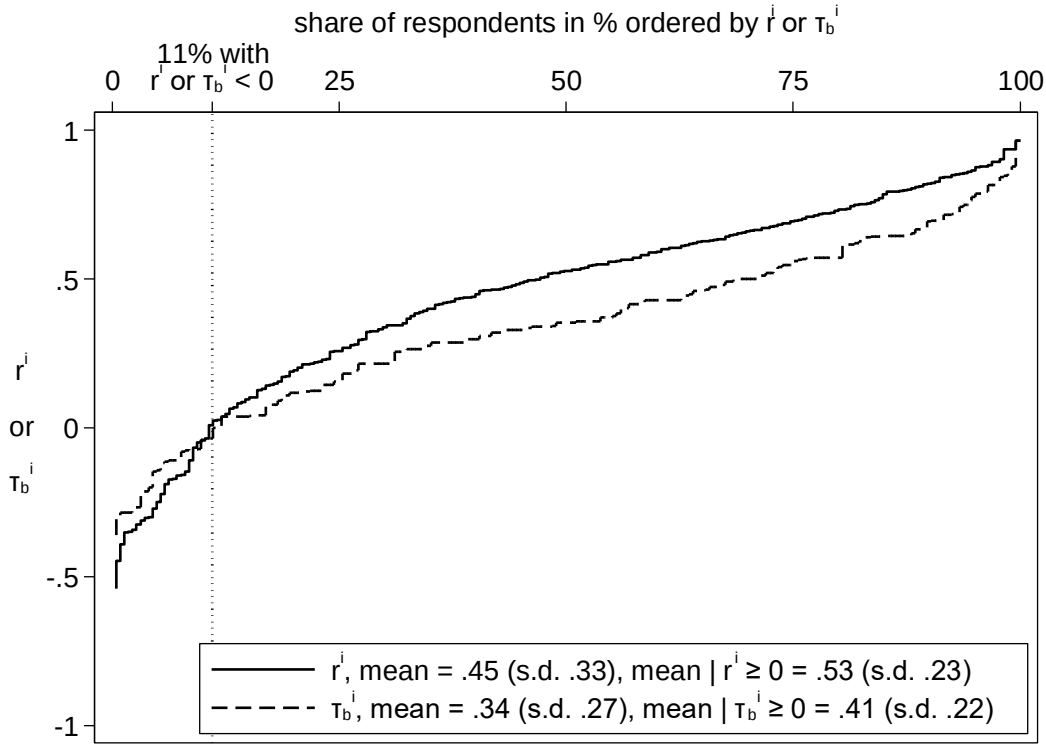

Figure 2. CDFs of the interval scaling level indicated by r^i or τ_b^i in the sample.
Source: Lang (2018), Figure 2 ($N_{respondents}$ = 225)

Overall, the findings are in line with H1, the expectation that there is a substantial share of non-interval scaled rating in FSEs, as response behavior is guided by simple heuristics. The high prevalence of non-interval scaled ratings in the sample, also supports the idea, that the use of simple heuristics in evaluation tasks is widespread or even part of the human condition (Gigerenzer 2008). In a standard FSE design, such a mismatch of ratings with the interval scaling assumption would go unnoticed.

Second, I look at the evidence regarding heaping and censoring. A response instrument which differentiates between steps of the rating process was implemented in the FSE on earnings justice attitudes included in the SOEP-Pretest 2008 (Sauer et al.

2009). The data of this FSE is also used in my dissertation (Lang and Groß 2020). In a first step respondents have to classify the earnings stated in vignettes as "just" or "unjust", in a second step as "unjustly too low" or "unjustly too high". In a third step they must quantify the amount of injustice they perceive with a natural number between 1 and 100.[5] While comparably complex, this instrument in principle gives respondents the possibility to express their evaluations very detailed.

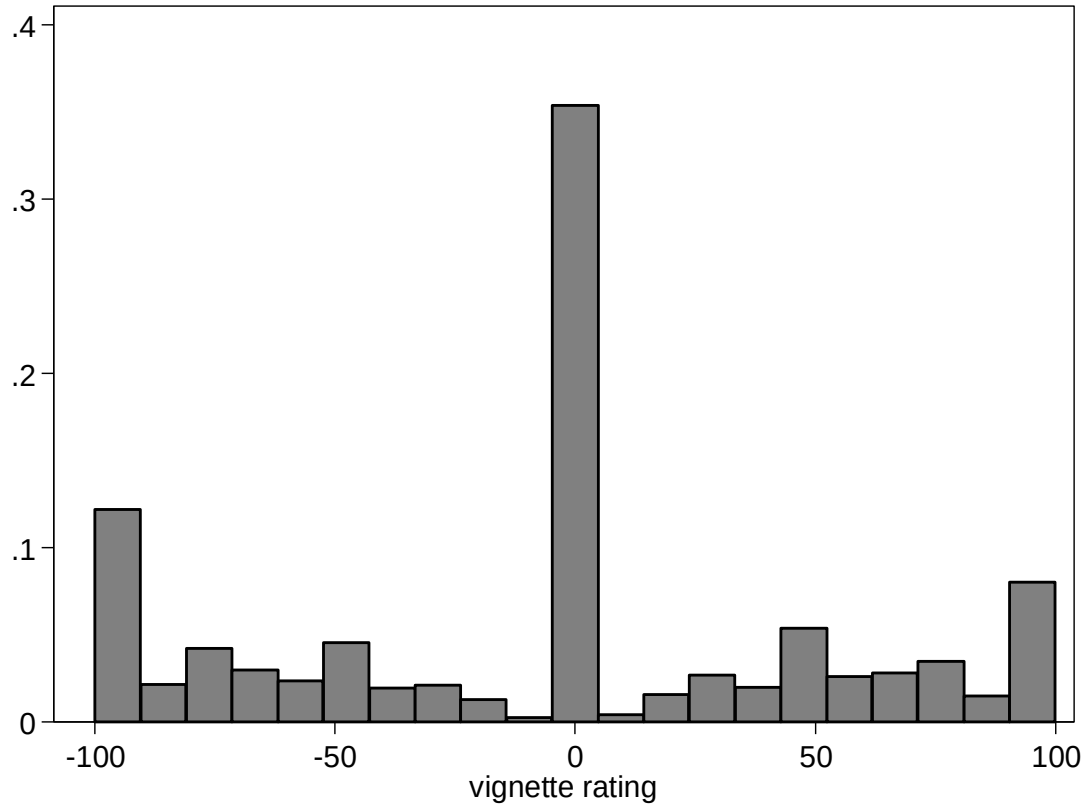

Figure 3. Distribution of vignette ratings in the SOEP-Pretest 2008
Source: Lang and Groß (2020), Figure 1 ($N_{ratings}$ = 26,650)

Figure 3 displays the rating distribution of this FSE. The mean percentage of vignettes rated just is 35 and the mode of just ratings per respondent is 7 out of 25 (28 %). 4 respondents consider all of the 25 vignettes just. A heap of about 20 % of the ratings on the midpoint of the response scale is common in FSEs on earnings justice attitudes

[5] In contrast to typical magnitude response instruments (Jasso 2006), in which respondents also use numbers to express the amount of injustice, the response instrument used in the SOEP-Pretest 2008 predefines the range and type of numbers used. Thus, the scale of this instrument is a priori fixed while magnitude response instruments have to use an anchor vignette instead.

using close-ended category rating instruments (Auspurg et al. 2009a). In comparison there are clearly more just ratings in the SOEP-Pretest 2008.[6]

With respect to heaps at the extreme values of the rating distribution—which point to potential censoring—12 % of the vignettes are rated as -100 ("unjustly much too low") and 8 % of the vignettes are rated as 100 ("unjustly much too high"). 30 % of the respondents do not use the extreme values at all; the median of such ratings per respondent is 3 out of 25 (12 %) and the mean 5 (20 %).[7] The amount of ratings on the endpoints of the response scale is comparable to other FSEs on earnings justice attitudes (Auspurg and Hinz 2015) and also to FSEs on other topics (e.g., Lang 2018).

An additional expectation regarding response behavior in FSEs following simple heuristics was, that if such heuristics represents a norm that prescribes an unconditional opinion towards a topic, this could explain a heaping of ratings on extreme values of the response scales (see sub-section 3.2.). The FSE on attitudes towards inheritance taxation in my dissertation (Groß and Lang 2018) represents a topic for which it is known that a substantial share of persons holds an unconditional opposed opinion. In this FSE 649 of 1930 ratings (34 %) were heaped on the endpoint of the response scale indicating that the tax proposed in the vignette was much too high.[8] 22 % of the respondents considered the tax much too high in at least four of the five vignettes they evaluated. This group of respondents accounts for 62 % of the censored ratings.

[6] Possible explanations are first a stronger focus on "just" ratings induced by the first step of the implemented response instrument or second that respondents use "just" ratings in order to skip the later steps of the instrument and speed up the evaluation task. However, the finding that the percentage of "just" ratings is not larger for vignettes rated later in the FSE contradicts the latter "speed up"-explanation (Sauer et al. 2009).

[7] Furthermore, previous analyses of the same FSE found that 65 % of the respondents only use values with tens digits—like 10, 20, 30 and so forth—for their evaluations (Sauer et al. 2009). On average respondents only use 8.5 different values for their ratings in the SOEP-Pretest 2008.

[8] A heap at the tail of the rating distribution pointing towards a too high taxation is also found in another FSE on this topic (Gross et al. 2017). Similar heaps of ratings are present in FSEs on other topics where a consensus of a subgroup on an unconditional opinion is plausible (Auspurg and Gundert 2015, Boots et al. 2003).

Taken together, the findings support H2, the expectation that ratings are heaped at salient points of the response scale, because response behavior is based on simple heuristics. In the FSEs analyzed in my dissertation—as well as in other studies using FSEs—a relevant share of ratings is heaped at the midpoint and endpoints of the response scale. The results also show that a relevant share of respondents following a simple cognitive heuristic which describes an unconditional opinion on a topic can lead to a censored rating distribution, i.e., a heap of responses on an endpoint the scale.

4.2. Factors influencing the Scaling of Responses, Heaping and Censoring in FSEs

In this section I analyze which factors have effects on the scaling level, heaping and censoring of ratings in FSEs. First, I assess factors influence on the scaling level indicators r and τ_b (see sub-section 3.1.) based on the FSE regarding student's internship preferences conducted in my dissertation (Lang 2018). The design of this FSE contains a split differentiating between an 11-category, a 21-category and a slider rating instrument. This design split can be used to test H3. In addition, the ordering of the single and paired vignette rating tasks was exchanged for a third of the respondents assigned to the 21-category or the slider rating instrument.

In addition, I use the following measures to evaluate influences of respondent's characteristics on the scaling of their ratings: an indicator variable for being a foreign student or having a first generation migration background, because of its association with less German language skills, and z-standardized sum indexes for the constructs expressiveness, need for cognition and conscientiousness based on five, four and two items (Rammstedt and Beierlein 2014). Expressiveness is a facet of the Big Five-aspect extraversion and describes the willingness to convey information. Need for cognition measures thoughtful and conscientiousness diligent response behavior.

Table 1 shows the results of regressions of r and τ_b on design features of the FSE and respondent characteristics. The upper part of Table 1 displays the effects related to the experimental design. Here, I find no relevant influences on r and τ_b for the 21-category or the slider compared to the 11-category rating instrument.[9] Moreover, the amounts of heaping and censoring in the design splits using the 21- or the slider compared to the 11-category rating instrument are not lower (Lang 2018, Figure A3). Taken together, these findings support H3. FSE designs which implement fine grained rating instruments neither show more interval scaled response behavior, nor less heaping and thus, point to the use of simple cognitive heuristics by respondents.

Table 1. Regressions of r and τ_b on factors influencing the scaling level of ratings

	Pearson's r		Kendall's τ_b	
	b (r.s.e.)	t-value	b (r.s.e.)	t-value
study design aspects:				
21-category rating instrument[a]	.07 (.06)	1.16	.10 (.06)	1.69†
slider instrument	-.05 (.07)	-.73	~0 (.06)	-.04
vignette pairings 1st X 21-cat. instrument	-.22 (.07)	-3.12**	-.18 (.05)	-3.23**
vignette pairings 1st X slider instrument	-.02 (.08)	-.25	-.04 (.06)	-.67
respondent characteristics:				
foreign student or 1st generation migrant[b]	-.28 (.12)	-2.36*	-.14 (.07)	-2.08*
expressiveness	.08 (.03)	2.50*	.07 (.03)	2.71**
need for cognition	.05 (.02)	2.06*	.04 (.02)	1.88†
conscientiousness	.02 (.02)	1.08	.03 (.02)	1.85†
constant	.45 (.08)	5.39**	.32 (.07)	4.51**
R^2 in %	18.5		18.3	

Note: Vignette deck and text order indicators are included as control variables.
Reference: [a]11-category rating instrument; [b]no foreign student or 1st generation migrant
Legend: **: P(T>|t|) < .01; *: P(T>|t|) < .05; †: P(T>|t|) < .10
Source: Own calculations based on Lang (2018) ($N_{respondents}$ = 225)

The lower part of Table 1 describes the effects of respondent characteristics on r and τ_b. Here, r and τ_b are significantly lower for foreign students or students with a first

[9] This finding is in agreement with results showing that respondents only use a limited number of response categories in FSEs (Sauer et al. 2009) as well as with previous experiments which found no substantial effects of instruments offering more response options on evaluations (Orth 1982, Wegener 1983). In addition, I find task order effects for the 21-category rating instrument in the paired difference rating design of the implemented scaling level test. If vignette pairings are rated first—which is the more complex part of the two-part rating task—r and τ_b are significantly lower. In comparison, there is no task order effect for the slider.

generation migration background. This finding indicates that respondents with lower language skills have more difficulties to produce interval scaled ratings. The analysis also shows substantial positive effects of expressiveness on r and τ_b. Thus, ratings more often match the interval scaling assumption if raters are more willing to convey information. Furthermore, I find a significant positive effect of the need for cognition on r and close to significant positive effects of the need for cognition and conscientiousness on τ_b. These results point to positive correlations between a thoughtful as well as diligent personality and interval scaled response behavior.[10]

Next, I look at effects of respondent's characteristics on heaping and censoring. This analysis is based on the SOEP-Pretest 2008 which is also used in my dissertation (Lang and Groß 2020). With respect to socio-demographic indicators, I look at respondent's gender, age, their education in years as, and whether German is their native language.[11] Additionally, I incorporate z-standardized sum indexes for the importance of earnings comparisons based on nine items, earnings equality preferences based on two items and for the personality facets agreeableness, conscientiousness and neuroticism based on two items each (Siegel et al. 2009). The willingness to participate in the FSE is measured by a z-standardized interviewer rating.[12]

[10] In addition to the heterogeneities in response behavior between respondents shown in Table 1, I tested the influences of university entrance diploma grades and parental education as well as occupation on r and τ_b. I find no significant effects for these indicators. A reason for these null-findings could be that tertiary students are a positively selected group with less variance in secondary school performance and social background compared to the general population.

[11] The specification for age, which includes a squared term, and the indicator for being a woman are correlated with labor force participation. Due to this multicollinearity, I use no variables related to respondent's employment situation in the presented specification. Models including indicators for the employment status and earnings of respondents lead to similar conclusions than the models in Table 2, but have less explanatory power.

[12] The variables for age and education are centered and age is scaled in 10 years. Descriptive statistics for all indicators are shown in Table A1 in the annex. Correlations of the explanatory variables are not substantial, the strongest is -0.3 between conscientiousness and neuroticism.

Table 2. Regressions of percentage of just ratings

	b (r.s.e.)	t-value	b (r.s.e.)	t-value
socio-demographic indicators:				
age in 10 years	-7.7 (1.6)	-4.82**	-7.0 (1.6)	-4.48**
age in 10 years squared	.7 (0.2)	4.48**	.7 (0.2)	4.29**
preferences and personality aspects:				
importance of earnings comparisons	.	.	2.2 (0.6)	3.89**
earnings equality preference	.	.	-3.2 (0.6)	-5.60**
conscientiousness	.	.	-2.7 (0.6)	-4.57**
neuroticism	.	.	-1.4 (0.5)	2.66**
constant	30.9 (1.6)	19.30**	30.4 (1.5)	19.79**
R^2 in %	7.0		14.6	

Note: Vignette deck indicators are included as control variables.

Legend: **: P(T>|t|) < .01; *: P(T>|t|) < .05; †: P(T>|t|) < .10

Source: Own calculations based on SOEP-Pretest 2008 ($N_{respondents}$ = 1,025)

Table 2 presents the analysis of factors influencing the percentage of just ratings heaped on the midpoint of the scale. The amount of just ratings is lower in the middle age range—in the age group which tends to be strongly integrated into the labor market. The percentage of just ratings is significantly larger for respondents which consider earnings comparisons more important and lower for participants with a stronger preference for equal earning as well as for more neurotic and more conscientious persons.

Table 3. Regressions of logged percentage of -100 and 100 ratings

	b (r.s.e.)	t-value	b (r.s.e.)	t-value
socio-demographic indicators:				
woman[a]	.29 (.10)	2.94**	.22 (.09)	2.32*
age in 10 years	.42 (.15)	2.83**	.38 (.15)	2.62**
age in 10 years squared	-.04 (.01)	-2.49*	-.03 (.01)	-2.31*
education in years	-.03 (.02)	-1.97*	-.03 (.02)	-1.71†
other native language[b]	.40 (.19)	2.13*	.43 (.17)	2.58*
preferences and personality aspects:				
importance of earnings comparisons	.	.	-.15 (.05)	-3.91**
earnings equality preference	.	.	.23 (.05)	4.71**
agreeableness	.	.	.18 (.05)	3.77**
conscientiousness	.	.	.19 (.05)	3.81**
willingness to participate	.	.	.22(.05)	4.44**
constant	2.20 (.16)	13.77**	2.29 (.15)	15.39**
R^2 in %	3.7		14.2	

Note: Vignette deck indicators are included as control variables.

Reference categories: [a]man; [b]German native language

Legend: **: P(T>|t|) < .01; *: P(T>|t|) < .05; †: P(T>|t|) < .10

Source: Own calculations based on SOEP-Pretest 2008 ($N_{respondents}$ = 1,021)

Table 3 shows the analysis of factors affecting the percentage of ratings heaped on the endpoints of the response scale which indicate censoring.[13] In contrast to the analysis of heaping on the midpoint (see Table 2), the amount of censored ratings is higher in the middle age range while the importance of earnings comparisons, a preference for equal earnings and conscientiousness lead to fewer censored ratings. Taking both analyses together, the differently directed effects for heaping on the midpoint compared to the endpoints show, that these factors affect the spread of ratings over the response scale.

Furthermore, the analysis in Table 3 shows that being a woman is associated with more often evaluating earnings in vignette scenarios as extremely unjust.[14] With regard to factors related to competence, more education is associated with less censored ratings, but after controlling for differences in preferences and personality this influence is only weakly significant. Being a non-native German speaker is associated with more censored ratings likely due to these respondents taking details which differ between the scenarios less into account. Regarding factors related to the motivation of respondents, agreeableness, conscientiousness and the task specific willingness to participate lead to more vignettes being rated extremely unjust.

In addition, based on the FSE on inheritance taxation conducted in my dissertation (Groß and Lang 2018, Tabelle 4), I look at factors that affect the amount of censored ratings which indicate that the proposed tax is deemed much too high. Here, respondents who report to have a rightwing political attitude and to have less trust in political institutions more often consider the suggested taxes much too high.

[13] Since the distribution of censored ratings is right skewed, I use the natural logarithm of the percentages in the analysis.

[14] This result corresponds with the "content female worker paradox"—the finding that women are often satisfied with lower wages compared to men (Mueller and Wallace 1996, Davison 2014). One potential explanation attributes this finding to a stronger preference for equal earnings among women. By contrast, the higher amount of extremely unjust ratings by female respondents remains in my analysis after controlling for earnings equality preferences.

Overall, my analyses of the respondent related factors influencing heaping and censoring of ratings in the FSEs which are used for the studies in my dissertation show that knowledge, attitudes and motivation related to the topic of the experiments are the most relevant predictors. These topic specific characteristics of respondents are much more important for the prevalence of heaped and censored ratings compared to general characteristics like education and language skills. As a consequence, while heaping and censoring are common FSEs (see section 4.1.), the amount of these types of ratings is dependent on respondent's interest and knowledge regarding the topic at hand.

4.3. Consequences for Parameter Estimates in FSEs

In this section I look at the consequences of non-interval scaled ratings, heaping and censoring for estimates of parameters in FSEs. First, I analyze how not interval scaled ratings effect the efficiency of the estimates. Second, I assess bias in the parameters due to response behavior following fluency heuristics (see sub-section 3.3.). Finally, I discuss differences in robustness of the effects associated with vignette dimension compared to observational factors given not interval scaled or heaped ratings.

Using the paired difference rating design to assess the scaling level of ratings I found that there is an interval scaled and a non-interval scaled response mode in the FSE on student's internship preferences in my dissertation (see Lang 2018, Figure 2). In this publication I also developed a model for scaling sensitive factorial survey analysis. This model is a structural equation model (Rabe-Hesketh et al. 2004) which processes the interval scaled ratings with a linear additive equation and the ordinal among the non-interval scaled ratings based on an ordered logit equation The ratings of respondents are assigned to each of these two equations using weights which are constructed based on the interval scaling indicators r or τ_b introduced in section 4.1. (for further details see Lang 2018). Thus, the scaling sensitive model treats the ratings according to their

respective scaling level. In contrast to other methods of analysis for FSEs neither assumes that all ratings are interval scaled, nor neglects the information contained in the distances between the ratings which are really interval scaled.

Regarding the efficiency of the parameter estimates, it is remarkable that the z-values of the parameters in the scaling sensitive factorial analysis conducted in my dissertation are consistently larger compared to a standard hierarchical linear model. For the effects of the vignette dimensions—the experimental factors—the z-values are on average 18 % larger with weights based on r and 14 % larger with weights based on τ_b (Lang 2018, Table 2, column 7, Table 3, column 7). Even more striking are the differences in effects of respondent characteristics: here, the z-values are on average 37% larger for weights based on either r or τ_b compared to a hierarchical linear model. Furthermore notable, is the finding that a standard analysis excluding the 11 % of ratings which are not even ordinal scaled—i.e., for which r or τ_b smaller than zero—yields z-values of comparable size to an analysis including these ratings (Lang 2018, Table 2, columns 1 and 2, Table 3, column 2). This result indicates that these ratings are not informative for the analysis.

Taken together, these findings are in line with H4. Given some ratings in a FSE are interval scale while others are not interval scaled, a method of analysis which takes the scaling level of the ratings into account is clearly more efficient. To apply such a scaling sensitive method, an indicator of the scaling of ratings must be constructed and used to classify ratings as either interval scaled or not. In addition, dropping the ratings which are not even ordinal scaled does not decrease the efficiency of a FSE analysis.

Next, I look in how far response behavior following fluency heuristics leads to biased parameter estimates. A fluency heuristics prescribes that response behavior in FSEs follows a step wise evaluation process which starts off with focusing on salient

aspects (see sub-section 3.3.).[15] In my dissertation a so called generalized Craggit model which differentiates between three evaluation steps is developed (Lang and Groß 2020a). This generalized Craggit model is a structural equation model which enables coefficients of explanatory factors to differ over steps of the evaluation process. In my dissertation this model is applied to analyze the FSE on earnings justice attitudes which is part of the SOEP-Pretest 2008 (Lang and Groß 2020).

In the most parsimonious specification—referred to as optimized generalized Craggit model—the coefficients of the dimensions (log) earnings and occupation are larger at the beginning of the evaluation.[16] Specifically, the coefficient associated with the vignette dimension earnings in the second step of the evaluation process is about two-thirds as large compared to the first step (0.68 compared to 1; Lang and Groß 2020, Table 3, column 5), and in the third rating step it is about two-fifths as large compared to the first step (0.39 compared to 1; Lang and Groß 2020, Table 3, column 5). The effect of the vignette dimension occupational status in the second and third step of the rating process is about half as large compared to the first step (-0.06 compared to -0.13; Lang and Groß 2020, Table 3, column 5). Earnings and occupation are the dimensions with the most variance in this FSE and thus, the most salient experimental factors. In line with H5, these effects of these vignette dimensions on the ratings are stronger in

[15] If response behavior is focused on salient aspects has also been studied for FSEs in which respondents have to rate several vignettes (Auspurg et al. 2009, Sauer et al. 2011). These studies show no decreasing consistency of ratings over vignettes—measured by a lower share of explained variance—which indicates that respondents do not focus on more salient dimensions in vignettes they rate later in the experiment, a so called fatigue effect (Sauer et al. 2011). By contrast, an increasing consistency would demonstrate learning effects. Moreover, these studies show no influence of the complexity of vignettes—measured by the number of dimensions—on the consistency of the ratings. The absence of complexity and fatigue effects is in line with the idea of a simple heuristic guiding response behavior from the outset to avoid cognitive load. In difference to these studies, in the following I assess if respondents start off with focusing on the salient dimensions of an FSE in the evaluation process of each vignette, not if they focus more or less on salient dimensions over the course of an FSE consisting of multiple vignettes.

[16] The model fit for this application was optimized using the Bayesian Information Criteria (BIC) as index (Lang and Groß 2020a, Table 2).

earlier steps of the evaluation process. These findings point to the use of fluency heuristics by respondents which start by focusing on the salient aspects of FSEs.

Finally, I assess if non-interval scaled and heap ratings affect the robustness of the effects associated with vignette dimension—which are set by the experimenter—differently compared to observational factors. Setting aside the differences between stepwise and standard methods of analysis due to response behavior focusing on salient aspects discussed above, the sizes of coefficients associated with vignette dimensions are stable over different modeling approaches which either address or ignore the different scaling level or the heaping of ratings (see Lang 2018, Tables 2 and 3; Lang and Groß 2020, Table 2). Thus, the robustness of the vignette coefficients is neither influenced by non-interval scaled ratings (Lang 2018), nor by the heaping of ratings which is especially pronounced on the midpoint of the response scale (Lang and Groß 2020). For the application on earnings justice attitudes the sizes of the coefficients of the vignette dimensions are even similar over different FSEs (Lang and Groß 2020, Table 2; Auspurg et al. 2017). Only for the FSE with a rating distribution which is censored at an endpoint of the response scale in my dissertation (Groß and Lang 2018) the sizes of coefficients for some vignette dimensions differ between a standard hierarchical linear model and a Craggit model which takes the censored rating distribution into account (Groß and Lang 2018, Tabelle 3).

Compared to the effects of vignette dimensions, the coefficients of variables based on observational data are not robust to the heaping of ratings. In the analyses carried out in my dissertation (Groß and Lang 2018, Lang and Groß 2020), some of the substantive conclusions drawn in relation to observational indicators differ between methods of analysis which consider the heaping and censoring of ratings and standard methods which do not. For example, the difference in the just gender pay gap between

eastern and western federal states of Germany in the analysis of earning justice attitudes carried out is significant, if the FSE is analyzed with a generalized Craggit model, but not significant if a hierarchical linear model is used (Lang and Groß 2020). According to the review of Wallander (2009) more than 90 % of the studies using FSEs are also interested in research questions related to observational variables. Importantly for these studies, as the findings of my dissertation draw attention to, is the match of response behavior with the strategy used to analyze the FSE. If the rating distribution in a FSE shows heaps or is censored and a research question involves observational indicators, specialized methods of analysis for such data have to be used.

5. Conclusion

I began this frame paper for my dissertation, with the expectation that there is a mismatch between response behavior in FSEs and the ways ratings in FSEs are typically recorded and analyzed. The conceptual literature on FSEs assumes that, aside from idiosyncratic deviations, response behavior in FSEs can be adequately described by an additive model in which all information of a vignette scenario—especially all vignette dimensions—is simultaneously considered and weighted against each other (Rossi and Anderson 1982, Jasso 2006). Alternatively, I hypothesize that survey responses—or even evaluations and decisions of humans in general—are guided by simple cognitive heuristics which likely process information in a stepwise manner, and the structures of these heuristics lead to ratings which are not interval scaled and heaped at salient values of response scales. In this frame paper I summarized findings based on the peer - reviewed publications contained in my dissertation as well as previous research to assess my hypothesis.

With respect to the scaling level of ratings, a FSE I conducted among tertiary students—a group of comparably well-educated and highly skilled respondents—

showed that the majority of ratings was not interval but rather ordinal scaled (Lang 2018). This was the first study testing the scaling level of ratings in a FSE. To assess the scaling level of the ratings I developed a so called paired difference rating design based on conjoint measurement methods. Furthermore, the prevalence of interval scaled ratings in this FSE was not higher in design splits which implemented response instruments with more options. This result indicated that the non-interval scaled ratings were not a consequence of offering too few response options. Moreover, I developed a model for scaling sensitive factorial survey analysis which takes the scaling level of ratings into account. An analysis of the FSE on student's internship preferences using this model yielded more efficient estimates compared to standard methods of analysis. On average, the z-values of vignette coefficients were about 15 % larger while those of coefficients related to observational factors were more than 30 % larger.

Regarding the heaping and censoring of ratings, previous research had already established that rating distributions with heaps at salient values are common in FSEs (e.g., Auspurg and Hinz 2015). Even response instruments which in principle enable the expression of evaluations in detail sometimes generate rating distributions with heaps. For example, the stepwise instrument of the FSE in the SOEP-Pretest 2008 led to substantial heaping of ratings on the midpoint of the response scale (Sauer et al. 2009). Related, in my dissertation a generalized Craggit model to analyze stepwise evaluation processes was developed (Lang and Groß 2020a). Modeling the stepwise rating process in this FSE revealed that respondents first focused on the salient dimensions of the FSE (Lang and Groß 2020). Moreover, my analyses of the response behavior in the FSEs used for my dissertation showed that factors related to the topic of the FSEs are more relevant to explain heaping and censoring in comparison to general characteristics of the respondents like education.

Overall, these findings on the scaling level, heaping and censoring of ratings in FSEs support my hypothesis that response behavior is guided by simple heuristics. These heuristics lead to classifications rather than continuous evaluations and likely prescribe to start off evaluations with a focus on salient dimensions of the FSEs. In addition, aside of the above discussed biases in vignette effects estimated with methods of analysis which do not take a stepwise response behavior into account, the studies carried out in my dissertation demonstrated that vignette coefficients in FSEs are robust pertaining to a systematic mismatch of response behavior with the assumptions of interval scaled and continuous—not heaped or censored—ratings. However, the analyses in my dissertation also clearly showed that the coefficients of observational factors—which are also of interest in many FSEs—are not robust to heaping or censoring. For related research questions an adequate modeling of response behavior is especially important.

References

Abraham, M., Auspurg, K., Bähr, S., Frodermann, C., Gundert, S., and Hinz, T. 2013. Unemployment and willingness to accept job offers: Results of a factorial survey experiment. *Journal of Labor Market Research* 46(4):283-305.

Anderson, N. H. 1974. Information integration theory: A brief survey. In: Krantz, D. H., Atkinson, R., Luce, D., and Suppes, P. (Eds.). Contemporary developments in mathematical psychology. San Francisco: F. H. Freeman.

Auspurg, K. and Gundert, S. 2015. Precarious employment and bargaining power: Results of a factorial survey analysis. *Zeitschrift für Soziologie (Journal of Sociology)* 44(2):99-117.

Auspurg, K. and Hinz, T. 2015. Factorial survey experiments. Quantitative Applications in the Social Sciences 175. Thousand Oaks, CA: Sage.

Auspurg, K. and Hinz, T. 2015a. Multifactorial Experiments in Surveys : Conjoint Analysis, Choice Experiments, and Factorial Surveys. In: Keuschnigg, M. and Wolbring, T. (Eds.). Experimente in den Sozialwissenschaften (*Soziale Welt*, Sonderband 22). Baden-Baden: Nomos. Pp. 291-315.

Auspurg, K., Hinz, T., and Liebig, S. 2009. Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey. (Complexity, Learning Effects and Plausibility of Vignettes in the Factorial Survey Design.) *Methods, Data, Analyses* 3(1):59-96.

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

Auspurg, K., Hinz, T., Liebig, S., and Sauer, C. 2009a. Auf das Design kommt es an: Experimentelle Befunde zu komplexen Settings in Faktoriellen Surveys. soFid Methoden und Instrumente der Sozialwissenschaften 2/2009:23-39. Bonn: GESIS - Leibniz-Institut für Sozialwissenschaften.

Auspurg, K., Hinz, T., and Sauer, C. 2017. Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments. *American Sociological Review* 82(1). Pp.179-210.

Boots, D. P., Cochran, J. K., and Heide, K. M. 2003. Capital punishment preferences for special offender populations. *Journal of Criminal Justice* 31(6):553-565.

Davison, H. K. 2014. The paradox of the contented female worker: Why are women satisfied with lower pay? *Employee Responsibilities and Rights Journal* 26(3):195-216.

de Wolf, I. and van der Velden, R. 2001. Selection processes for three types of academic jobs. An experiment among Dutch employers of social sciences graduates. *European Sociological Review* 17(3):317-330.

Dülmer, H. 2007. Experimental plans in factorial surveys: Random or quota design? *Sociological Methods and Research* 35(3):382-409.

Emerson, M. O., Yancey, G., and Chai, K. J. 2001. Does race matter in residential segregation? Exploring the preferences of white Americans. *American Sociological Review* 66(6):922-935.

Esser, H. 1993. The rationality of everyday behavior. A rational choice reconstruction of the theory of action by Alfred Schütz. *Rationality and Society* 5(1):7-31.

Faia, M.1980. The vagaries of the vignette world: A comment on Alves and Rossi. *American Journal of Sociology* 85(1):951-954.

Gigerenzer, G. 2008. Why heuristics work. *Perspectives on Psychological Science* 3(1):20-29.

Gigerenzer, G. and Todd, P. M. (Eds.). 1999. Simple heuristics that make us smart. New York: Oxford University Press.

Gross, C., Lorek, K., and Richter, F. 2017. Attitudes towards inheritance taxation: Results from a survey experiment. *Journal of Economic Inequality* 15(1):93-112.

Groß, M. and Lang, V. 2018. Warum Bürger gegen die Erhebung von Erbschaftssteuern sind – auch wenn sie keine zahlen müssen: Ergebnisse einer Vignettenstudie. (Why citizens oppose inheritance taxes – Even if they do not have to pay them: Results of a factorial survey.) *Zeitschrift für Soziologie (Journal of Sociology)* 47(3):200-207.

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

Imbens, G. W. and Rubin, D. B. 2015. Causal inference in statistics, social, and biomedical sciences. Cambridge: Cambridge University Press.

Jann, B. 2003. Lohngerechtigkeit und Geschlechterdiskriminierung: Experimentelle Evidenz. Working Paper of the Institute of Sociology. Zürich: ETH Zürich.

Jasso, G. 2006. Factorial survey methods for studying beliefs and judgments. *Sociological Methods and Research* 34(3):334-423.

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

Jasso, G. and Webster, M. 1999. Assessing the gender gap in just earnings and its underlying mechanisms. *Social Psychology Quaterly* 62(4):367-380.

Kahneman, D. and Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263-292.

Kroneberg, C. 2005. Die Definition der Situation und die variable Rationalität der Akteure. Ein allgemeines Modell des Handelns. (The definition of the situation and the variable rationality of actors. A general model of action.) *Zeitschrift für Soziologie (Journal of Sociology)* 34(5):344-363.

Kuhfeld, W.F. 2010. Marketing research methods in SAS: Experimental design, choice, conjoint and graphical techniques. Cary, NC: SAS Institute Inc.

Lang, V. 2018. Scaling sensitive factorial survey analysis. *Sociological Methods and Research*, online first.

Lang, V. and Groß, M. 2020. The just gender pay gap in Germany revisited: The male breadwinner model and regional differences in gender-specific role ascriptions. *Research in Social Stratification and Mobility* 65, online first.

Lang, V. and Groß, M. 2020a. Analyzing rating distributions with heaps and censoring points using the generalized Craggit model. *MethodsX* 7, online first.

Liebig, S. 2001. Lessons from philosophy? Interdisciplinary justice research and two classes of justice judgements. *Social Justice Research* 14(3):265-287.

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

Louviere, J. J., Hensher, D. A., and Swait, J.D. 2000. Stated choice methods: Analysis and applications. Cambridge: Cambridge University Press.

Markovsky, B. and Eriksson, K. 2012. Comparing direct and indirect measures of just rewards: What have we learned? *Sociological Methods and Research* 41(1):240-245.

Martignon, L., Vitouch, O., Takezawa, M., and Forster, M. 2003. Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. In: Hardman, D. and Macchi, L. (Eds.). Thinking: Psychological perspectives on reasoning, judgment and decision making. West Sussex: John Wiley and Sons. Pp.189-211.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63(2):81-97.

Mueller, C. W., and Wallace, J. E. 1996. Justice and the paradox of the contented female worker. *Social Psychology Quarterly* 59(4):338-349.

Orth, B. 1982. A theoretical and empirical study of scale properties of magnitude-estimation and category-rating scales. In: Wegener, B. Social attitudes and psychophysical measurement. Hillsdale, NJ: Lawrence Erlbaum. Pp.351-377.

Orth, B. and Wegener, B. 1983. Scaling occupational prestige by magnitude estimation and category rating methods: A comparison with the sensory domain. *European Journal of Social Psychology* 13(4):417-431.

Petzold, K. and Wolbring, T. 2019. What can we learn from factorial surveys about human behavior? A validation study comparing field and survey experiments on discrimination. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 15(1):19-30.

Petzold, K. and Wolbring, T. 2018. Zur Verhaltensvalidität von Vignettenexperimenten. Theoretische Grundlagen, Forschungsstrategien und Befunde. In: Menold, N. and Wolbring, T. (Eds.). Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente. Wiesbaden: Springer VS. Pp.307-338.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. 2004. Generalized multilevel structural equation modeling. *Psychometrika* 69(2):167-190.

Rammstedt, B. and Beierlein, C. 2014. Can't we make it any shorter? The limits of personality assessment and ways to overcome them. *Journal of Individual Differences* 35(4):212-220.

Rossi, P. H. and Anderson, A. B. 1982. The factorial survey approach: An introduction. In: Rossi, P. H. and Nock, S. L. (Eds.). Measuring social judgments: The factorial survey approach. Beverly Hills, CA: Sage. Pp.15-67.

Sauer, C., Auspurg, K., Hinz, T., and Liebig, S. 2011. The application of factorial surveys in general population samples: The effects of respondent age and education on response times and response consistency. *Survey Research Methods* 5(3):89-102.

Sauer, C., Liebig, S., Auspurg, K., Hinz, T., Donaubauer, A., and Schupp, J. 2009. A factorial survey on the justice of earnings within the SOEP-Pretest 2008. Institute of Labor Economics (IZA) Discussion Paper No. 4664. Bonn: IZA.

Schooler, L.J., and Hertwig, R. 2005. How forgetting aids heuristic inference. *Psychological Review* 112(3):610-628.

Schrenker, M. and Wegener, B. 2007. Was ist gerecht? Ausgewählte Ergebnisse aus dem International Social Justice Project 1991–2007. International Social Justice Project (ISJP) Arbeitsbericht 150. Berlin: Humboldt-Universität.

Siegel, N. A., Stocker, A., and Warnholz, S. 2009. SOEP Testerhebung 2008: Persönlichkeit, Gerechtigkeitsempfinden und Alltagsstimmung. Methodenbericht. München: TNS Infratest Sozialforschung.

Simon, H. A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69(1):99-118.

Tourangeau, R., Rips, L. J., and Rasinski, K. A. 2000. The psychology of survey response. Cambridge: Cambridge University Press.

Tversky, A. and Kahneman, D. 1974. Judgement under Uncertainty: Heuristics and biases. *Science* 185(4157):1124-1131.

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

Wegener, B. 1983. Category-rating and magnitude estimation scaling techniques: An empirical comparison. *Sociological Methods and Research* 12(1):31-75.

Appendix 1

Table A1. Descriptive statistics of indicators based on SOEP-Pretest 2008

	N	mean	s.d.	min.	max.
percentage of ..					
.. just ratings	1,066	34.8	18.1	0	100
.. "-100" and "100" ratings	1,066	19.5	20.9	0	96
.. ratings not only using tens digits	1,066	5.8	12.7	0	100
woman	1,066	0.53	0.50	0	1
age in years	1,066	51.7	1.9	16	92
education in years	1,064	11.0	3.1	8	21
other native language	1,066	0.07	0.26	0	1
earnings equality preference	1,058	0	1	-3.1	1.0
importance of earnings comparisons	1,041	0	1	-1.0	3.5
agreeableness	1,060	0	1	-3.8	2.2
conscientiousness	1,054	0	1	-3.3	1.4
neuroticism	1,061	0	1	-1.7	2.8
willingness to participate	1,066	0	1	-4.4	0.9

Source: Own calculations based on SOEP-Pretest 2008

Appendix 2

Abstract

(of the cumulative dissertation: Lang, V. 2020. Response behavior in factorial survey experiments: Challenges and innovative solutions. Tübingen: University of Tübingen. Link: https://rds-tue.ibs-bw.de/opac/RDSIndexrecord/1725134705)

This cumulative dissertation consists of four peer-reviewed publications and a frame paper. All publications address some aspects of response behavior in factorial survey experiments (FSEs) which lead to heaped, censored and non-interval scaled ratings. The frame paper of this dissertation (Lang 2020) proposes a new concept of response behavior in FSEs in terms of simple cognitive heuristics, which can explain the common occurrence of heaped and not interval scaled ratings. Inasmuch as response behavior is guided by such heuristics, respondents likely evaluate vignettes in a stepwise manner, start with a focus on salient aspects, and produce ordered categorizations instead of interval scaled ratings. The findings summarized in Lang (2020) based on the peer-reviewed publications of this cumulative dissertation support this alternative conceptualization of response behavior in FSEs.

The first peer-reviewed publication in this dissertation is the first study to test the scaling level of ratings in a factorial survey experiment (Lang 2018). To this end, a so called paired difference rating test design based on conjoint measurement methods is developed. In an exemplary FSE on tertiary student's internship preferences, the test showed that around 60 % of the ratings were not interval scaled. This result indicates that there is an interval and a non-interval scaled response mode in FSEs. To take these different types of response behavior into account, a model for scaling sensitive factorial survey analysis is developed. In an analysis of the exemplary FSE using this new model

the parameter estimates are much more efficient. The z-values of vignette effects are about 15 % larger and those of observational characteristics are over 30 % larger.

The second and the third peer-reviewed publications of this dissertation each explain a research puzzle related to the social stratification of justice attitudes on wealth and earnings (Groß and Lang 2018, Lang and Groß 2020). Using a FSE, Groß and Lang (2018) examine why several surveys which used direct questioning techniques have found that a majority of Germans oppose taxes on inheritances even so they would profit from related redistribution effects. The results indicate that the activation of a self-interested reference frame is decisive to explain why people more often principally oppose the taxation of inheritances if they are directly asked about it. By contrast, the analysis of the FSE shows that respondents favor a progressive inheritance taxation and exemptions if company assets are inherited. Furthermore, less trust in political institutions and a more right-wing political attitude of respondents lead to a higher probability for a principal opposition towards the taxation of inheritances.

Lang and Groß (2020) is the first study which can explain the just gender pay gap in Germany repeatedly found by FSEs. Based on the so called “male breadwinner model” of family labor division, it is hypothesized that female as well as male respondents favor higher earnings for men with children compared to childless men and women. In line with this expectation a just gender pay gap of about 8 % is found if there are children in the vignette scenarios, whilst there is no just gender pay gap for scenarios without children. The influence of the male breadwinner model on earnings justice attitudes tends to be stronger in the eastern compared to the western federal states. In Lang and Groß (2020a) a so called “generalized Craggit model” used to analyze stepwise response behavior in FSEs is developed. This model is applied in the study of earnings justice attitudes in this dissertation (Lang and Groß 2020).

All in all, this dissertation contributes new insights with respect to the conceptualization of response behavior in FSEs and in regards to the social stratification of justice attitudes on taxation and earnings. In addition, it provides researchers using FSEs with new methods of analysis to adequately handle non-interval scaled, heaped and censored ratings.

References

Groß, M. and Lang, V. 2018. Warum Bürger gegen die Erhebung von Erbschaftssteuern sind – auch wenn sie keine zahlen müssen: Ergebnisse einer Vignettenstudie. (Why citizens oppose inheritance taxes – Even if they do not have to pay them: Results of a factorial survey.) *Zeitschrift für Soziologie (Journal of Sociology)* 47(3):200-207.

Lang, V. 2020. Scaling level of responses, heaping and censoring in factorial surveys: Expectations and evidence in view of a simple cognitive model. Tübingen: University of Tübingen.

Lang, V. 2018. Scaling sensitive factorial survey analysis. *Sociological Methods and Research*, online first.

Lang, V. and Groß, M. 2020. The just gender pay gap in Germany revisited: The male breadwinner model and regional differences in gender-specific role ascriptions. *Research in Social Stratification and Mobility* 65, online first.

Lang, V. and Groß, M. 2020a. Analyzing rating distributions with heaps and censoring points using the generalized Craggit model. *MethodsX* 7, online first.

Appendix 3

Concluding Remarks

(of the cumulative dissertation: Lang, V. 2020. Response behavior in factorial survey experiments: Challenges and innovative solutions. Tübingen: University of Tübingen. Link: https://rds-tue.ibs-bw.de/opac/RDSIndexrecord/1725134705)

1. Overview

This cumulative dissertation contains four peer-reviewed publications and a frame paper. The first of the peer-reviewed publications was the first study to test the scaling level of ratings in a factorial survey experiment (FSE, Lang 2018). The second peer-reviewed publication assessed the paradox that a majority of Germans opposes taxes on inheritances even so they would profit from related redistribution effects (Groß and Lang 2018). The third peer-reviewed publication explained the just gender pay gap in Germany by the influence of the so called male breadwinner model of family labor division on earnings justice attitudes using data of a FSE (Lang and Groß 2020). Thus, Groß and Lang (2018) and Lang and Groß (2020) addressed two puzzles in the field of research on the social stratification of justice attitudes. In the fourth peer-reviewed publication a so called generalized Craggit model to analyze stepwise evaluation processes was developed (Lang and Groß 2020a). This model was applied in the analysis of earnings justice attitudes in Lang and Groß (2020). Response behavior in FSEs which leads to heaped, censored and non-interval scaled ratings is a topic which was encountered in all publications. The frame paper of my dissertation discussed different conceptualization of response behavior in FSEs and related expectations as well as findings based on the four peer-reviewed publications with respect to the scaling level, heaping and censoring of ratings (Lang 2020).

In the following I give an overview of the content and contributions of each publication in my dissertation: first, with respect to novel findings on the social stratification of justice attitudes; second, regarding improvements to the toolbox of methods which can be used to analyze FSEs; and third, with regard to an alternative conceptualization of response behavior in FSEs which can explain non-interval scaled and heaped ratings.

2. Contributions regarding the social stratification of justice attitudes

Groß and Lang (2018) and Lang and Groß (2020) dealt with substantive research questions on the social stratification of justice attitudes related to inequalities in wealth and earnings. In Groß and Lang (2018) the surprising findings of several surveys using direct questioning techniques that a majority of Germans completely opposes inheritance taxation were assessed using the indirect response format of a FSE. Since most citizens would gain from the redistribution effects associated with taxing inheritances this oppositional attitude is paradoxical. To activate an impartial frame of reference in evaluating the scenarios, the vignettes clearly stated that the heir is not the respondent—in addition to other conditions of the inheritance.

In this FSE only 11 % of the respondents strictly opposed the taxation of inheritances. This result indicates that the activation of a self-interested reference frame is critical to explain why more people oppose inheritance taxation in principal if they are directly asked about it. In general, respondents favored a progressive taxation of inheritances and tax exemptions for the inheritance of company assets, especially if the inherited firm had many employees. With respect to explanatory factors on the respondent level, less trust in political institutions and a more right-wing political attitude were factors associated with an unconditional rejection of inheritance taxes. The principal rejection of inheritance taxes by part of the respondents led to a censored

distribution of ratings which was addressed by a Craggit model. To adequately assess respondent level effects, this study was, to my knowledge, the first to apply a random intercept version of the Craggit model in an analysis of a FSE.

An often replicated finding of the research on the social stratification of justice attitudes using FSEs is that both women and men consider lower earnings for women just, even after differences in job, qualification, and performance related characteristics are accounted for. This result is called just gender pay gap and for Germany different studies estimate this gap to be about 6 %. Lang and Groß (2020) was the first study to provide an explanation for this gender difference: a system of values prescribing a traditional family division called the "male breadwinner model". Specifically, the male breadwinner model entails the belief that fathers should be gainfully employed to provide for the material needs of their family while mothers attend to the unpaid family work, which led to the hypothesis that the just gender pay gap is larger if there are children in the family. This expectation was tested using a FSE conducted with the population-representative sample of the SOEP-Pretest 2008.

In line with the male breadwinner model explanation, the results showed a just gender pay gap of about 8 % if there are children in a vignette scenario while there is no just pay gap between childless women and men. In addition, the analysis indicated that the relevance of the male breadwinner model explanation depended on macro social conditions. In the eastern federal states of Germany—where women have been more integrated into the labor market over the last decades—the gender differences in the evaluation of situations with and without children as well as the just gender pay gap overall were smaller. Given the changes in family policies and family labor division over the last decade it would be interesting to look at potential trends in the just gender pay gap and the relevance of the male breadwinner model explanation using recent data.

Similarly, future research should assess the importance of the male breadwinner model explanation for just gender pay gaps in other countries.

Taken together, Groß and Lang (2018) and Lang and Groß (2020) explained two research puzzles related to the social stratification of justice attitudes. In both chapters, the findings highlight the ability of FSEs to disclose attitudinal structures which in direct response formats would likely be not detected due to social desirability bias.

3. Contributions to factorial survey research methods

While heaping and censoring are observable characteristics of a distribution, the scaling level of responses has to be assessed using additional instruments which are typically variants of conjoint measurement methods. Lang (2018) was the first study which developed and implemented such a scaling level test for the ratings in a FSE, a so called paired difference rating design. I implemented this design in a FSE on internship preferences of tertiary students. While this was a group of comparably well educated respondents, the test showed that around 60 % of the ratings were not interval scaled. 11 % of the ratings were not ordinal but rather nominal scaled. In addition, the amount of not interval scaled ratings was not lower in design splits using response instruments with more options like a slider. These findings clearly contradict the assumption that ratings in FSEs are interval scaled and indicate that there is an interval scaled and a not interval scaled response mode in the FSE.

To take these differences in response behavior into account for analyses of FSEs, I developed a model for scaling sensitive factorial survey analysis. In contrast to standard methods of analysis for FSEs, this model processes ratings according to their respective scaling level. In an analysis of the FSE on students' internship preferences the z-values of the vignette coefficients in this model were around 15 % larger compared to standard methods, and those of the coefficients related to observational

factors were more than 30 % larger on average. These findings indicate huge gains in the efficiency of studies based on FSEs if the scaling level of ratings is adequately considered by analyses.

Compared to previous laboratory experiments on the scaling level of ratings the paired difference rating design I developed for FSEs is parsimonious. However, it is more laborious than a standard FSE design which is a major reason why the scaling level of ratings has so far not been tested in a FSE with a population representative sample. To conduct such a study as part of future research would be necessary to assess in how far the scaling level of ratings in FSEs is socially stratified and in which ways such stratification affects substantive conclusions. The biases found due to heaped ratings in Groß and Lang (2018) and Lang and Groß (2020) highlight the importance of additional research in this direction. Furthermore, using a scaling sensitive analysis such a study could assess if there are similar gains in the efficiency of estimates based on a representative sample.

In the FSE included in the SOEP-Pretest 2008 which was analyzed in Lang and Groß (2020) a response instrument with several evaluation steps was used. This instrument led to a rating distribution with multiple heaps, especially pronounced was the heap of just ratings on the midpoint of the response scale. To capture this evaluation process adequately, a so called generalized Craggit model was developed in Lang and Groß (2020a). This model combines a random intercept Craggit model with a random intercept generalized ordered probit model. It enables to analyses of stepwise evaluations which start off with a coarse classification—leading to heaps of ratings—while it also considers the more differentiated ratings resulting of later and more detailed evaluation steps. The model fitted the data of the FSE in the SOEP-Pretest 2008 much better than standard methods of analysis and demonstrated that respondents

started their vignette evaluations with a focus on salient dimensions, specifically, the earnings and occupations of vignette persons.

Overall, the methods of analysis developed in Lang (2018) and Lang and Groß (2020a) provide researchers using FSEs a wide range of tools to deal with different types of rating distributions.

4. An alternative conceptualization of response behavior in factorial surveys

Finally, some aspects of all publications in my dissertation dealt with the response behavior in FSEs. Specifically, with methods to address the consequences of a response behavior which leads to heaped, censored and not interval scaled ratings. The frame paper of my dissertation (Lang 2020) presented conceptual considerations to explain the response behavior in FSEs and related findings based on the peer-reviewed publications of my dissertation. While the conceptual literature so far describes response behavior in FSEs with an additive model, which assumes that ratings are continuous and interval scaled, I alternatively hypothesized that response behavior in FSEs is guided by simple cognitive heuristics. These heuristics likely produce an ordered categorization. In consequence, the resulting ratings are not interval scaled and heaped at salient values of response scales. In line with a response behavior guided by simple heuristics the studies in my dissertation showed that non-interval scaled, heaped and censored ratings are common in FSEs, and furthermore, that respondents tend to evaluate vignettes stepwise and start with a focus on salient aspects. In addition, models which took a stepwise response behavior and non-interval scaled ratings into account yielded more efficient parameter estimates.

At first glance, the robustness of vignette effects in FSEs supports the idea that standards methods of analysis like an additive model can be used to estimate the "social components" of evaluations based on FSEs. For example, the estimates of the just

gender pay gap for Germany in Lang and Groß (2020) were similar using different modeling approaches and samples. However, the differences in vignette effects I found using models which suppose a stepwise response behavior are highly relevant to the external validity of FSEs. First, since the aim of studies using FSEs is to uncover the regularities guiding evaluations in real life situations, a match of the heuristics supposed to guide response behavior in FSE analyses with these regularities is important. Also, if research aims to predict actual behavior using estimates of behavioral intentions based on FSEs the vignette effects have to be similar to the influences of comparable factors on decisions in real life.

Second, in contrast to the vignette effects, the estimates of the effects related to observational factors in the studies for my dissertation often differed substantially depending on the methods of analysis used. Therefore, to adequately capture the evaluation processes in real life situations based on FSEs, it would be useful to apply the more sophisticated methods of analysis developed in my dissertation in future studies. In addition, more specific assessments of the types of heuristics guiding response behavior in FSEs and which kind of respondents follow certain heuristics would be relevant contributions of future research. More detailed knowledge about the structure and prevalence of these heuristics would further our understanding of the formation and social stratification of human attitudes, evaluations and decisions.

References

Groß, M. and Lang, V. 2018. Warum Bürger gegen die Erhebung von Erbschaftssteuern sind – auch wenn sie keine zahlen müssen: Ergebnisse einer Vignettenstudie. (Why citizens oppose inheritance taxes – Even if they do not have to pay them: Results of a factorial survey.) *Zeitschrift für Soziologie (Journal of Sociology)* 47(3):200-207.

Lang, V. 2020. Scaling level of responses, heaping and censoring in factorial surveys: Expectations and evidence in view of a simple cognitive model. Tübingen: University of Tübingen.

Lang, V. 2018. Scaling sensitive factorial survey analysis. *Sociological Methods and Research*, online first.

Lang, V. and Groß, M. 2020. The just gender pay gap in Germany revisited: The male breadwinner model and regional differences in gender-specific role ascriptions. *Research in Social Stratification and Mobility* 65, online first.

Lang, V. and Groß, M. 2020a. Analyzing rating distributions with heaps and censoring points using the generalized Craggit model. *MethodsX* 7, online first.