ISER Working Paper Series

First equals most important? Order effects in vignette-based measurement

Katrin Auspurg
Department of History and Sociology
University of Konstanz

Annette Jäckle
Institute for Social and Economic Research
University of Essex

No. 2012-01
January 2012

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

www.iser.essex.ac.uk

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Non-technical summary

Vignettes are increasingly used in social science surveys, to measure how people make decisions and what determines their attitudes. A vignette typically describes a hypothetical situation or object, about which respondents are asked to make a judgment. The object is described as having various characteristics. For example in our research, the vignettes describe full-time employees. The characteristics of the vignettes are experimentally varied, so that researchers can estimate the impact of individual characteristics on respondents' judgments. In our research we use vignettes to investigate which characteristics of an employee, or the organization they are working for, should determine how much they earn, in order for their earnings to be judged as fair.

Drawing on the literature in cognitive psychology and survey methodology, we examine the following research questions: Does the order in which characteristics are presented in the vignette affect respondents' judgments? Does this affect research conclusions? Under which conditions are order effects mostly likely to occur?

We use data from a web survey of 300 students to analyze several possible conditions: features of the vignette design, characteristics of respondents, and interactions between these. Our results show that strong order effects can occur, which alter conclusions about which characteristics respondents think should determine how much an employee earns. The order however only matters when the vignettes are complex, that is, when employees are described with 12 rather than 8 different characteristics – or when respondents are asked two questions about each vignette rather than just one. Order effects are more likely for respondents who have little knowledge or weak attitudes about the topic the vignettes are describing. Contrary to expectations respondents' cognitive ability did not appear to matter. The results have implications for how best to design vignettes, in order to avoid order effects that could impact results.

First equals most important?
Order effects in vignette-based measurement

Katrin Auspurg (University of Konstanz) and Annette Jäckle (University of Essex)

Abstract

A vignette typically describes a hypothetical situation or object which respondents are asked to judge. The object is described as having different dimensions, the values of which are experimentally varied, so that their impact on respondents' judgments can be estimated. We examine 1) whether the order in which dimensions are presented impacts estimates, and 2) under which conditions order effects are mostly likely. Using data from a web survey of students we analyze several possible conditions: features of the vignette design, characteristics of respondents, and interactions between these. Our results show that strong order effects can occur, but only when the vignettes are complex.

Keywords: Factorial survey design; conjoint analysis; dimension order; survey research.

JEL classification: C83

Acknowledgements: This paper uses data from the research project "The Factorial Survey as a Method for Measuring Attitudes in Population Surveys", funded by the German Research Foundation (DFG) at the Universities of Konstanz and Bielefeld (Priority Programme 1292 on Survey Methodology; grant number HI 680/7-1). We thank Thomas Hinz, Stefan Liebig, Carsten Sauer, and especially Stefan Wehrli for their valuable support with data collection, and Thomas Hinz for comments on an earlier draft of this paper.

Contact: Katrin Auspurg, Department of History and Sociology, University of Konstanz, Universitätsstr. 10, Postfach D 40, 78457 Konstanz, Germany. Email: Katrin.Auspurg@uni-konstanz.de

Introduction

The factorial survey method is well established in the social sciences as a method of assessing respondents' beliefs about the world, judgment principles, or decision rules (see Wallander 2009 for a review of applications). Instead of single-item questions, respondents are confronted with multi-dimensional stimuli (*vignettes*) that resemble real-life judgments or decision making situations. Within these vignettes some attributes (*dimensions*) are experimentally varied in their values (*levels*). This experimental variation allows the researchers to assess the exact impact of each of the dimensions on the evaluation task (Alexander and Becker 1978; Jasso 2006a; Rossi and Anderson 1982). In this article we examine whether the order in which dimensions are presented to the respondent has any effect on the evaluations, and test hypotheses about the moderators of order effects.

Factorial surveys are used increasingly in academic and non-academic research, including the social sciences, law studies, and consumer research. Classical applications consist of the evaluation of fairness of income (Alves and Rossi 1978; Hermkens and Boerman 1989; Jasso and Rossi 1977; Jasso and Webster Jr. 1997; Jasso and Webster Jr. 1999; Shepelak and Alwin 1986), the criteria for welfare payments and fair tax rates (Liebig and Mau 2002; Liebig and Mau 2005), and the rating of social status of households (Meudell 1982; Nock 1982; Rossi 1979; Rossi, Sampson, Bose, Jasso, and Passel 1974). In addition, factorial surveys have been used to reveal respondents' definitions of sexual harassment (Garret 1982; O'Toole, Webster, O'Toole, and Lucal 1999), appropriate sentences for criminals (Berk and Rossi 1977; Hembroff 1987; Miller, Rossi, and Simpson 1986), criteria for the desirability of immigrants (Jasso 1988) and for deserving medical treatment (Hechter, Ranger-Moore, Jasso, and Horne 1999). Further applications are decision rules of professionals like teachers or nurses (Ludwick, Wright, Zeller, Dowding, Lauder, and Winchell 2004; O'Toole, Webster, O'Toole, and Lucal 1999), the preconditions for social norms (Diefenbach and Opp 2007; Jasso and Opp 1997), trust (Buskens and Weesie 2000), or discriminating behavior (John and Bates 1990), or even possibilities of overcoming social dilemmas (Abraham, Auspurg, and Hinz 2010).

Although factorial surveys are frequently used, there is only little research on methodological issues. Issues that have been studied include the effects of complexity (i.e. the number of vignettes and dimensions) on the consistency of responses, learning and fatigue effects (Sauer, Auspurg, Hinz, and Liebig 2011; Sauer, Liebig, Auspurg, Hinz, Donaubauer, and Schupp 2009), the effects of illogical combinations of vignette dimensions (Auspurg, Hinz, and Liebig 2009), the impact of the range of levels of dimensions (Jasso 2006b) and strategies for sampling vignettes (Dülmer 2007; Steiner and Atzmüller 2006). The effects of the order in which respondents evaluate vignettes have been demonstrated and are sometimes addressed by first giving the respondents some base vignettes that are the same for all respondents (Garret 1982; O'Toole, Webster, O'Toole, and Lucal 1999) or by randomizing the order in which vignettes are presented (Rossi and Anderson 1982). The order in which dimensions are used within vignettes has to our knowledge not received any attention. This is astonishing since order effects are considered one of the main aspects of questionnaires design that potentially impair data quality.

Survey researchers have worried for years about how the order of single survey questions, or of response categories, might affect responses. Despite hundreds of experiments varying the question or response order, there is still little understanding of the conditions leading to order effects: sometimes they occur, sometimes not; sometimes with expected, sometimes with unexpected patterns (Schuman 1992; Tourangeau 1999; Tourangeau, Singer, and Presser 2003). Studying order effects in factorial survey designs might not only serve to derive practical guidelines for further improvements of this particular method, but also help gain a deeper understanding of the causal mechanisms underlying order effects in general. The design of factorial surveys is extraordinarily well suited for experimentally varying the complexity of the evaluation task. Prior studies on the relationship between task difficulty and order effects suffer from the limitation that the response task differed not only in complexity but also in other features like the question format or answer scales (Malhotra 2009). Factorial surveys offer the possibility of varying complexity while keeping the question format and response options fixed. Furthermore, the repeated evaluation of similar vignettes by each respondent offers unique opportunities for studying the interaction of order effects with learning and fatigue of respondents.

The present article contributes both to the methodological literature on the design of factorial surveys, and to the more general survey literature on the mechanisms underlying order effects. We address two main research problems. *First*, we examine whether the order of vignette dimensions matters for the results obtained from factorial survey designs: Does the order of dimensions impact on the *absolute* effect sizes, and therefore on the statistical significance of the effects of vignette dimensions on evaluations? If this were the case, hypothesis testing, which is one of the main aims of factorial survey design, would be compromised. Does the order of dimensions have any effect on the *relative* importance of vignette dimensions? Several applications use the factorial survey to rank the importance of dimensions (see, for instance, Hermkens and Boerman 1989; Miller, Rossi, and Simpson 1986). The relative importance affects the calculation of trade-offs between dimensions, and of common estimates such as just gender pay gaps or willingness to pay. What are the practical implications of order effects – are they strong enough to change substantive conclusions? *Second*, we examine under which conditions order effects are mostly likely to occur. Drawing on the literature in cognitive psychology and survey methodology, we assume that order effects are most likely when the vignettes are very complex, when respondents have lower cognitive ability, have less knowledge of the topic, and are less certain in their attitudes. In addition, we examine the possible impact of dimension-importance in moderating order effects.

Theoretical Background and State of Research

Order effects can be defined as changes in answers to survey questions that are produced by varying the order in which questions or response options are presented (Krosnick and Alwin 1987: 202). A large number of survey experiments have demonstrated the existence of order effects. Yet order effects are not easy to predict and there is still no general explanation for their occurrence. In the following we summarize the main theoretical assumptions and empirical findings from cognitive psychology and survey research, from methodological research on factorial survey designs, and from research on other related experimental survey methods (conjoint analyses and choice experiments).

Order Effects in Cognitive Psychology and Survey Research

Several potential mechanisms have been identified that could trigger the occurrence of order effects: limitations of cognitive memory, context effects and satisficing behavior. The research has further identified potential moderator variables that may strengthen or dampen order effects. Empirical evidence testing these hypotheses is however mixed.

Mechanisms Leading to Order Effects. Limitations of cognitive memory were first theorized to cause order effects: items presented early in a list are more likely to enter long-term memory, while items presented at the end are more likely to enter short-term memory. Therefore items presented at the beginning or end of long lists are more likely to be recalled than items in the middle (see Krosnick 1992: 205). Theories on the functioning of our working memory further suggest that when questions are read out to respondents, recency effects are most likely, meaning that respondents are more likely to select items listed last. In contrast, when questions are presented visually primacy effects are most likely, meaning that respondents are more likely to select items listed first (Krosnick and Alwin 1987; Schwarz, Hippler, and Noelle-Neumann 1992). Applied to factorial surveys, where vignettes are typically presented visually, limitations of cognitive memory could mean that respondents attach more importance to a dimension if it is placed at the beginning of the vignette text.

Context effects can take on different forms relevant to order effects. Priming effects occur when information presented earlier establishes a cognitive framework or reference point that guides the interpretation of later information. Preceding survey context can prime schemata – sets of closely related arguments – that lead to different interpretations of later items (Bradburn 1992: 319). The context also determines what information the respondent has in mind when evaluating a survey question (Sudman, Bradburn, and Schwarz 1996; Tourangeau, Singer, and Presser 2003). Applied to vignette evaluations, context effects could mean that respondents may interpret a dimension differently, depending on the order in which they read – and cognitively process – the dimensions.

Satisficing behavior is a further potential explanation for order effects. Information presented earlier might be subject to deeper cognitive processing than information presented later and therefore have more impact on evaluations. Based on Simon's principle of *satisficing* (Simon 1957), Krosnick theorized that respondents do not necessarily make sufficient effort to answer survey questions optimally, but in some circumstances shortcut the response process to provide satisfactory answers requiring least effort (Krosnick 1991; Krosnick 1992). Applied to factorial survey designs, satisficing means that respondents might sometimes base their evaluations on only few – and possibly only the first – dimensions, and not take account of further dimensions.

Moderators of Order Effects. Respondents' cognitive ability is relevant to all three potential mechanisms through which order effects can occur. Memory limitations increase with age and as a result primacy and recency effects are stronger with older respondents (Schwarz and Knäuper 2000; although Holbrook, Krosnick, Moore, and Tourangeau 2007 find no such relation). Satisficing is more likely for respondents with low cognitive ability or low educational background, presumably because the response task is more burdensome for them (Holbrook, Green, and Krosnick 2003; Narayan and Krosnick 1996). However, low educational background or older age is not always a predictor of order effects (see, e.g., McClendon 1991). Increasing memory limitations can mean that respondents store less contextual information in their working memory and that the order of questions has less effect on evaluations for older respondents (Knäuper, Schwarz, Park, and Fritsch 2007).

Respondent motivation and fatigue is a further relevant moderator. In long questionnaires order effects have been found to be more pronounced for questions placed late in the questionnaire (Holbrook, Krosnick, Moore, and Tourangeau 2007). Bishop and Smith (2001) found no such association, but used a short questionnaire of only 20 questions.

The difficulty of the response task is a further important moderator of order effects. The risk of satisficing is generally hypothesized to be the higher, the more complex, and therefore burdensome, the response task is (Schwarz, Hippler, and Noelle-Neumann 1992: 189). Previous studies have shown that order effects are more pronounced in questions including more sentences, words or letters than other ones (Bishop and Smith 2001; Holbrook, Krosnick, Moore, and Tourangeau 2007; Payne 1949; Schuman and Presser 1981). These studies however rely on comparisons of different questions, which inherently vary not only in their wording, but also in content. To our knowledge, no study so far has examined the effect of task difficulty by experimentally varying difficulty, while keeping other aspects of the task the same.

How important the respondent finds a particular piece of information may also determine whether order effects occur. Some authors suggest that important items or questions are likely to be immune against order effects (Krosnick 1991). For example, in a study by Krosnick (1988) attitudes people considered to be personally important were in general more resilient to context effects. Questions that are important to the respondent may be less prone to order effects, since the relevant information for answering the question is likely to be more easily accessible in memory (reflected in shorter processing times; see Krosnick 1989).

The strength of respondents' attitudes on the topic of evaluation might be a further determinant of order effects (Schwarz 2007; Tourangeau and Rasinski 1988). Strong and previously formed attitudes should in general be more resistant to context influences than attitudes that are formed on the spot (Lavine, Huff, Wagner, and Sweeney 1998: 359). Indifferent respondents, or those with weak attitudes and judgment rules, are more likely to draw on context information that is momentarily salient or accessible and are therefore probably more sensitive to order effects (Hippler and Schwarz 1986; Lavine, Huff, Wagner, and Sweeney 1998). The overall evidence so far is, however, mixed. According to Lavine et al.'s (1988) review of the state of research, only one in seven tests provided evidence for attitude strength significantly moderating the occurrence of question order effects, although their own experiments showed strong evidence of weak attitudes triggering order effects.

The respondent's knowledge on and familiarity with the substantive issue is a further potential moderator of order effects (Bradburn 1992: 321; McClendon 1991; Tourangeau, Rips, and Rasinski 2000). Experts on a topic should need less cognitive effort and might find thinking about the issue more interesting and as a result be influenced less by context information than novices. The more knowledge respondents have on a topic, the less they should therefore be susceptible to order effects. There is some evidence that context effects are indeed less pronounced for respondents with good topic knowledge (Bickart 1992; Smith 1992).

To conclude, there are no straightforward predictions and explanations of order effects and so far, no single theory can predict their occurrence (McClendon 1991). Although each of the moderator variables referred to are related to mechanisms that could plausibly cause order effects, the empirical evidence so far is mixed (see, e.g., Schuman 1992; Schwarz 2007; Smith 1992; Tourangeau 1999 for other reviews). One reason for the stagnation might be that it is necessary to specify more precisely the underlying cognitive mechanisms (memory problems versus context effects versus satisficing). Another explanation could be that some of the mechanisms interact with each other or some moderator variables define necessary preconditions for order effects. For example, respondents' cognitive ability might matter only for very complex evaluations tasks. There have been few attempts to test competing mechanisms or to specify the concrete conditions for order effects. The special design of factorial surveys offers promising opportunities to address these research questions.

Methodological Research on the Design of Factorial Surveys

How the order of dimensions affects results in factorial survey designs has, to our knowledge, not been examined. Previous studies have however examined how the complexity of vignettes affects responses, learning and fatigue effects. These previous studies are relevant to our research in that they provide a useful background for studying conditions under which order effects are more or less likely to occur in factorial survey designs.

Sauer et al. (2011) and Auspurg et al. (2009) experimentally varied the number of dimensions used per vignette, and in some cases additionally varied the number of vignettes presented to each respondent. The results suggested that vignettes consisting of about 8 dimensions were in general well manageable, while vignettes consisting of 12 dimensions produced signs of inconsistent evaluations, especially for respondents with lower education. Learning effects were apparent in the first 10 vignettes: respondents evaluated the vignettes in an increasingly consistent way and with increasing speed. Fatigue or boredom effects became apparent after the 10th vignette: in particular respondents with low educational level gave less consistent responses than other ones. Learning and fatigue effects were more pronounced with very complex vignettes, consisting of 12 rather than 8 or 5 dimensions. Response heuristics have also been found in factorial surveys: when evaluating the first vignettes respondents seemed to take account of a large number of dimensions; when evaluating later vignettes respondents seemed to concentrate on a restricted number of more salient dimensions and to ignore less salient ones (Sauer et al. 2009).

Given the strong evidence of order effects in both general survey research and other experimental survey methods (see below), the lack of research on order effects in factorial surveys is astonishing. One reason why the order of dimensions has not received any attention may be the practice of designing vignettes as running text. In designing text vignettes, researchers aim to place and combine dimensions such that the flow of the text is as natural and smooth as possible. Some dimensions, like the gender of vignette persons or their earnings, may fit more logically at the beginning or end of a text vignette, than somewhere between other dimensions. This implies that rotating the order of vignette dimensions, which could be a way of mitigating order effects, conflicts with the aim of designing smooth vignette texts. This also implies that the optimal order of dimensions in vignette texts may be different in different languages, if they have a different logic of ordering words and phrases within sentences and paragraphs. If the order of dimensions does impact evaluations, this could affect the validity of international comparisons. An alternative to text vignettes is presenting the vignette dimensions in tabular form. Vignette dimensions can easily be rotated in a tabular format and the order is no longer specific to the syntax of a language. Tabular and text vignettes have so far not been contrasted empirically. Our own initial analyses of an experiment related to this study suggest that tabular vignettes produce similar evaluations to text vignettes, when vignettes are not overly complex (consisting of 8 dimensions).

Order Effects in Conjoint Analysis and Choice Experiments

Even though order effects have not received any attention for factorial surveys, order effects have been studied in other related experimental survey methods. Conjoint analysis and choice experiments resemble factorial survey designs in that respondents are asked to evaluate several short descriptions of objects or situations, consisting of dimensions that vary in their levels. The main difference between the three experimental methods is the nature of the response task: in conjoint analyses respondents are typically asked to rank different alternatives (*profile cards*); with choice-experiments respondents have to choose one out of several alternatives that are jointly presented in a *choice-set*; in factorial surveys the respondents are typically asked to evaluate each example case (*vignette*) sequentially on a rating scale. Conjoint analyses are primarily used to assess the utility of product features, and willingness to pay for these, in marketing research (Carrol and Green 1995; Orme 2006). Choice experiments are mainly used in transportation research, health and environmental economics to assess the willingness to pay for public goods, such as transportation services or recreation areas, and in general for objects not (yet) traded in markets (Bennett and Blamey 2001; Louviere, Hensher, and Swait 2000; Ryan, Gerard, and Amaya-Amaya 2008).

Research on the effects of dimension order in conjoint analysis and choice experiments may be informative for factorial surveys, even if the methods differ in the nature of the response answer tasks, fields of application, research aims, and statistical methods used for data analyses. For conjoint analyses there have been several experiments demonstrating the occurrence of order effects. The effects were often large enough to considerably change the relative impact of single dimensions and to alter the estimated monetary values of dimensions. Johnson (1981) varied the dimension order within the profile cards presented to each respondent. He demonstrated in two experiments that respondents' evaluations were less reliable (i.e. less correlated) when the dimension order was changed between the first and last profile card, than when the order of dimensions was fixed (Johnson 1981; Johnson 1989). Johnson's experimental design however not only varied the order of dimensions, but also the complexity of the evaluation task: for one group of respondents the dimension order was fixed; for the other group it varied, possibly making the task more difficult. The effects could therefore in part be the result of respondent confusion about the task. Other studies using between-respondent designs have sometimes found order effects, but not always (e.g. Perrey 1996 found positive evidence, while Orme, Alpert and Christensen 1997 did not). Some of the order effects did not reveal any systematical pattern, while others were in line with primacy or recency effects (e.g. Perrey 1996). Kumar and Gaeth (1991) found familiarity with the object of evaluation to be a strong moderator: order effects only occurred when respondents evaluated unfamiliar products. More systematic research attempting to deepen the theoretical and empirical knowledge about the mechanisms causing order effects in conjoint analysis is lacking. Most applications and methodological guidelines seem to accept the fact that order effects occur and merely attempt to neutralize them by randomizing the order of dimensions.

For choice experiments the situation is similar. Most studies varying the order of dimensions do so to neutralize possible order effects and not to deepen knowledge on the causal mechanisms. All in all the evidence is mixed. In some studies results were not affected by dimension order (Borghans, Romans, and Sauermann 2010; Farrar and Ryan 1999; Olsen, Ladenburg, Petersen, Lopdrup, Hansen, and Dubgaard 2005), while in other studies dimension order had substantial impact on the importance of single dimensions and derived monetary values (Chrzan 1994; Scott and Vick 1999). Where order effects were apparent, their pattern in part corresponded to primacy and recency effects (that is, higher importance of dimensions when placed in the first or last position compared to middle positions; e.g. Glenk 2006; Kjaer, Bech, Gyrd-Hansen, and Hart-Hansen 2006; Scott and Vick 1999), but in part did not conform to any (expected) pattern (e.g. Chrzan 1994).

Two studies on choice-experiments examined the conditions under which order effects occur. Glenk (2006, 2007) examined the role of respondents' educational background, understanding of the response task, and dimension importance in moderating order effects. Surprisingly, order effects were stronger for respondents with higher levels of education, but also for respondents with lower choice task-specific capability (measured by a 5-point self-evaluation rating of how well the respondent had understood the choice-task). Respondents for whom the income dimension was likely to be more important (i.e. respondents with low income), also exhibited more pronounced order effects. Glenk nevertheless concluded that he was "not able to explain in detail why the ordering effects […] may have occurred" (Glenk 2007: 25) and stressed the need of further research.

The second study by Kjaer et al. (2006) focused on a special feature of choice experiments, which is the repeated evaluation of very similar stimuli by each respondent. According to the theory of satisficing respondents are likely to evaluate the choice sets using "rules of thumb" or response heuristics, instead of optimizing their decision rules. For instance, they might employ dominant decisions (also known as "lexicographic" or "non-compensatory" decision making) and solely pay attention to the most important dimensions, considering further dimensions only to differentiate if two choice options have similar utility. While Kjaer et al. (2006) found a recency effect (the price dimension had more impact on evaluations when placed last than first), they did not detect any relationship between the order effect and the use of dominant decision rules. The authors however questioned the generalizability of their results, due to the prominent role of the monetary dimension in choice evaluations.

In sum the studies on conjoint analyses and choice experiments also suggest that the occurrence and magnitude of order effects may be caused by different mechanisms and depend on moderator variables. There are also possible interactions of order effects with response heuristics.

Hypotheses

Theory and empirical research suggest the existence of order effects. When questions are presented visually, as they are in factorial surveys, cognitive psychology and empirical evidence further predict that primacy effects are more likely than recency effects. Hence we derive the following two main hypotheses:

H_{1a}: The order in which vignette dimensions are presented influences their impact on vignette evaluations.

H_{1b}: Dimensions have a larger impact on the vignette evaluations when they are placed in the first position than when they are placed in a middle position.

We further examine under which conditions order effects are more or less likely to occur. The practical implications differ, depending on whether the magnitude of order effects mainly depends on questionnaire design or respondent characteristics. Therefore we organize our analyses into hypotheses that are primarily related to the design of the factorial survey, hypotheses related to respondent characteristics, and hypotheses about how different moderator variables interact to produce order effects.

Impact of the Complexity of Factorial Survey Modules

Theory and empirical findings suggest that order effects may be caused by some combination of memory limitations and respondent satisficing, both of which are more likely the more complex an evaluation task is. Methodological research on factorial surveys has shown that respondents are able to evaluate vignettes with 8 dimensions consistently. When the number of dimensions is increased to 12, there are signs of cognitive overburdening. Therefore we expect the following:

H_{2a}: Order effects are more pronounced when vignettes consist of 12 rather than 8 dimensions.

Vignettes consisting of different numbers of dimensions inevitably differ not only in their complexity, but also in the substantial information presented to respondents. To verify that it is really the complexity of the evaluation task that matters, and not just the additional information, we examine further aspects. Survey research proposes that complexity varies with the nature of the response format. Rating closed answer scales has been shown to be an easier task than, for example, answering open questions about fair amounts of income or willingness to pay (e.g. Bijlenga, Bonsel, and Birnie 2011). Similarly, answering two instead of one target question for each vignette is likely to be a more complex task. We therefore expect that:

H_{2b}: Order effects are more pronounced when respondents are asked two instead of one target questions about each vignette.

The vignette evaluation task is also more cognitively demanding the more vignettes each respondent has to evaluate. Previous research has shown that fatigue or boredom effects occur after the respondent has evaluated about 10 vignettes. After this point respondents appear to focus on the most salient dimensions and to ignore other less salient ones when computing their evaluations. This is consistent with the idea that respondents develop response heuristics in the course of evaluating a series of vignettes. It is difficult to predict how possible response heuristics and order effects might interact. Presumably, the resulting response patterns depend on which dimensions enter into response heuristics and if these dimensions are *per se* more or less prone to order effects. For example, if respondents focus increasingly on the most important dimensions when evaluating a series of vignettes, and only important dimensions are prone to order effects, then the overall order effects will be stronger for later than earlier vignettes. Similar effects might be true for fatigue leading to less concentrated response behavior that is in general more prone to order effects. In addition, the order of dimensions may itself influence which dimensions are most important to respondents. In this case, not only would some dimensions gain importance during the sequence of vignettes, but the extent of order effects for these specific variables would also increase.

In sum, the current state of research does not enable clear predictions about the likely nature of interactions between order effects and response heuristics or fatigue effects.

Nonetheless we expect the magnitude of order effects to vary during the course of vignette evaluations and in particular to change after about 10 vignettes have been evaluated:

H_{2c}: The magnitude of order effects depends on the sequential position of a vignette and in particular changes after about 10 vignettes have been evaluated.

Impact of Respondent Characteristics

One of the main predictions from cognitive psychology is that order effects are more pronounced the less cognitively sophisticated respondents are. Furthermore, the extent of order effects seems to depend on the strength of attitudes and on how familiar respondents are with the survey topic. We therefore expect the following:

H_{3a}: The more cognitively able respondents are, the less we expect the order of dimensions to affect evaluations.

H_{3b}: The stronger the attitudes are that respondents have on the topic of the evaluation task, the less we expect the order of dimensions to affect evaluations.

H_{3c}: The more knowledge respondents have on the topic of the vignettes, the less we expect the order of dimensions to affect evaluations.

Additionally, the survey literature suggests that the magnitude of order effects depends on whether questions or items are personally important to respondents. Correspondingly, one might expect order effects within factorial surveys to be moderated by the extent to which respondents consider single dimensions to be essential for the evaluation task. We therefore expect:

H_{3d}: Dimensions that are personally important to the respondents are less prone to order effects than other dimensions.

Interaction of Moderator Variables

Finally, there is evidence from prior research that single moderator variables interact with each other. The theory of satisficing predicts an interaction of respondents' cognitive ability with the task difficulty, and other interactions have sometimes been demonstrated or supposed. The current state of research only allows a very general prediction:

H_4: Order effects are more pronounced the more conditions apply: the response task is complex, respondents have low cognitive ability, weak attitudes, or only little knowledge of the substantive topic.

Survey and Experimental Design

Sample, Survey Mode and Questionnaire Content

Studies testing for order effects have mostly employed data from very heterogeneous respondent samples. With heterogeneous samples socio-demographic characteristics such as age and education can be used as proxies for cognitive ability. The socio-demographic characteristics might however also be indicative of other constructs such as selection criteria for educational tracking (Krosnick 1989: 206). Even more problematic, age cohorts and educational levels are certainly related to respondents' opinions and attitudes. Thus, it is difficult to disentangle the "true" opinions and attitudes the researcher is interested in and methodological effects that might invalidate their measurement. The strong interrelatedness of socio-demographics with both attitudes and moderator variables involved in the causation of order effects might explain why existing research has so far failed to detect clear causal patterns. To overcome these problems we used a more homogeneous sample. Social science students at three German universities were recruited for an online survey in 2009. The topic of the survey was the fairness of earnings, which is probably the most common subject of factorial survey designs (see Alves and Rossi 1978; Hermkens and Boerman 1989; Jasso and Rossi 1977; Jasso and Webster Jr. 1997; Jasso and Webster Jr. 1999; Shepelak and Alwin 1986 for some applications). The questionnaire contained a series of socio-demographic questions, the factorial survey module, some items about attitudes related to justice and the importance of single dimensions for fair earnings, attitude strength, knowledge about the earnings distribution in Germany and further socio-demographic questions.[1]

Factorial Survey Module and Experimental Design

Within the factorial survey module all respondents had to evaluate 20 different vignettes describing fictive employees. The employees were characterized by 8 or 12 variable dimensions (including gross earnings) that were known to influence justice evaluations from prior factorial surveys, justice and labor market research (see Appendix Table A1 for the definition of dimensions and levels). Respondents had to rate each vignette, evaluating how fair they thought the gross earnings were. Within the vignettes descriptions of the occupations (e.g. ‘medical doctor’, ‘hairdresser’) were used, but for the analyses occupational prestige scores were employed (the magnitude prestige scale, MPS, see Christoph 2005; as it is common in factorial survey designs).[1] The gender of the vignette person was also signaled by the occupation, since for female vignette persons the feminine forms were used (for instance “Ärztin” instead of “Arzt”; only for the clerk the masculine and feminine forms are the same in German).

As the combination of all dimension levels gave rise to more than one million possible combinations, we generated a D-efficient sample of 240 vignettes excluding illogical cases (e.g. medical doctors without a university degree). D-efficient samples optimize two desirable characteristics of experimental designs: maximal ‘orthogonality’ (i.e., a minimal intercorrelation between all single dimensions) and maximal variance and ‘level balance’ of the single dimensions (i.e., all dimension levels occur with about the same frequency). Both design features allow estimating the influence of single dimensions with maximum precision and therefore offer maximum statistical power to detect their impact on evaluations (for details: Kuhfeld 2009; Kuhfeld, Tobias, and Garrat 1994; Steiner and Atzmüller 2006). The vignette sample selected had a D-efficiency measurement of 90.7. In a second step we applied the two target criteria for D-efficient designs to the 240 vignettes to combine them into 12 sets of 20 vignettes. The sets of vignettes were then randomly allocated to respondents. This means that all experimental splits were based on the same sample of vignettes, which guaranteed that correlations and variances of vignette dimensions were similar for all experimental treatments. For respondents allocated to the 8-dimension version, the dimensions health status, firm size, firm success and job performance were simply deleted.

The order of the 20 vignettes was randomized for each respondent. This randomization neutralized possible effects of vignette order and facilitated the identification of learning and fatigue effects. The dimensions were listed in tabular format, since this allowed a more flexible variation of the dimension order than presenting them as running text.

To test the methodological research questions we used a 2x2x2 multi-factorial experimental design. Two different dimension orders (order 1 and order 2) were crossed with two versions of vignette complexity (8 and 12 dimensions) and two versions of evaluation complexity (one and two target questions about the vignette).

The table containing the vignette dimensions was always split into two columns for convenience and clarity. Therefore, four dimensions appeared either in first or last position and might have been especially eye-catching (see Table 1).

In order 1 the vignette dimensions were sorted in a similar way as the text vignettes used in a parallel study. This was a typical order of vignette dimensions and is summarized in Table 1: the vignettes started with gender and age, followed by dimensions describing the education level, employment characteristics and socio-demographic background of the vignette person. The last cell in the table contained the gross earnings.

Table 1: Order of dimensions

12 Dimensions

Order1	
Sex	Health status
Age	Tenure
Education	Firm size
Children	Firm success
Occupation	Job performance
Experience	Gross earnings

Order 2	
Education	Job performance
Occupation	Age
Experience	Sex
Tenure	Children
Firm size	Health status
Firm success	Gross earnings

8 Dimensions

Order 1	
Sex	Occupation
Age	Experience
Education	Tenure
Children	Gross earnings

Order 2	
Education	Age
Occupation	Sex
Experience	Children
Tenure	Gross earnings

Within order 2, first all dimensions about employment related characteristics of the vignette person were listed, followed by dimensions describing the socio-demographic background of the vignette person. Only the earnings dimension was again listed in the last cell, since it seemed more natural to end the vignettes with this dimension and since it was expected to have an extraordinarily high impact on the justice evaluations regardless of its position. Order 2 was also chosen to avoid confounding the position of dimensions with their importance. Both dimensions known to be of more or less importance, were placed in first or last row in one order, and in the middle in the other order condition.

Two further experimental splits were used that varied the complexity of the vignettes and the complexity of the evaluation task. To increase the complexity of the vignettes, 12 rather than the more common 8 dimensions (Sauer, Auspurg, Hinz, and Liebig 2011) were used per vignette. The additional 4 dimensions were chosen to reflect characteristics that were already well known to have substantial impact on justice evaluations (i.e. health status, firm size, firm success and job performance). This was done to ensure that the added information was indeed of relevance to the respondents.

The complexity of the evaluation task was manipulated by using one versus two target questions about the vignette. The first target question asked respondents to evaluate the fairness of gross earnings: “Are the monthly gross earnings of this person fair or are they, in your view, unfairly high or low?”, using an 11-point rating scale ranging from -5 ”far too low” to 0 “fair” to +5 “far too high”. For our analyses we use only this rating scale to assess the robustness of vignette evaluations to order effects. The second target question was added for a random half of respondents. Respondents who had rated the earnings as “unfair” were asked an open question about what they thought a fair earnings would be: “If you rated the earnings as unfair, what do you think would be a fair amount of earnings for the described person?” All experimental splits were fully crossed (full orthogonal design) and each allocated to about half of respondents. That is, we employed a between respondent design, whereby the 20 vignettes presented to a respondent were all of the same experimental design.

Table 2 documents the sample sizes for respondents and vignette evaluations. For each of the eight experimental cells at least 30 respondents and about 600 to 700 vignette evaluations were achieved. The eight experimental groups were balanced in terms of case numbers and respondent characteristics, including sex, degree studied for, partnership status, location, and average time studying (see Appendix Table A2), although there were some differences in mean ages and income levels between groups. Only few students had children or were born outside Germany, but those that did were distributed across treatment groups. The balanced numbers and characteristics confirm that the random allocations to treatment groups were successfully implemented.

Table 2: Number of respondents, scale and open evaluations per treatment group

Order of dimensions	Number of evaluations	8 Dimensions			12 Dimensions		
		Respondents	Scale	Open	Respondents	Scale	Open
1	1	38	754	—	33	657	—
	2	31	613	529	37	730	622
2	1	38	758	—	38	760	—
	2	37	732	646	30	591	586
Total		144	2857	1175	138	2738	1208

Notes: Missing evaluations excluded. All respondents answered at least one vignette.

Operationalization of Respondent Characteristics

All other constructs we use to test our hypotheses were measured by item-questions that followed the factorial survey module. These were:

- *Cognitive ability*: Respondents were asked to indicate whether their performance at university was above or below average: “What do you think, how is your performance at university?” using an 11-point rating scale ranging from -5 “below average” to 0 “average” to +5 “above average”. For our analyses we classified students as either “high” or “low” ability based on a median-split. We used a self-assessment instead of actual grades since performance measurements based on grades are hardly standardized across Germany universities (even within the field of social sciences: Müller-Benedict and Tsarouha 2011).

- *Strength of attitudes*: A sub-item belonging to the scale “justice ideologies” (Stark, Liebig, and Wegener 2008) was used as a proxy for the extent to which students have previously and strongly fixed norms in regard to their justice evaluations. Respondents were asked to rate the statement “The way things are these days, it is hard to know what is just anymore” on a 5-point rating scale. Respondent who either “strongly” or “somewhat” agreed with the statement were classified as having weak attitudes (36%); respondents who neither agreed nor disagreed, or disagreed (somewhat or strongly) were classified as having strong attitudes (64%).[2]

- *Knowledge of the subjective matter*: Respondents were asked an open question about mean gross earnings in Germany to assess their knowledge on the subject matter of fair earnings: “What do you think is the average monthly gross salary for full-time employees in Germany?” The actual value was 3141 Euros per month in 2009 (Statistisches Bundesamt 2010). 50% of respondents indicated amounts within 641 Euros from the true value, and were coded as having “good knowledge”. The remaining 50% indicated amounts outside this range or refused to answer the question and were classified as having “little knowledge”.

- *Importance respondents ascribe to each vignette dimension*: Towards the end of the questionnaire, respondents were asked to rate how much impact each of the vignette dimensions should have in order to achieve a fair distribution of earnings: “In your opinion, what impact should the following items have for fair levels of gross earnings? – Age of employees, gender, …”. The response scale was a 7-point rating scale ranging

> from 0 “no impact at all” to 6 “very large impact”. To test our hypotheses we were interested in which dimensions respondents personally considered more or less important. Therefore, we determined each respondent’s median rating across all dimensions, and then classified each dimension as being important to the respondent if the valuation was above the respondent’s median, and unimportant if the dimension was rated below or equal to the respondent’s median.[3]

In addition, after completing the factorial survey module respondents were asked to assess the overall complexity of the evaluation task: “All in all, how easy or difficult was the evaluation of these exemplary cases for you?” using an 11-point rating scale ranging from -5 “very difficult” to +5 “very easy”. We used a median-split to classify respondents as either finding the task difficult (59%) or easy (41%). Answers to this question were expected to be strongly influenced by respondents’ cognitive abilities, strength of attitudes, and knowledge since respondents short of at least one of these features should have relatively high difficulty with the evaluation task. We use this measurement as an additional possibility to assess the overall impact of respondents’ characteristics for moderating dimension order effects.

The correlations between respondent characteristics were low: *r* = .06 for strength of attitude and ability, *r* = .06 for knowledge and ability, and *r* = .08 for strength of attitude and knowledge. Even if the single constructs are partially related to each other, they clearly measure different things. The respondents’ rating of how difficult they found the vignette task was not correlated with knowledge (*r* = -.06) or strength of attitude (*r* = -.02), but was negatively correlated with ability (*r* = -.16, *p* = .007). Table A3 in the Appendix documents the number of closed evaluations achieved for the different combinations of respondent characteristics, separating respondents allocated to the 8 and 12 dimension treatments.

Results

Does the Order of Vignette Dimensions Matter?

Impact on Absolute Effect Sizes. Table 3 shows the results of Ordinary Least Squares (OLS) regression models estimating the influence of vignette dimensions on the vignette evaluation for the 12 dimension condition. As the data have a hierarchical structure (several evaluations stem from single respondents; see Hox, Kreft, and Hermkens 1991 for details) we estimated robust standard errors. The first two columns report regression coefficients and standard errors from separate models for the order 1 and order 2 conditions. Positive (negative) coefficients denote that the earnings of the vignette person were judged unfairly too high (low). Results are all in all plausible and in line with justice theories. For example, persons with university degree should have higher earnings than persons without any vocational degree (β = -0.625 in order 1; β = -0.612 in order 2); or the more children a person has, the higher their earnings should be (β = -0.074 in order 1; β = -0.099 in order 2). Earnings of the vignette person were integrated as a logarithmic variable to model the non-linear relationship with fairness evaluations.

The order of dimensions mattered for some estimates. The dimensions ‘age of vignette person’, ‘experience’, and ‘risk of bankruptcy’ only reached statistical significance (5%-level) in the order 1 condition. Because the case numbers were similar for both models and the vignette sample was exactly the same, the only explanation for these differences apart from sampling variation are order effects. Other dimensions had stronger effects in the order 2 than order 1 condition. For example, the absolute value for the regression coefficient of ‘performance above average’ was about 1.5 times larger in the order 2 than order 1 condition. Other coefficients even changed signs, for instance those for age, tenure and firm size.

To test whether the differences caused by the order of dimensions are statistically significant, we estimated a joint model of order 1 and order 2 vignettes, including interaction terms of each dimension with a binary indicator of the order, and tested the null-hypothesis H_0 that the interaction terms were all jointly zero applying a Wald test (this 'omnibus' hypotheses test that there are no differences at all is also known as 'Chow test', for details: Wooldridge 2003). The result suggests that the order of dimensions does matter and that evaluations differed across order 1 and order 2 vignettes (*F* = 2.06; *p* = .013). We further tested for order effects of individual vignette dimensions by employing Wald tests for the interactions of single dimensions with the order indicator. For dimensions with 3 categorical values, the results are from joint tests of the interactions for both dummy variables included in the model. The *F*- and *p*-values of these Wald tests are displayed in the column labeled ‘Interactions” in Table 3. The order effects were statistically significant for three dimensions (‘age’, ‘experience’, and ‘long tenure’), that is, for a quarter of all vignette dimensions. Thus, we find support for the first hypothesis: The order in which vignette dimensions are presented influences their impact on vignette evaluations (H_{1a}).

There was no clear evidence that dimensions have more impact on evaluations when they are placed in first position. For ‘gender’ the results suggested weak primacy effects: the regression coefficient was larger in order 1 where the dimension was in first position, than in order 2. Neither coefficient was however statistically significant, and neither was the interaction of gender with order in the pooled model. For ‘education’ there were no differences in coefficients or significance levels between order 2, where it was in first position, and order 1. For the other dimensions which appeared in the first or last cell of a column in either order there were also no clear patterns. Sometimes the absolute coefficient values were larger when dimensions were placed in first or last position, sometimes not. Significant order effects only appeared for dimensions placed in middle positions. Hence, we do not find support for the hypothesized primacy effects (H_{1b}).

Impact on Relative Importance. One of the main advantages of factorial survey designs is that they reveal information on the relative importance of dimensions. Respondents are forced to trade-off the impact of different dimensions in their evaluations. Usually semi-partial R^2-values, that measure the proportion of variance explained by each dimension, are used to draw conclusions about the relative importance of different dimensions (Wallander 2009).[4]

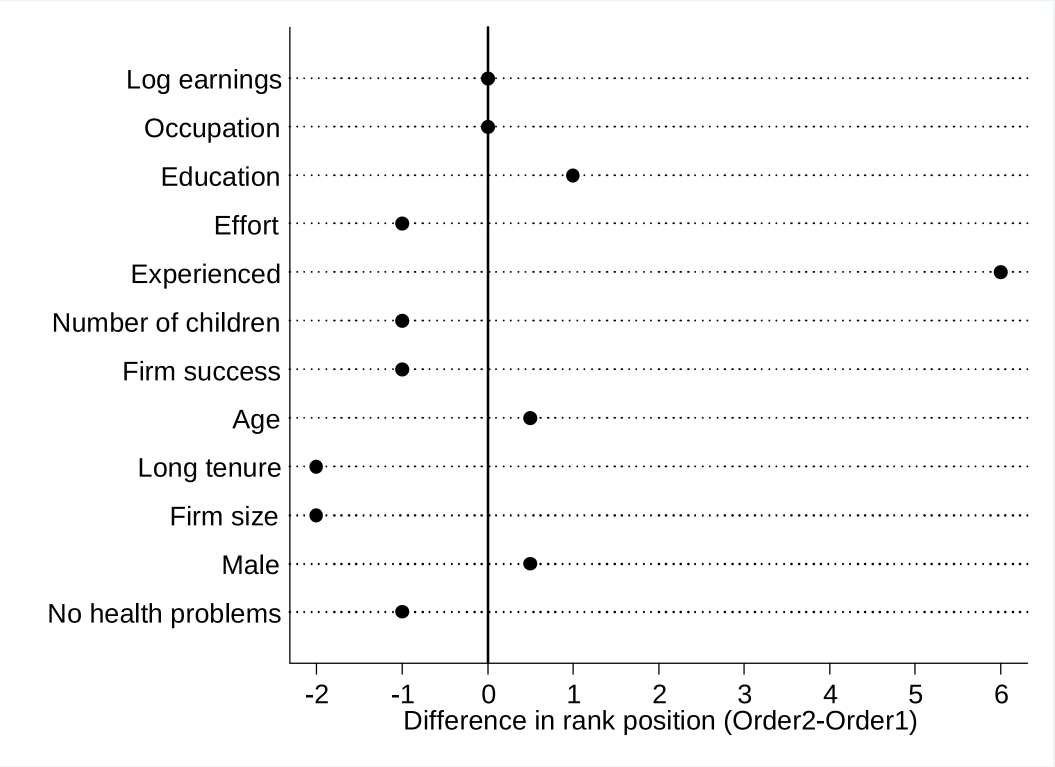
The last two columns in Table 3 report the semi-partial R^2-value for each dimension based on separate models for order 1 and order 2. Figure 1 illustrates the difference in rank position of each dimension between order 1 and order 2 (using mean values for ties). The dimensions are listed from top to bottom on the y-axis, according to their rank position in order 1. The plots show the difference in rank position between order 2 and order 1. The rank correlation (Spearman’s ρ) between the semi-partial R^2-values of order 1 and order 2 is .83 (p = .001). In both orders the dimension ‘earning’ was by far the most important, followed by the dimension ‘occupation’. For the other dimensions order effects were large enough to change the relative importance: for seven dimensions the relative importance differed by maximally one rank position; for two dimensions (firm size and tenure) it differed by two positions; for the dimension ‘experience’ the difference amounted to six rank positions (position 5 versus 11). These findings provide further support for the hypotheses that the order in which dimensions are presented influences their effect (H_{1a}). Where the dimensions were placed revealed no clear pattern, again not providing any support for the hypothesized primacy effects (H_{1b}).

Table 3: Effect of order on coefficients, standard errors and semi-partial R^2-Values

	Regression coefficients (standard errors)		Interactions		Semi-partial R2	
	Order 1	Order 2	*F*	*p*	Order 1	Order 2
Male	-0.157	-0.017	0.99	.3220	.0008	.0000
[Female]	(0.0955)	(0.1045)				
Age (10 years)	-0.008*	0.006	6.67	.0108	.0011	.0005
	(0.0037)	(0.0039)				
Vocational training	-0.260*	-0.273**	0.02	.9846	.0094	.0090
	(0.1061)	(0.0991)				
University degree	-0.625***	-0.612***				
[No vocational training]	(0.1232)	(0.1347)				
Number of children	-0.074*	-0.099**	0.26	.6121	.0016	.0028
	(0.0366)	(0.0344)				
Occupation	-0.013***	-0.017***	2.66	.1052	.0012	.0490
(10 MPS-Scores)	(0.0013)	(0.0016)				
Experienced	-0.376**	0.001	5.00	.0270	.0027	.0000
[Litte experience]	(0.1178)	(0.1214)				
No health problems	-0.050	-0.035	0.01	.9061	.0001	.0000
[Long-term problems]	(0.0956)	(0.0896)				
Long tenure	0.211*	-0.306*	11.71	.0008	.0010	.0020
[Short tenure]	(0.0874)	(0.1240)				
Medium sized firm	-0.123	0.027	1.96	.1450	.0008	.0005
	(0.1114)	(0.1160)				
Large firm	-0.144	0.157				
[Small firm]	(0.0995)	(0.1179)				
Risk of bankruptcy	0.192*	0.121	0.35	.7039	.0014	.0021
	(0.0888)	(0.1159)				
High profit	-0.174*	-0.293*				
[Solid]	(0.0860)	(0.1130)				
Performance below average	0.441***	0.486***	0.76	.4685	.0062	.0093
	(0.0820)	(0.1164)				
Performance above average	-0.294**	-0.443***				
[Performance average]	(0.1020)	(0.1043)				
Log earnings	2.266***	2.307***	0.09	.7701	.5237	.5309
	(0.0936)	(0.1039)				
Constant	-15.704***	-16.608***				
	(0.7371)	(0.7450)				
N	1387	1351	2738		1387	1351
R^2	0.6050	0.6061	0.6057			

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. The *F*- and *p*-values are from Wald tests of the interaction of each variable with the order of dimensions. For variables with more than two categories the statistics are from joint tests of all interactions related to the variable. Degrees of freedom (1, 137) for all tests of two-category variables, and (2, 137) for all tests of three category variables. Estimations are only based on the 12 dimension split. MPS = Magnitude Prestige Scale.

Figure 1: Impact of dimension order on relative dimension importance



Notes: Dimensions are listed from top to bottom according to their rank position in order 1. Mean rank position used for ties.

Practical Relevance. A common target measure in factorial surveys is the fair amount of earnings derived from the vignette evaluations. Some applications estimate fair returns to different levels of education or fair pay gaps between equally qualified men and women ('just gender pay gaps', shortened JGPGs; see, e.g., Jasso and Rossi 1977; Jasso and Webster Jr. 1999 for applications).The intuitive logic is to ask which amount of earnings one group has to be paid more or less, compared to the reference group, for the mean fairness evaluation to be the same across both groups. To estimate the fair differences between groups the regression estimates are used to calculate the trade-offs between single dimensions. When using a binary indicator for the groups compared and a logarithmic transformation of earnings the just pay gap in percent (%JPG) can be calculated from the following formula (see the technical appendix for details):

$$\%\text{JPG} = (\exp\,(\frac{-\beta_{groupvariable}}{\beta_{earnings}}) - 1) * 100 \qquad (1)$$

with $\exp(\cdot)$ denoting the exponential function and β the regression coefficients from the OLS regression of the vignette evaluations on all vignette dimensions. In both order conditions there is evidence of a JGPG favoring male vignette persons: in order 1 respondents evaluated an earnings gap of 341 Euros between men and women as fair (which is equivalent to 7.2% of mean vignette earnings). In order 2 the JGPG is much smaller at only 35 Euros (0.7%). With order 2 vignettes one would conclude that there is no substantial JGPG, while the magnitude of the JGPG with order 1 vignettes is considerable. Note that the differences between male and female vignette persons were not caused by differences in their labor market or demographic characteristics. Due to the D-efficient vignette design, male and female vignette persons were on average described as having the same characteristics. For just returns to education, the order effects are smaller: the just return to a vocational qualification (compared to no vocational qualification) amounts to 579 Euros per month with order 1 vignettes (12.2%) and 605 Euros (12.6%) with order 2 vignettes. For a university degree, the estimated just rates of return are 1511 Euros (31.8%) respectively 1459 Euros (30.4%).

For the JGPG the results suggest that order effects can have large practical implications for the conclusions drawn from vignette evaluations. Since the JPGs are based on two dimensions, already small differences in the impact of single dimensions can substantially affect conclusions.

Which Conditions Trigger Order Effects?

Effects of Factorial Survey Design. To test under which conditions order effects are more likely to occur, we again estimated regression models pooling the evaluations for order 1 and order 2 vignettes, including interaction terms for each dimension with the order indicator. The models were estimated separately for the less and more complex vignette conditions. For each model we used Wald tests to assess whether the interaction terms are jointly zero, i.e. testing the null hypotheses of no order effects. In the results reported here we did not include the main effect of order in the joint tests, since most applications of vignette studies focus on the impact of the dimensions and not the absolute level of evaluations. Estimates including the main effect of order however produced very similar results. When the vignette complexity conditions resemble each other in their degrees of freedom (numbers of observations), the *F*-values resulting from the Wald tests can be used to compare the magnitude of order effects across models: the higher the *F*-value, the larger the difference in the impact of dimensions between the order groups. In other cases the significance levels (*p*-values) give some indication under which conditions order effects occur.

Table 4 shows the results separately for the less complex (8 dimensions) and more complex vignettes (12 dimensions). As expected (H_{2a}), order effects were in general more pronounced with more complex vignettes (*F* = 0.82; *p* = .594 for 8 dimensions and *F* = 2.06; *p* = .013 for 12 dimensions). When complexity was added by asking two instead of one target question, order effects appeared also in the 8 dimension condition (*F* = 2.03; *p* = .049). This provides evidence for the hypothesis that order effects are more pronounced when the nature of the target question is more complex (H_{2b}). For the already complex vignettes using 12 dimensions, adding further complexity through the second target question did not make any difference.

To analyze how order effects change in the course of evaluating the sequential vignettes, we pooled vignettes in positions 1-5, 6-10, 11-15 and 16-20, as there were too few vignettes to produce stable results for individual vignette positions. Both with 8 and 12 dimensions order effects were most pronounced in the first and last 5 vignettes evaluated (Table 4). For the middle 10 vignettes there were no significant order effects. This provides some support for the hypothesis that the magnitude of order effects depends on learning and fatigue effects (H_{2c}).[5]

Table 4: Impact of design characteristics on strength of order effects

	8 Dimensions			12 Dimensions		
	N	*F*	*p*	*N*	*F*	*p*
All	2857	0.82	.5943	2738	2.06	.0134
1 Evaluation	1512	0.80	.6160	1417	2.51	.0044
2 Evaluations	1345	2.03	.0486	1321	1.92	.0337
Vignettes 1-5	717	2.10	.0329	689	1.76	.0424
Vignettes 6-10	716	0.87	.5501	684	1.10	.3612
Vignettes 11-15	714	1.00	.4438	687	1.29	.2138
Vignettes 16-20	710	1.53	.1411	678	1.97	.0194

Notes: Test statistics from separate OLS models. *N* = number of observations, *F*-statistics and *p*-values from joint Wald tests of all interactions.

Effects of Respondent Characteristics. Table 5 shows the results of Wald tests of the joint significance of all dimensions interacted with order, estimated separately for the 8 and 12 dimension conditions and for respondents sharing different characteristics. With 8 dimensions there are no significant order effects at all. The 12 vignette condition shows results that are mostly in line with our hypotheses: order effects were more pronounced when respondents had little knowledge of the substantive matter (H_{3c}) and when respondents had only a comparatively weak attitude regarding the evaluation task (H_{3b}). Contrary to our expectations order effects were not stronger for respondents with lower cognitive ability (H_{3a}). The last two rows in Table 5 indicate the strength of order effects separately by the respondents' own ratings of how difficult they found the task of evaluating the vignettes. With 12 dimensions the respondents' self-assessment is predictive of order effects. While respondents who thought evaluating the vignettes was a relatively easy task showed no significant ordering effects (*F* = 1.13; *p* = .355), respondents who found the task difficult produced highly significant order effects (*F* = 2.75; *p* = .002). These findings provide support for the hypothesis that moderators interact to produce order effects and that order effects are more pronounced if several conditions are met (H_4). A minimal level of complexity of the vignettes seems to be one necessary precondition. Whether a complex vignette design leads to pronounced order effects however depends on the respondents' knowledge and strength of attitudes regarding the evaluation task.

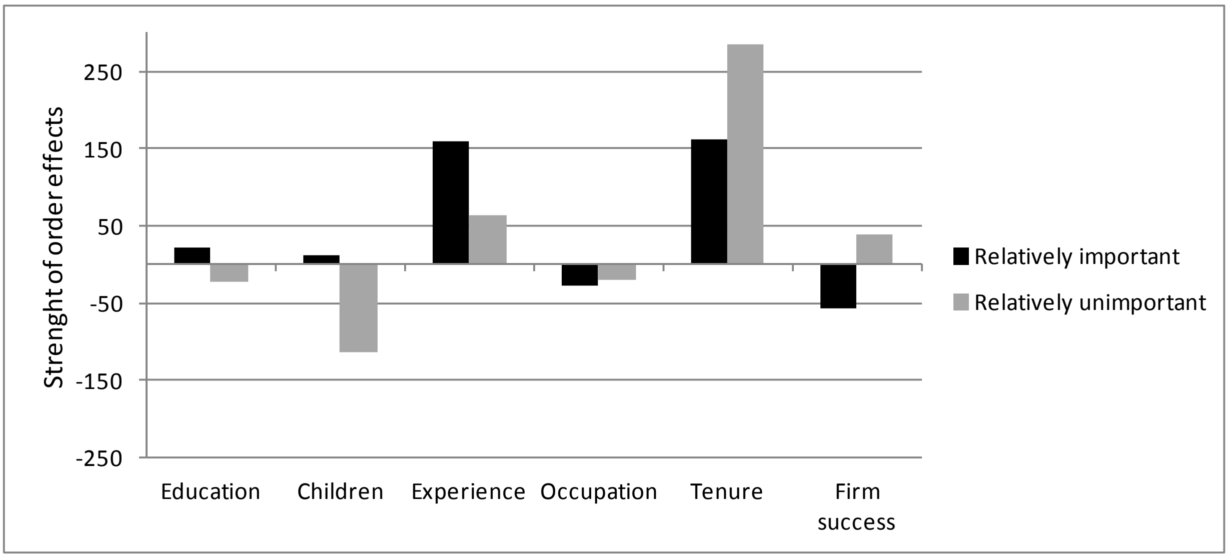
Table 5: Impact of respondent characteristics on strength of order effects

	8 Dimensions			12 Dimensions		
	N	*F*	*p*	*N*	*F*	*p*
Little knowledge	1290	0.58	.8064	1505	2.13	.0152
Good knowledge	1567	0.68	.7276	1233	1.79	.0531
Weak attitude	992	1.29	.2652	997	2.62	.0050
Strong attitude	1825	0.93	.5000	1721	1.86	.0360
Low ability	1342	0.78	.6334	1112	1.40	.1773
High ability	1475	1.34	.2303	1626	1.54	.1065
Difficult task	1600	1.22	.2939	1646	2.75	.0015
Easy task	1217	1.07	.3979	1092	1.13	.3547

Notes: Test statistics from separate OLS models. *N* = number of observations, *F*-statistics and *p*-values from joint Wald tests of all interactions.

Impact of Dimension Importance. The analysis of semi-partial R^2-values (Figure 1) provides some initial evidence that the importance of dimensions matters: the most important dimensions seemed robust against order effects. Dimension importance was however derived as an aggregate measure across all respondents, whereas the literature suggests that what matters is how important a dimension is personally to a respondent. For each dimension we therefore estimated the strength of order effects separately for respondents who rated that particular dimension as relatively unimportant or important. For both respondents groups we measured the strength of order effects by estimating the impact of each dimension on vignette evaluations (i.e. the OLS coefficients) for order 1 and 2 vignettes separately, and computing the percentage difference between the two estimates. Some dimensions were rated unimportant (e.g. sex) or important (e.g. performance) by nearly all respondents. For these dimensions the case numbers were too small to estimate stable regression coefficients for both groups of respondents and orders of dimensions. We therefore analyzed the six dimensions for which there was enough variance between respondents in the importance they attached to the dimension. For four of those dimensions the order effects were slightly larger for respondents who rated the dimension as relatively important (black bars in Figure 2). For the dimensions 'children' and 'tenure', however, the opposite was true: the order effects were larger for respondents who thought children or tenure were relatively unimportant (grey bars). None of the differences in order effects between respondents who thought the dimension was relatively important or unimportant were statistically significant (tested by estimating the three-way interaction of dimension x personal relative importance of that dimension x dimension order). All in all the results do not suggest a clear pattern and do not provide support for the hypothesis that dimensions that are personally important to respondents are less prone to order effects (H_{3d}).[6]

Figure 2: Strength of order effects by dimension importance



Notes: Strength of order effects was measured as the percentage difference in OLS regression coefficients (i.e. the impact of a dimension on the fairness evaluation) in order 1 and order 2: $((\beta_{\text{order2}} - \beta_{\text{order1}}) / \beta_{\text{order1}})*100$. The black bars are estimated for respondents, for whom the given dimension was relatively important, the grey bars for respondents for whom the dimension was relatively unimportant.

Interactions of Respondent Characteristics

The results above suggest that complexity of vignettes is a necessary condition for order effects. We further tested whether any of the risk factors on the part of respondents are also necessary conditions, and whether they interact at all. Table 6 displays the *F*- and *p*- values from joint Wald tests of the interactions of each dimension with dimension order, estimated separately for all combinations of respondents. That is, for respondents with low or high ability, little or good knowledge, and weak or strong attitudes, using the 12 dimension condition.

For all eight combinations of respondent characteristics there were significant order effects, except for the combination of low ability, little knowledge, and strong attitudes (*F* = 1.33; *p* = .271). We therefore conclude that none of the respondent characteristics is a necessary condition for order effects. Weak attitudes however represented a strong risk factor: for all combinations of respondent ability and knowledge the order effects tended to be stronger for respondents with weak compared to strong attitudes. Additionally there were some signs that low ability and weak attitudes interact with each other (see the extraordinarily high *F*- and low *p*-values for low ability and weak attitudes: *F* = 14.82; *p* = .001 for respondents with little knowledge resp. *F* = 94.66; *p* = .000 for respondents with more knowledge). The results therefore provide partial support for the hypothesis that order effects are more pronounced when more conditions apply (H_4).

Table 6: Interaction of respondent characteristics in triggering order effects

	Low ability			High ability			All		
	N	*F*	*p*	*N*	*F*	*p*	*N*	*F*	*p*
Little knowledge									
Weak attitudes	179	14.82	.0005	280	5.32	0.0027	459	6.04	.0001
Strong attitudes	415	1.33	.2709	611	5.10	0.0001	1026	2.01	.0302
More knowledge									
Weak attitudes	298	94.66	.0000	240	7.97	0.0010	538	2.93	.0073
Strong attitudes	220	4.54	.0135	475	6.78	0.0000	695	2.28	.0215

Notes: Estimations are based on the 12 dimension condition. Test statistics from separate OLS models.
N = number of observations, *F*-statistics and *p*-values from joint Wald tests of all interactions.

Robustness Check

As a robustness test for our results we verified that the effect of vignette complexity was not merely an artifact of the larger number of statistical tests performed with 12 rather than 8 dimensions. The dimensions for which order mattered according to Table 3 were all part of the core 8 dimensions – which suggests that the higher sensitivity to order effects in the 12 dimension condition was not merely caused by the larger number of dimensions tested (and therefore the increased risk of Type I errors in significance testing). We additionally re-estimated Tables 4, 5 and 6 using OLS regressions predicting fairness of earnings for the 12 dimension group, but testing the null hypothesis of no order effects (i.e. the joint Wald test of dimensions and their interaction with the order indicator) using only the core 8 dimensions to make the test (and degrees of freedom) more comparable to the 8 dimension condition. The results were comparable to those reported here, suggesting that the impact of vignette complexity is not a statistical artifact.

Conclusions

This paper examines whether the order of vignette dimensions affects research conclusions from factorial surveys – and under which conditions order effects are likely. We make several contributions. To our knowledge this is the first study of order effects in surveys that examines the role of task difficulty by experimentally varying difficulty. Previous studies (e.g. Holbrook, Krosnick, Moore, and Tourangeau 2007; Malhotra 2009) have relied on comparisons of different questions that varied in their content as well as difficulty.

This is also one of few studies that systematically examine the interaction of several risk factors, by examining how characteristics of the survey design and respondents interact to produce order effects. Although we study order effects in a factorial survey, the results are relevant to other vignette-based methods such as conjoint analysis and choice experiments and extend previous studies on those methods. The risk of order effects has so far been ignored for factorial surveys and our results have practical implications, which we discuss below.

The results *first* show that the order in which vignette dimensions are presented can affect research conclusions from factorial surveys. For a quarter of dimensions the absolute impact on vignette evaluations changed significantly when dimensions were presented in an alternative order, and for some dimensions the relative importance also changed. Estimates of just pay gaps between men and women were ten-fold larger with one order than the other; just returns to education where not affected. The order of dimensions however only mattered when the response task was complex and vignettes either consisted of 12 instead of 8 dimensions, or included a second target question about each vignette.

The results *second* support hypotheses suggesting that respondent characteristics matter. Order effects were stronger for respondents who had weak attitudes or little knowledge on the subject matter – but respondent characteristics only mattered when the vignette task was complex. Contrary to expectations and previous findings respondents' cognitive ability was not associated with order effects. Combinations of certain characteristics however increased the magnitude of order effects, especially for respondents with low ability and weak attitudes. That is, the results suggest that task complexity is a precondition for order effects, and that the effects are stronger if several risk factors coincide.

Third, the magnitude of order effects varied depending on the sequential position of a vignette. Order effects tended to be larger for the first and last five vignettes respondents evaluated. This suggests a link between order effects and respondent learning and fatigue in the course of answering the 20 vignettes. Contrary to expectations we found no evidence of primacy effects, that is, respondents did not appear to attach more importance to dimensions when they were listed first. We also found no clear evidence that the importance of dimensions matters. Overall, dimensions considered most important by all respondents were immune to the order. The personal importance attached to dimensions was however not associated with order effects.

These findings have important implications for the interpretation of results from factorial surveys. If our results are replicated in other studies, researchers should be cautious when interpreting the effects of dimensions that are of minor importance, and when reporting trade-offs between single dimensions. For respondents with strong attitudes (e.g. experts on a topic), there appears to be little risk of order effects. Factorial surveys are however typically used for heterogeneous samples, since they enable an easy implementation of experimental approaches with population samples (Sauer, Auspurg, Hinz, and Liebig 2011; Wallander 2009). In such applications order effects could be particularly problematic. Differences in the evaluations of respondents with weak or strong attitudes might partly represent their different sensitivity to order effects – instead of the differences in attitudes the researcher is interested in. Similarly, comparisons across different (international) surveys or trend studies might be impaired by order effects.

The findings also have important practical implications for the design of factorial surveys. In order to reduce the risk of order effects, it is advisable to minimize the complexity of the evaluation task such that it is manageable for all respondents. Previous research suggests that vignettes consisting of about 8 dimensions are cognitively well manageable by respondents, not only for student samples, but also heterogeneous respondent samples (Sauer, Auspurg, Hinz, and Liebig 2011; Sauer et al. 2009). Similarly, asking only one question about each vignette, rather than two, reduces the risk of order effects. Alternatively, one could routinely randomize the order of vignette dimensions to neutralize any potential order effects. Randomizing the order of dimensions may, however, conflict with a smooth flow of vignette texts, especially when text instead of tabular vignettes are used. We return to this issue.

Our study has some limitations that point to the need for further research. *First*, as in other previous studies (e.g. Holbrook, Krosnick, Moore, and Tourangeau 2007), we were only able to examine conditions under which order effects occur. The experimental design with only two alternative dimension orders did not allow testing more concrete hypotheses about the underlying causal mechanisms.

Second, the relatively small number of respondents meant that we were not able to perform detailed analyses of how order effects, learning and fatigue effects evolve, and interact, as respondents progress through the 20 vignettes. Initial analyses (grouping the vignettes into sets of 5 to increase sample sizes) suggested that there are some interactions: as respondents learn or become fatigued, they concentrate on fewer dimensions. At the same time the extent of order effects decreases. For the final 5 vignettes these effects tend to reverse again, with respondents taking account of a larger number of dimensions, and order effects increasing again. These changes in respondent behavior are consistent with our main results, suggesting that order effects are related to cognitive overburdening, which occurs when respondents try to incorporate more information into their decision making.

Third, the fact that cognitive ability was not related to order effects is surprising given our findings that task complexity matters, and given previous research testing the effects of ability (e.g. Holbrook, Krosnick, Moore, and Tourangeau 2007; Narayan and Krosnick 1996). This could either be the result of our homogenous student sample, or of our measure of ability. The student sample was used deliberately, to avoid potential confounds in a population sample, where differences in cognitive ability are likely to be related both to true differences in the attitudes measured, and to differences in the susceptibility to order effects. This nevertheless suggests the need to replicate a study of order effects using a general population sample, which would vary more in terms of cognitive ability. The measure of ability, using a self-assessment of student performance, was chosen deliberately because university grades are not necessarily comparable between the universities from which the sample was drawn. However, students may not have much information about their performance yet, as they were on average only in their third semester, and asking about their relative performance may be too vague a question to capture true differences in underlying capabilities. This suggests the need for using better measures of cognitive ability in future studies.

Fourth, the vignettes in our study were presented in tabular format (as typically used in choice experiments and conjoint analysis), although factorial surveys typically present vignettes as running text. The tabular format was chosen because varying the order of dimensions is easier than in text format. It remains to be tested whether our results replicate when vignettes are presented in text format. More generally, it remains to be tested whether presenting vignettes in tabular format produces comparable data to text vignettes. To our knowledge this has not been studied, neither for factorial surveys, nor conjoint analysis or choice experiments. Our own initial analysis of an experiment related to the data used in this paper suggests that with vignettes consisting of 8 dimensions, evaluations based on text and tabular formats are comparable. If confirmed and replicated, this finding would have important implications. Instead of the current standard text vignettes, it may be advisable to design factorial surveys using tabular vignettes, since that would allow mitigating any potential order effects by routinely randomizing the order of dimensions. All in all, our results suggest that researchers may need to be more concerned about order effects, not only in standard surveys, but also with experimental vignette-based measurement.

Notes

1. Another motivation for this was to limit the number of variables that enter the regression analyses, by using the continuous prestige scale, instead of categorical indicators of occupations.
2. The main results are robust using an alternative operationalization, based on the number of justice attitude items which the respondent either did not answer, or answered using the middle ("neither agree nor disagree") category, as an indicator of weak attitudes.
3. Alternative operationalizations, such as coding only the two dimensions most important to the respondent as "important", do not provide sufficient between-respondent variation for meaningful analyses. The importance of gross earnings for fair earnings was excluded from this question on dimensions' importance since this makes no sense.
4. There has been some criticism of these measures since they depend on the order in which variables enter the regression model (Soofi, Retzer, and Yasai-Ardekani 2000). This problem however does not apply to orthogonal designs, which we achieved almost perfectly with our D-efficient sample of vignettes.
5. To test whether the pattern of order effects is related to the use of heuristics, consisting in respondents increasingly focusing on the most important dimensions while fading out other ones, we plotted the semi-partial R^2-values (measuring the relative importance) for each dimension against the sequential positions of vignettes. The results (not shown here) provided some evidence for respondents using heuristics. After about 10 vignettes some dimensions (e.g. 'firm success') gained in relative importance, while other dimensions (e.g. 'performance' and 'education) lost impact. For the last part of vignette evaluations there was again a change in answer behavior and the importance of most dimensions returned close to the initial level. This might suggest that respondents lost concentration after about 5 vignettes, but concentrated on the last vignettes again, as has been found in other studies (Sauer et al. 2009). If this interpretation holds, order effects would be more pronounced when respondents try to include more dimensions into their judgments. However, the pattern was not completely clear.
6. Dimension importance was measured after the vignette module and might therefore be influenced by the experimental split. *t*-tests show that the evaluations differ for two (age and firm success) out of eleven dimensions by order condition to a 5% significance level (as already mentioned the importance of earnings for fair incomes was not questioned). Further studies should randomize the order of the importance questions and factorial survey module in the questionnaire.

Appendix

Table A1: Vignette dimensions and levels

#	Dimension	Levels
1	Age	30/40/50/60 years
2	Sex	Male/female
3	Education	No vocational training/vocational training/university degree
4	Occupation	Unskilled worker/doorman/engine driver/clerk/hairdresser/social worker/software engineer/electrical engineer/manager/medical doctor
5	Gross earnings/month	Ten values ranging from 500 to 15000 Euros
6	Experience	Little/a lot of experience
7	Job tenure	Short/long tenure
8	Children	No child/1 child/2 children/3 children/4 children
9	Health status	No health problems/long-term health problems
10	Job performance	Below average/average/above average
11	Firm success	High profits/threatened by bankruptcy/solid
12	Firm size	Small/medium/large enterprise

Notes: The levels ‘no child’ and ‘no health problems’ were oversampled to achieve a more realistic distribution of vignette persons’ characteristics.

Table A2: Randomization check

	Treatment groups									
	dim8/eval1/ order1	dim12/eval1/ order1	dim8/eval1/ order2	dim12/eval1/ order2	dim8/eval2/ order1	dim12/eval2/ order1	dim8/eval2/ order2	dim12/eval2/ order2	*N*	*p*
Male (%)	14.2	16.0	10.4	8.5	11.3	14.2	13.2	12.3		
Female (%)	13.4	9.3	14.5	16.9	10.5	12.2	13.4	9.9	278	0.393
Sociology (%)	14.5	9.8	11.6	15.6	13.3	13.9	11.6	9.8		
Other (%)	12.4	15.2	17.1	10.5	5.7	12.4	16.2	10.5	278	0.228
Partner (%)	11.1	16.7	15.3	12.5	9.7	16.7	12.5	5.6		
No partner (%)	14.4	10.1	12.9	13.9	11.0	12.0	13.4	12.4	281	0.528
Mainz (%)	16.9	7.0	15.5	12.7	12.7	14.1	14.1	7.0		
Konstanz (%)	13.7	12.0	12.0	14.5	12.0	12.8	11.1	12.0		
Bielefeld (%)	10.6	14.9	13.8	12.8	8.5	12.8	14.9	11.7	282	0.950
Semesters (mean)	3.6	3.8	3.8	3.3	2.8	3.3	3.1	3.0	276	0.336
Income (mean)	2777	2311	2592	2413	2885	2446	2832	2738	266	0.021
Birthyear (mean)	1984.4	1984.6	1984.4	1986.7	1985.4	1985.3	1984.9	1985.3	278	0.015

Notes: Row percentages. dim = number of dimensions, eval = number of questions per vignette, order = order of dimensions. *p*-values for categorical values from *Chi*^{2}-tests; for continuous variables from joint tests of the equality of means across treatment groups.

Table A3: Number of vignette evaluations by respondent characteristics and number of dimensions

	8 Dimensions		12 Dimensions	
	Low ability	High ability	Low ability	High ability
Little knowledge				
Weak attitude	295	220	179	280
Strong attitude	375	360	415	611
More knowledge				
Weak attitude	238	239	298	240
Strong attitude	414	636	220	475

Technical Appendix: Estimation of Just Pay Gaps

The estimation of just pay gaps (JPG) is based on the vignette evaluations, which reveal how respondents trade off single dimensions with the earnings dimension when evaluating fairness. Technically, the JPG is based on multivariate regression estimates of the vignette evaluations on vignette dimensions. Here we demonstrate the procedure for the just gender pay gap (JGPG). To calculate this value, first a multivariate OLS regression of the fairness evaluation Y_i for vignette *i* on the sex of the vignette person and the *m* other vignette dimensions is estimated, as formalized in equation (1). The earnings are entered in a logarithmic specification to model their non-linear relationship with the fairness evaluations:

$$Y_i = \beta_0 + \beta_{earnings}\ln(earnings_i) + \beta_{sex}sex_i + \ldots. + \beta_k x_{ki} + \varepsilon_i \qquad i = 1, \ldots, n;\ k = 1, \ldots m \qquad (1)$$

Y_i = fairness evaluation of vignette i

X_k = other vignette dimensions

β_k = regression coefficients

ε_i = random error in judgment

The mean fairness evaluation for a vignette described with the dimensions x_k is estimated as:

$$Y = \beta_0 + \beta_{earnings}\ln(earnings) + \beta_{sex}sex + \ldots. + \beta_k x_k \qquad (2)$$

To determine the JGPG one has intuitively to ask which amount of earnings “neutralizes” the influence of a female instead of a male vignette person (that is, β_{sex}) in regard to the fairness evaluations. This relationship is formalized in equation (3a) for the just absolute earnings difference (JGPG), in equation (3b) for the just percentage difference (%JGPG):

$$\beta_{earnings}\ln(earnings + JGPG) + \beta_{sex} = \beta_{earnings}\ln(earnings) \qquad (3a)$$

$$\beta_{earnings}\ln[earnings \cdot (1 + \%JGPG / 100)] + \beta_{sex} = \beta_{earnings}\ln(earnings) \qquad (3b)$$

After simple transformation one obtains the formulas in equation 4b and 4c which can be easily used to calculate the percentage gap (%JGPG) respectively absolute gap (JGPG):

$$\%JGPG = (\exp(-\frac{\beta\ \mathrm{sex}}{\beta\ \mathrm{earnings}}) - 1) * 100 \qquad (4a)$$

$$JGPG = \mathrm{mean}(earnings) \cdot \frac{\%\mathrm{JGPG}}{100} \qquad (4b)$$

Other JPGs, like just returns to a university degree, are obtained by replacing the coefficient β_{sex} with the coefficient of the respective group dimension (e.g. with $\beta_{university}$). Note that the correct estimation of all JPGs relies on a correct specification of the regression model.

References

Abraham, Martin, Katrin Auspurg, and Thomas Hinz. 2010. "Migration Decisions Within Dual-Earner Partnerships: A Test of Bargaining Theory." *Journal of Marriage and Family* 72:876-892.

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

Alves, Wayne M. and Peter H. Rossi. 1978. "Who Should Get What? Fairness Judgments of the Distribution of Earnings." *American Journal of Sociology* 84:541-564.

Auspurg, Katrin, Thomas Hinz, and Stefan Liebig. 2009. "Complexity, Learning Effects, and Plausibility of Vignettes in Factorial Surveys." Paper presented at *104th Annual Meeting of the American Sociological Association (ASA)*. San Francisco.

Bennett, Jeff and Russell Blamey. 2001. *The Choice Modelling Approach to Environmental Valuation*. Cheltenham, Northmapton: Edward Elgar.

Berk, Richard A. and Peter H. Rossi. 1977. *Prision Reform and State Elites*. Cambridge, Mass.: Ballinger.

Bickart, Barbara A. 1992. "Question-Order Effects and Brand Evaluations: The Moderating Role of Consumer Knowledge." Pp. 63-79 in *Context Effects in Social and Psychological Research*, edited by N. Schwarz and S. Sudman. New York: Springer.

Bijlenga, Denise, Gouke J. Bonsel, and Erwin Birnie. 2011. "Eliciting Willingness to Pay in Obstetrics: Comparing a Direct and an Indirect Valuation Method for Complex Health Outcomes." *Health Economics* 20:1392-1406.

Bishop, George and Andrew Smith. 2001. "Response-Order Effects and the Early Gallup Split-Ballots." *Public Opinion Quarterly* 65:479-505.

Borghans, Lex, Margo Romans, and Jan Sauermann. 2010. "What Makes a Good Conference? Analysing the Preferences of Labour Economists." *Labour Economics* 17:868-874.

Bradburn, Norman M. 1992. "What Have We Learned?" Pp. 315-323 in *Context Effects in Social and Psychological Research*, edited by N. Schwarz and S. Sudman. New York: Springer.

Buskens, Vincent and Jeroen Weesie. 2000. "An Experiment on the Effects of Embeddedness in Trust Situations." *Rationality and Society* 12:227-253.

Carrol, Douglas J. and Paul E. Green. 1995. "Psychometric Methods in Marketing Research. Part I, Conjoint Analysis." *Journal of Marketing Research* 32:358-391.

Christoph, Bernhard. 2005. "Zur Messung des Berufsprestiges. Aktualisierung der Magnitude-Prestigeskala auf die Berufsklassifikation ISCO88." *ZUMA-Nachrichten* 57:79-127.

Chrzan, Keith. 1994. "Three Kinds of Order Effects in Choice-Based Conjoint Analysis." *Marketing Letters* 5:165-172.

Diefenbach, Heike and Karl-Dieter Opp. 2007. "When and Why Do People Think There Should Be a Divorce?" *Rationality and Society* 19:485-517.

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

Farrar, Shelley and Mandy Ryan. 1999. "Response-Ordering Effects: A Methodological Issue in Conjoint Analysis." *Health Economics* 8:75-79.

Garret, Karen. 1982. "Child Abuse: Problems of Definition." Pp. 177-204 in *Measuring Social Judgments. The Factorial Survey Approach*, edited by P. H. Rossi and S. L. Nock. Beverly Hills: Sage.

Glenk, Klaus. 2006. "Economic Valuation of Biological Diversity. Exploring Non-market Perspectives in the Vicinity of the Lore-Lindu National Park in Indonesia's Central

Sulawesi Region." Dissertation Thesis, Fakultät für Agrarwissenschaften, Georg-August-Universität Göttingen, Göttingen.

—. 2007. "Effects of Attribute Order in Choice Experiments: Evidence from Rural Indonesia." *Working Paper of the Macaulay Land Use Research Institute.* Aberdeen: The Macaulay Land Use Research Institute.

Hechter, Michael, James Ranger-Moore, Guillermina Jasso, and Christine Horne. 1999. "Do Values Matter? An Analysis of Advance Directives for Medical Treatment." *European Sociological Review* 15:405-430.

Hembroff, Larry A. 1987. "The Seriousness of Acts and Social Contexts: A Test of Black's Theory of the Behavior of Law." *American Journal of Sociology* 93:322-347.

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

Hippler, Hans-J. and Norbert Schwarz. 1986. "Not Forbidding Isn't Allowing: The Cognitive Basis of the Forbid-Allow Asymmetry." *Public Opinion Quarterly* 50:87-96.

Holbrook, Allyson L., Melanie C. Green, and Jon A. Krosnick. 2003. "Telephone Versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires." *Public Opinion Quarterly* 67:79-125.

Holbrook, Allyson L., Jon A. Krosnick, David Moore, and Roger Tourangeau. 2007. "Response Order Effects in Dichotomous Categorical Questions Presented Orally." *Public Opinion Quarterly* 71:325-348.

Hox, Joop, Ita Kreft, and Piet Hermkens. 1991. "The Analysis of Factorial Surveys." *Sociological Methods & Research* 19:493-510.

Jasso, Guillermina. 1988. "Whom Shall We Welcome? Elite Judgments of the Criteria for the Selection of Immigrants." *American Sociological Review* 53:919-932.

—. 2006a. "Factorial Survey Methods for Studying Beliefs and Judgments." *Sociological Methods & Research* 34:334-423.

—. 2006b. "Homans and the Study of Justice." Pp. 203-227 in *George C. Homans: History, Theory, and Method*, edited by J. Trevino. Boulder, Colorado: Paradigm Press.

Jasso, Guillermina and Karl-Dieter Opp. 1997. "Probing the Character of Norms: A Factorial Survey Analysis of the Norms of Political Action." *American Sociological Review* 62:947-964.

Jasso, Guillermina and Peter H. Rossi. 1977. "Distributive Justice and Earned Income." *American Sociological Review* 42:639-651.

Jasso, Guillermina and Murray Webster Jr. 1997. "Double Standards in Just Earnings for Male and Female Workers." *Social Psychology Quarterly* 60:66-78.

—. 1999. "Assessing the Gender Gap in Just Earnings and Its Underlying Mechanisms." *Social Psychology Quarterly* 62:367-380.

John, Craig St. and Nancy A. Bates. 1990. "Racial Composition and Neighborhood Evaluation." *Social Science Research* 19:47-61.

Johnson, Richard M. 1981. "Problems in Applying Conjoint Analysis." Paper presented at *Conference on Analytic Approaches to Product and Marketing Planning*. Vanderbit University, October 1981.

—. 1989. "Assessing the Validity of Conjoint Analysis." Pp. 273-280 in *Gaining A Competitive Advantage Through PC-Based Interviewing and Analysis.*, vol. 1, edited by M. Metegrano. Sun Valley: Sawtooth Software.

Kjaer, Trine, Mickael Bech, Dorte Gyrd-Hansen, and Kristian Hart-Hansen. 2006. "Ordering Effect and Price Sensitivity in Discrete Choice Experiments: Need We Worry?" *Health Economics* 15:1217-1228.

Knäuper, Bärbel, Norbert Schwarz, Denise Park, and Andreas Fritsch. 2007. "The Perils of Interpreting Age Differences in Attitude Reports: Question Order Effects Decrease with Age." *Journal of Official Statistics* 23:515-528.

Krosnick, Jon A. 1988. "Attitude Importance and Attitude Change." *Journal of Experimental Social Psychology* 24:240-255.

—. 1989. "Attitude Importance and Attitude Accessibility." *Personality and Social Psychology Bulletin* 15:297-308.

—. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213-236.

—. 1992. "The Impact of Cognitive Sophistication and Attitude Importance on Response-Order and Question-Order Effects." Pp. 203-218 in *Context Effects in Social and Psychological Research*, edited by N. Schwarz and S. Sudman. New York: Springer.

Krosnick, Jon A. and Duane F. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement." *Public Opinion Quarterly* 51:201-219.

Kuhfeld, Warren F. 2009. "Marketing Research Methods in SAS. Experimental Design, Choice, Conjoint, and Graphical Techniques." Cary, NC: SAS Institut.

Kuhfeld, Warren F., Randall D. Tobias, and Mark Garrat. 1994. "Efficient Experimental Design with Marketing Research Applications." *Journal of Marketing Research* 31:545-557.

Kumar, V. and Gary J. Gaeth. 1991. "Attribute Order and Product Familiarity Effects in Decision Tasks Using Conjoint Analysis." *International Journal of Research in Marketing* 8:113-124.

Lavine, Howard, Joseph W. Huff, Steven H. Wagner, and Donna Sweeney. 1998. "The Moderating Influence of Attitude Strength on the Susceptibility to Context Effects in Attitude Surveys." *Journal of Personality and Social Psychology* 75:359-373.

Liebig, Stefan and Steffen Mau. 2002. "Einstellungen zur sozialen Mindestsicherung. Ein Vorschlag zur differenzierten Erfassung normativer Urteile." Kölner Zeitschrift für Soziologie und Sozialpsychologie.

—. 2005. "Wann ist ein Steuersystem gerecht? Einstellungen zu allgemeinen Prinzipien der Besteuerung und zur Gerechtigkeit der eigenen Steuerlast." *Zeitschrift für Soziologie* 34:468-491.

Louviere, Jordan, David A. Hensher, and Joffre D. Swait. 2000. *Stated Choice Methods. Analysis and Application*. Cambridge: Cambridge University Press.

Ludwick, Ruth, Marion E. Wright, Richard A. Zeller, Dawn W. Dowding, William Lauder, and Janice Winchell. 2004. "An Improved Methodology for Advancing Nursing Research: Factorial Surveys." *Advances in Nursing Science Advances in Research Methods* 27:224-238.

Malhotra, Neil. 2009. "Order Effects in Complex and Simple Tasks." *Public Opinion Quarterly* 73:180-198.

McClendon, McKee J. 1991. "Acquiescence and Recency Response-Order Effects in Interview Surveys." *Sociological Methods & Research* 20:60-103.

Meudell, Bonner M. 1982. "Household and Social Standing: Dynamic and Static Dimensions." Pp. 69-94 in *Measuring Social Judgments. The Factorial Survey Approach* edited by P. H. Rossi and S. L. Nock. Beverly Hills: Sage.

Miller, J. L., Peter H. Rossi, and Jon E. Simpson. 1986. "Perceptions of Justice: Race and Gender Differences in Judgments of Appropriate Prison Sentences." *Law & Society Review* 20:313-334.

Müller-Benedict, Volker and Elena Tsarouha. 2011. "Können Examensnoten verglichen werden? Eine Analyse von Einflüssen des sozialen Kontextes auf Hochschulprüfungen." *Zeitschrift für Soziologie* 40:388-409.

Narayan, Sowmya and Jon A. Krosnick. 1996. "Education Moderates Some Response Effects in Attitude Measurement." *Public Opinion Quarterly* 60:58-88.

Nock, Steven L. 1982. "Family Social Standing: Consensus on Characteristics." Pp. 95-118 in *Measuring Social Judgments. The Factorial Survey Apporach*, edited by P. H. Rossi and S. L. Nock. Beverly Hills: Sage.

O'Toole, Richard, Stephen W. Webster, Anita W. O'Toole, and Betsy Lucal. 1999. "Teachers' Recognition and Reporting of Child Abuse: a Factorial Survey." *Child Abuse & Neglect* 23:1083-1101.

Olsen, Søren Bøye, Jacob Ladenburg, Mads L. Petersen, Ulrich Lopdrup, Anja S. Hansen, and Alex Dubgaard. 2005. "Motorways versus Nature. A Welfare Economic Valuation of Impacts." Report from the Environmental Assessment Institute Copenhagen.

Orme, Bryan K. 2006. *Getting Started with Conjoint Analysis. Strategies for Product Design and Pricing Research*. Madison/Wisconsin: Research Publishers LLC.

Orme, Bryan K., Mark I. Alpert, and Ethan Christensen. 1997. "Assessing the Validity of Conjoint Analysis – Continued." Pp. 209-225 in *Proceedings of the Sawtooth Software Conference,* edited by S. Software. Seattle: Sawtooth Software.

Payne, Stanley L. 1949. "Case Study in Question Complexity." *Public Opinion Quarterly* 13:653-658.

Perrey, Jesko. 1996. "Erhebungsdesign-Effekte bei der Conjoint-Analyse." *Marketing. Zeitschrift für Forschung und Praxis* 2:105-116.

Rossi, Peter H. 1979. "Vignette Analysis: Uncovering the Normative Structure of Complex Judgments." Pp. 176-186 in *Qualitative and Quantiative Social Research. Papers in Honor of Paul F. Lazarsfeld,* edited by R. K. Merton, J. S. Coleman, and P. H. Rossi. New York: The Free Press.

Rossi, Peter H. and Andy B. Anderson. 1982. "The Factorial Survey Approach: An Introduction." Pp. 15-67 in *Measuring Social Judgements: The Factorial Survey Approach*, edited by P. H. Rossi and S. L. Nock. Beverly Hills: Sage.

Rossi, Peter H., William A. Sampson, Christine E. Bose, Guillermina Jasso, and Jeff Passel. 1974. "Measuring Household Social Standing." *Social Science Research* 3:169-190.

Ryan, Mandy, Karen Gerard, and Mabel Amaya-Amaya. 2008. *Discrete Choice Experiments to Value Health and Health Care*. Dodrecht: Springer.

Sauer, Carsten, Katrin Auspurg, Thomas Hinz, and Stefan Liebig. 2011. "The Application of Factorial Surveys in General Population Samples: The Effects of Respondent Age and Education on Response Times and Response Consistency." *Survey Research Methods* 5:89-102.

Sauer, Carsten, Stefan Liebig, Katrin Auspurg, Thomas Hinz, Andy Donaubauer, and Jürgen Schupp. 2009. "A Factorial Survey on the Justice of Earnings within the SOEP-Pretest 2008." *IZA Working Paper*. Bonn: Institute for the Study of Labor (IZA).

Schuman, Howard. 1992. "Context Effects: State of the Past/State of the Art." Pp. 5-20 in *Context Effects in Social and Psychological Research*, edited by N. Schwarz and S. Sudman. New York: Springer.

Schuman, Howard and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording, and Context.* New York: Academic Press.

Schwarz, Norbert. 2007. "Cognitive Aspects of Survey Methodology." *Applied Cognitive Psychology* 21:277-287.

Schwarz, Norbert, Hans-J. Hippler, and Elisabeth Noelle-Neumann. 1992. "A Cognitive Model of Response-Order Effects in Survey Measurement." Pp. 187-201 in *Context*

Effects in Social and Psychological Research, edited by N. Schwarz and S. Sudman. New York: Springer.

Schwarz, Norbert and Bärbel Knäuper. 2000. "Cognition, Aging and Self-Reports." Pp. 233-252 in *Cognitive Aging: A Primer*, edited by D. C. Park and N. Schwarz. Philadelphia: Psychology.

Scott, Anthony and Sandra Vick. 1999. "Patients, Doctors and Contracts: An Application of Principal-Agent Theory to the Doctor-Patient Relationship." *Scottish Journal of Political Economy* 46:111-134.

Shepelak, Norma J. and Duane F. Alwin. 1986. "Beliefs about Inequality and Perceptions of Distributive Justice." *American Sociological Review* 51:30-46.

Simon, Herbert. 1957. *Models of Man*. New York: Wiley.

Smith, Tom W. 1992. "Thoughts on the Nature of Context Effects " Pp. 163-184 in *Context Effects in Social and Psychological Research*, edited by N. Schwarz and S. Sudman. New York: Springer.

Soofi, Ehsan S., Joseph J. Retzer, and Masoud Yasai-Ardekani. 2000. "A Framework for Measuring the Importance of Variables with Applications to Management Research and Decision Models." *Decision Sciences* 31:595-625.

Stark, Gunnar, Stefan Liebig, and Bernd Wegener. 2008. "Gerechtigkeitsideoloigen." in *Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS Version 12.00,* edited by A. Glöckner-Rist. Bonn: GESIS.

Steiner, Peter and Christiane Atzmüller. 2006. "Experimentelle Vignettendesigns in Faktoriellen Surveys." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 58:117-146.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. New York: Jossey-Bass.

Tourangeau, Roger. 1999. "Context Effects on Answers to Attitude Questions." Pp. 111-132 in *Cognition and Survey Research*, edited by M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, and R. Tourangeau. New York: John Wiley & Sons.

Tourangeau, Roger and Kenneth A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103:299-314.

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

Tourangeau, Roger, Eleanor Singer, and Stanley Presser. 2003. "Context Effects in Attitude Surveys: Effects on Remote Items and Impact on Predictive Validity." *Sociological Methods & Research* 31:486-513.

Wallander, Lisa. 2009. "25 Years of Factorial Surveys in Sociology: A Review." *Social Science Research* 38:505-520.

Wooldridge, Jeffey M. 2003. *Introductory Econometrics. A Modern Approach*. Mason, Ohio: South Western.