Universität Bielefeld

Universität Konstanz

Working Paper # 4

Complexity, Learning Effects, and Plausibility of Vignettes in Factorial Surveys

Paper accepted for the ASA-Conference 2009

Katrin Auspurg, Thomas Hinz and Stefan Liebig

January 2009

Research Project: The Factorial Survey as a Method for Measuring Attitudes in Population Surveys

The research project *'The Factorial Survey as a Method for Measuring Attitudes in Population Surveys'* is funded by Deutsche Forschungsgemeinschaft (DFG).

Prof. Dr. Thomas Hinz
Dipl. Soz. Katrin Auspurg
Professur für empirische
Sozialforschung mit Schwerpunkt
Demoskopie
Universität Konstanz
Geisteswissenschaftliche Sektion
Postfach D-40
78457 Konstanz
Tel.: 07531 88-2349/-2167
Fax: 07531 88-4085
katrin.auspurg@uni-kostanz.de

Prof. Dr. Stefan Liebig
Carsten Sauer, M.A.
Professur Soziale Ungleichheit
und Sozialstrukturanalyse
Universität Bielefeld
Fakultät für Soziologie
Postfach 10 01 31
33501 Bielefeld
Tel.: 0521 106-4616/-6948
Fax: 0521 106-6479
carsten.sauer@uni-bielefeld.de

Homepages:

http://www.uni-konstanz.de/hinz/?cont=faktorieller_survey&lang=de

http://www.uni-bielefeld.de/soz/arbeitsbereiche/sozialstrukturanalyse/faktsurvey.html

Complexity, Learning Effects, and Plausibility of Vignettes in Factorial Surveys[1]

Katrin Auspurg, Thomas Hinz (University of Konstanz)
Stefan Liebig (Bielefeld University)

Abstract

The factorial survey is a research method of that combines the advantages of survey research with the advantages of experimental designs. Respondents react to hypothetical descriptions of objects or situations (vignettes) instead of answering single-item questions. By varying each dimension of the vignettes in an experimental design, the dimensions' impact on respondents' judgments or decisions can be estimated accurately. Thus, the method is able to isolate the weight of single factors which are often confounded in reality. So far, only a few methodological studies address questions of validity of measurement within factorial surveys. The paper provides an overview of the use of factorial surveys in social sciences and open methodological questions. Three of these methodological problems are discussed in more detail and studied empirically using data from an online-experiment: (1) the effects of different complex situations presented to respondents, (2) learning effects due to a repeated presentation of the vignettes and (3) the effects of implausible vignettes on respondents' reactions. According to our results all three aspects matter: A high complexity of the vignettes leads to weaker effects of single dimensions while the consistency of the judgments remains the same. Until the tenth vignette there is evidence for a learning effect: Response speed increases at a stable rate of consistent responses. Implausible combinations of vignette dimensions cause the respondents to neglect the respective dimensions and therefore lead to artificial judgments. Finally, we discuss the practical consequences of these findings.

[1] The research project "Factorial Survey Design as an Instrument for the Measurement of Attitudes" is funded by the German Science Foundation (DFG). Please all correspondence to: katrin.auspurg@uni-konstanz.de.

1. Introduction

The factorial survey is an experimental method used in survey studies which confronts respondents with hypothetical descriptions of objects or situations (vignettes).[2] The vignettes vary by certain characteristics (dimensions) which have different levels (variable values). Such hypothetical descriptions of cases and scenarios evaluated by respondents are more and more used in academic and non-academic research, including the social sciences, law studies, psychology as well as marketing research. Within sociology, the topics addressed by factorial surveys are very broad: norms and values, measurement of status and prestige of individuals and households (Rossi 1979; Meudell 1982; Nock 1982), evaluations of a fair labor market income (Alves/Rossi 1978; Hermkens/Boerman 1989; Jann 2003; Jasso 1994; Jasso/Webster 1997, 1999; Shepelak/Alwin 1986), the dimensions of poverty (Will 1993), criteria for welfare payments and fair tax rates (Liebig/Mau 2002, 2005) as well as acceptable criteria for layoffs in firms (Struck et al. 2008). There are further studies on different levels of sexual harassment (Garrett 1982; Rossi/Anderson 1982; O´Toole et al. 1999), on appropriate sentences for criminals (Berk/Rossi 1977; Hembroff 1987; Miller/Rossi/Simpson 1986), the criteria for the selection of immigrants (Jasso 1988), for medical treatment (Hechter et al. 1999) and for the activation of social norms (Beck/Opp 2001; Diefenbach/Opp 2007; Horne 2003; Jasso/Opp 1997). Additionally, research using the factorial survey focuses on discrimination (Jann 2002; John/Bates 1990), social embeddedness (Buskens/Weesie 2000) and the sociology of the family (Auspurg/Abraham 2007). Despite of the increasing use of this methodological instrument there are only a few studies discussing methodological issues (e.g. Dülmer 2001, 2007; Dülmer/Klein 2003; Steiner/Atzmüller 2006), and the few existing studies primarily intend to underline the advantages of the factorial survey design compared to item-based questionnaires and traditional experimental designs (Hechter et al. 2005; Jasso 1988). Methodological problems connected to this technique have been rarely studied in an explicit and systematic way. This is foremost true for the problems of planning and conducting factorial designs and using vignettes in survey research. A deeper methodological understanding is essential for appropriate constructions of vignettes, methods of data analysis, and interpretation of results.

We aim at three closely connected and important methodological issues and their empirical relevance: (1) the effects of different complex situations presented to respondents, (2) learning effects due to a repeated presentation of the vignettes and (3) the effects of implausible vignettes on respondents' reactions. The probability of such implausible cases rises with the extent of complexity. And higher complexity is related to the use of simplified heuristics in the evaluation of the presented situations.

[2] The factorial survey was established in the social sciences by Peter H. Rossi in his dissertation in 1951. It was used for the measurement of social status and prestige of households (Alves/Rossi 1978; Rossi 1979; Rossi/Nock 1982). Rossi's central goal was the development of a method of measurement that distinguishes between the relative relevance of several factors for social attitudes (Rossi/Anderson 1982:15 et seq..; Rossi/Nock 1982).

The paper proceeds in six steps: we start with a brief description of the factorial survey design (section 1) and an overlook on previous methodological research (section 2). Based on the current state of research, hypotheses on the mentioned problems are derived (section 3). We test the hypotheses with data from an experimental online-study (sections 4 und 5). Finally, empirical results are discussed and directions for further research are outlined (section 6).

2. Factorial Survey: Design, Motivation and Problems

By using a factorial survey design it is possible to determine the relative weights of single characteristics which describe an object or a situation (for detailed discussion: Jasso 2006; Rossi/Anderson 1982). Each factorial survey starts with the – theoretically driven – selection of variables describing the object or situation (dimensions and levels). The second step is the experimental variation of the dimensions and levels. By this variation the causal effects of each dimension on the respondents' reactions can be tested. Thus the factorial survey design reveals the systematic correlation of the situational factors (in the descriptions) with the evaluations (and decisions) of the respondents. Factorial surveys normally present several, randomly or systemically selected vignettes to each respondent.[3]

The factorial survey has a number of advantages compared to regular, item-based scales. (1) Objects and situations are constructed which are characterized by a *bundle* of variables. In reality, these variables are often confounded and as a consequence the estimation of separate factors is difficult. The experimental set-up of the factorial survey facilitates – in a technical sense – the isolation of each variable, they are set orthogonally. The statistical independence of the variables is created by design and enables thus a separate measurement of the influence of each variable. (2) In contrast to classical experimental studies in laboratories hypotheses can be tested on the basis of larger (random) samples of the population. (3) In case of several vignettes per respondent the hierarchical data structure can be used to identify between and within factors. It is possible to detect the covariation of variables in the vignettes with variables describing the respondents. (4) The factorial survey avoids a severe problem of conventional measurement of attitudes in survey research. The analysis of single items does not fit the complex structure of attitudes. The factorial survey simulates the complexity of real world decisions and evaluation problems by crossing a variety of dimensions (see Jasso/Opp 1997: 949; Liebig/Mau 2002: 114-116). This holds in particular for objects and situations with several relevant dimensions and when varying social contexts play an important role. For instance, the amount of an income perceived as fair or a just sentence for convicts is linked to varying conditions of the

[3] If *decisions* of respondents are at the focus, only a few vignettes (if not a single one) are presented. This is as well the standard approach of conjoint analysis which is mainly used in marketing research (Carroll/Green 1995). Respondents are confronted with real or simulated descriptions of products and asked to rank them. Thereby, the utility values of the dimensions are calculated. The products as well the vignettes have a multi factorial design (Klein 2002; Orme 2006). Given a random allocation of vignettes to respondents we see plausible arguments to present only one vignette per respondent: effects of social desirability as well as learning effects are no longer relevant. Such studies do not require a hierarchical data analysis either (Jann 2003).

social situations and the characteristics of the evaluating respondents. The factorial survey design takes exactly these conditions into account and thereby simulates more realistic situations. According to some authors (see Hechter/Kim/Baer 2005; Jasso 1988; Dülmer/Klein 2003) a measurement using several factors which are relevant for the attitudes or decisions of respondent has a higher validity compared to item-based scales. In factorial survey designs attitudes towards single dimensions are not asked sequentially but jointly. In addition, it is assumed that the repeated evaluation of a larger number of objects and situations prevents respondents from depicting a biased and very artificial attitude (Hechter et al. 1999). Comparisons of item-based and vignette-based measurements in fact reveal that attitudes measured by factorial surveys are less biased against social desirability (Jann 2003; Liebig/Mau 2002; Smith 1986). Given these results Dülmer and Klein (2003) resume that vignette analysis yields a more exact measurement of attitudes (see also Hechter/Kim/Baer 2005: 103; Jasso 1988).

Researchers who are more critical about the factorial survey name a number of disadvantages and shortcomings. Basic claims refer to the high expenses of survey time and the opportunity costs regarding alternative items (Sniderman/Grob 1996). The evaluation of ten and more vignettes is more time consuming than an analogously item-based module targeting at the single dimensions (Dülmer/Klein 2003; Liebig/Mau 2002). It is assumed as even more problematic that vignettes studies are accompanied by strong response effects resulting from the selection of presented cases (i.e. contrast effects), the sequence of cases (*carry-over*-effects) and the complexity of cases. Thus, the measured attitudes are probably instable and even artificial. The latter charge is partly due to a possible cognitive over-load for the respondents. They might align their responses to the variation of dimensions that are actually irrelevant. Further critical problems arise with a strong cognitive force towards consistent responses (Faia 1980; Seyde 2005). Moreover, using names and notions might evoke a systematical bias because the used terms are related to individual and idiosyncratic experience of respondents which is hard to control.[4] These objections have not been confirmed or refuted by methodological studies yet.

This paper addresses such gaps in previous methodological research and is based on data collected in the first phase of a much broader research project.[5] Presented analyses concern three topics: (1) It seems necessary to discuss and establish rules for a maximum of complexity to employ in the described situations and objects (Beck/Opp 2001: 287; Rossi/Anderson 1982: 59). We address this complexity by the *number of dimensions*. Possible cognitive over- or under-loads are closely related to

[4] Prognostic validity is at question (Rooks et al. 2000) because respondents only make hypothetical but no real world decisions (see Hechter et al. 2005; for studies on external validity: Eifler 2007; Groß/Börensen 2008; Nisic/Auspurg 2008).

[5] This project studies a large number of experimental variations on the complexity of the situation (number of dimensions, number of vignettes, effects of sequence), on the form of presentation (width of dimensions [„range effects"], effects of different forms of scaling and modes of presentation). Moreover, the project focuses at the stability of evaluations over time. Several series of experiments are conducted with a homogenous sample of students and with a standard population survey.

learning effects created through the repeated evaluation of vignettes. (2) Therefore, we focus at the *consistency of responses* dependent on the sequence of presented vignettes. (3) We study the impact of *implausible cases* which are again closely related to both other aspects (for instance, the probability of implausible cases rises with the complexity and at the same time they might cause the activation of simplified heuristics representing a special case of learning effects).

3. Previous Research and Hypotheses

Before we introduce our methodological approach and data set results from previous research concerning the three problems at focus are briefly discussed. This also leads to straightforward hypotheses on the response effects of interest. We consider the literature on similar methods in marketing research as well as environmental and health economics (conjoint analysis and choice experiments).

3.1 Complexity of Vignettes: Number of Dimensions

As argued above the factorial survey is best suited for research questions that need complex evaluations by the respondents. The intention to receive a detailed and most realistic description through a bundle of dimensions collides with a restriction in the cognitive capacity of the respondents. The decision for a certain number of dimensions is of far reaching importance (Rossi/Anderson 1982) because the number of dimensions determines the length of the descriptions and the complexity of the evaluation task. A high number of dimensions eventually creates a complexity which is very hard to handle for the respondents. Consequently, the evaluation might become more and more artificial – if the respondents do not even terminate the experiment. Jasso (2006) supposes to select only such dimensions that are known to be relevant for the judgment. This can be based on theoretical reasoning, previous studies and every day considerations. With reference to studies in cognitive psychology she argues to use only a few dimensions. Rossi and Anderson (1982) suggest a self restriction to six dimensions. Previously conducted factorial surveys show a variation from three (Berk/Rossi 1977) up to ten dimensions (Smith 1986). The majority of studies use five to seven dimensions. This empirical fact is more or less based on a rule of thumb from cognitive sciences saying that humans can deal best with seven plus/minus two pieces of information (Zimbardo 1988: 275). We see a broad variation in fixing the number of dimensions and levels. Suggestions delivered in the literature do not go beyond very general advice. Beck and Opp (2001: 287) suggest in accordance with Jasso (2006) to generate the dimensions out of theoretical hypotheses and to use only such dimensions that are actually expected to have an impact.[6]

[6] Apart from the number of dimensions the number of levels (values) is also of great importance because both parameters determine the size of the universe of vignettes, i.e. the total number of all possible variations in descriptions of situations and objects.

A plausible but very general preposition states that the cognitive demand on the side of the respondents increases with the number of dimensions – up to a situation where the situation might become too complex (Rossi/Anderson 1982; for choice experiments and conjoint analyses Melles 2001; DeShazo/Fermo 2002). It is less clear how the expected tendency for simplification is expressed. Besides of a complete dropout and item non-response there might be a trend towards inconsistent responses. But also the emergence of heuristics, e.g. in form of a complete neglect of dimensions which are of minor importance or attract less attention (e.g. because they vary with a lower number of levels) can be expected (Wason/Polonsy/Hyman 2002; for results in choice experiments and conjoint analyses Swait/Adamowicz 2001; Melles 2001; DeShazo/Fermo 2002). In choice experiments and conjoint analyses, even the oppositional proposition of a more consistent response behavior is supported. This is based on the presumption that vignettes with fewer dimensions miss relevant information. Therefore, respondents construct the missing piece of information by themselves.[7] Compared to the explicit information given by the researcher, the under-complex situation leads to a smaller degree of experimental control and, thereby, to a higher variance across respondents and less precise estimates (e.g. DeShazo/Fermo 2002; Caussade et al. 2005: 632; Johnson 2006: 46f.). Similarly, uncontrolled framing effects become more probable (see e.g. Melles 2001: 186). Finally, a lack of information is seen as a cognitive burden as well. This is plausible because a smaller number of dimensions make it more difficult to recognize differences between the vignettes (for this argument in choice experiments: Hensher 2006). Results from a panel study can be seen as preliminary support for this proposition of an information under-load. Students who have been confronted with exactly the same set of vignettes at three different times of measurement showed a higher stability of judgments when confronted with eight dimensions instead of five dimensions (Liebig/ Meyermann/Schulze 2006).

Anyway it is unclear which number of dimensions leads to the effects of information under-load or over-load. For this paper a strong contrast between five and twelve dimensions is chosen. The number of twelve dimensions is highly above average used in previous studies. This high number lets us expect a domination of the over-load effect. There are the following sub-hypotheses: *Twelve dimensions are more often accompanied with dropouts compared to vignettes with five dimensions (H_{1a}); the evaluations per respondent are more inconsistent in case of twelve than five dimensions (H_{1b}).* In contrast there might be a fall back to a simplified evaluation strategy by masking some (minor relevant) dimensions (Swait/Adamwicz 2001: 137): *In case of twelve dimensions single dimensions of vignettes become less relevant for judgments. That is, they have lower regression coefficients compared to a set-up with five dimensions. (H_{1c}).*

[7] In vignettes studies on the fairness of income such a lack of information could consist of missing data on the occupational experience.

3.2 Learning and Exhaustion Effects

The repeated presentation of vignettes per respondent leads to a sufficient number of cases for testing hypotheses even if the total number of respondents is rather low (Auspurg/Abraham/Hinz 2008). A second motivation for the repeated presentation is the detection of respondent specific rules of evaluation and decision („within subjects“). The repeated making of evaluations per respondent goes hand in hand with respondents' learning effects which are interdependent with other situational factors. This is obvious for the number of dimensions discussed in the previous section. Learning needs a longer sequence of vignettes if the number of dimensions is high. At the same time exhaustion could start earlier. Learning and exhaustion are two sides of the coin of complexity. Learning is connected to a more consistent response behavior and to the capacity to integrate a higher number of dimensions into the judgments. A growing relevance of single dimensions within the sequence of the vignettes is also to expect due to the fact that during the first presented vignettes respondents may view variables as redundant which are highly correlated in the real world. Not before respondents recognize after a series of vignettes that the variables co-vary independently they take these variables into account (for conjoint analysis: Melles 2001: 118). This does not hold however without limitations, because vignettes presented later in the sequence cause exhaustion and monotonicity, which in turn result in a reduced consistency and a stronger tendency to simplified evaluation heuristics (for choice experiments: Carson et al. 1994: 335f.).[8] So far, the role and the degree of learning and exhaustion effects have not been studied systematically for vignettes studies. It is also an open question from which number of vignettes a trade off between the quantity and the quality of judgments can be expected.

As a point of reference we use the experience made in choice experiments which shows that the consistency grows until about the tenth judgment and shrinks afterwards (e.g. Bradley/Daly 1994: 180; Caussade et al. 2005: 631f.). Because previous vignettes studies did not reveal major problems with the quality of judgment even using 50 vignettes and more (Jasso 2006), the given number of vignettes in this paper (10) let us expect a domination of learning effects. We test the following hypotheses: *The higher the position of vignettes within a set of vignettes the higher the consistency of responses and/or the number of dimensions which are taken into account (H_{2a}); this effect is stronger for twelve than for five dimensions (H_{2b}).*

[8] Graphically we expect an inverted u-shaped curve depicting the correlation of the number of vignettes with the consistency respectively the number of relevant dimensions.

3.3 Treatment of Implausible and Illogical Cases

Before a sample of vignettes which should be evaluated is drawn from the universe of all vignettes (from all possible combinations of variables, see Beck/Opp 2001; Steiner/Atzmüller 2006; Dülmer 2007) it is conventional to exclude vignettes with obviously unusual or absurd combinations („implausible" or "illogical cases").An "illogical case" would be a vignette person with a longer job tenure but without any occupational experience. An example for "implausible case" is given by a vignette person without school education working in an occupation which normally requires academic training (e.g. a judge or university professor). The elimination of such irritating vignettes is legitimated by the aim of avoiding unintended consequences for the response behavior. Absurd cases would obviously question the seriousness of the evaluation task and, therefore, raise the probability of non-responses and dropouts (Faia 1980; Jasso 2006).

At first glance, this argument is perfectly reasonable but the criteria for "implausible cases" seem pretty vague. Many factorial surveys aim at evaluations that are made most independently from accepted norms and laws in order to measure contra-factual attitudes and opinions. The normality of a "plausible case" is shaped by empirical regularities and related expectations. Factorial surveys provide the rare opportunity to confront the respondents with *deviant* cases – just the reaction to this deviance from normality is often an explicit research goal. In this respect, manipulations of variables combinations are problematic because they restrict the variation of descriptions of objects and situations a priori to an actually existing degree (Beck/Opp 2001).

Solid methodological knowledge exists only with regard to the statistical consequences of the exclusion of implausible and illogical cases. By the exclusion of such cases the orthogonal design of dimensions is reduced, thus multi-collinearity is forced by design (consequences for estimates: Greene 2003: 56-59; Wooldridge 2003: 96-100). It is meanwhile well documented how relevant the exclusion of cases is in regard to the balance and independence of samples of vignettes (Kuhfeld/Randall/Garratt 1994: 551; Dülmer 2007: 391f.). There are algorithms minimizing the reduction of efficiency if the exclusion of some cases is inevitable. When using a fractionalized design and, in particular, when using a small sample of vignettes it is absolutely necessary to apply these algorithms to avoid a significant loss of efficiency.

The impact of implausible cases on the responses is much more disputed – primarily due to the lack of research targeted at this problem.[9] If the assumption is true that implausible cases reduce the trust in the principal relevance of the survey and also reduce believe in the utility of participation, dropouts and – even more problematic – invalid responses are likely (response sets or superficial and inconsistent evaluations). Therefore, we derive the following hypotheses: *In case of implausible cases dropouts are more frequent (H_{4a}) and the consistency of responses is lower (H_{4b}) compared to situations without such cases.* Faia (1980) additionally expects that the dimensions which cause the

[9] The discussion between Rossi/Alves (1980) and Faia (1980) on the sense and the gain of implausible vignettes cannot be decided on the basis of empirical data yet.

implausibility of cases become more salient. The respondents would misinterpret the task of evaluation as a test of intelligence to de-mask anomalies. This produces an artificial result. We try to test the following hypothesis: *After a confrontation with implausible cases respondents tend to consider primarily the dimensions causing the implausibility, whereas the other dimensions become less relevant (H_{4c}).* An alternative explanation for this hypothesis is given by the learning processes mentioned above: An empirically rare case makes the respondents notice that variables co-vary independently. A shrinking willingness of cooperation might have a similar effect: The switch towards simplified response behavior with lower cognitive burden is accompanied by a loss of relevance of certain dimensions.

4. Methodology and Data

The three methodological problems cannot be solved by an analytical method but need an empirical approach. By conducting a methodological experiment, it is possible to study the relevance of design elements for response behavior and data quality. Most importantly, the methodological splits have to be assigned to respondents by random. Moreover, the splits must not correlate with the decks of vignettes. As in a true experimental setup, this randomization is a prerequisite in order to neutralize unobserved factors of the respondents (e.g. their cognitive attention) and to avoid a confounding of respondents' characteristics with the dimensions of the vignettes.[10] Therefore, the experimental setup needs to be thoroughly developed. In the experiments studied in this paper, the complexity of vignettes is mainly implemented by the number of dimensions. About one half of the respondents is asked to evaluate vignettes with five dimensions, whereas the other half is confronted with vignettes consisting of twelve dimensions (i.e. for testing our hypothesis we apply a between subject design).[11] In a first round, participants answered seven vignettes. Because of a very low dropout rate we increased the number of vignettes from seven to ten in the second (smaller) subsample.

The subject of our study is the well-known, nearly classical topic of a lot of vignettes studies: what is a fair amount of income given the characteristics described in the vignettes (z.B. Alves/Rossi 1978; Jasso/Webster 1997, 1999; Jann 2003; Hermkens/Boerman 1989, Shepelak/Alwin 1986). Respondents receive fictional examples representing persons with a number of variables that are relevant for the income position (e.g. sex, age, education, occupation). Additionally, each vignette gives a value for

[10] In choice experiments similar studies are labeled "design of design" (e.g. Hensher 2004, 2006; Caussade et al. 2005). Principally, it is a multi factorial extension of "split ballot designs". Several design elements are varied independently (Sniderman/Grob 1996).

[11] This is not quite correct. As a second experimental factor we varied the sex of the vignette person only for a subsample of the respondents (within variation). Other respondents only received vignettes where the sex of the vignettes persons was hold constant per respondent (between variation). They generally had to make evaluations for male or female (fictional) persons only. Thereby the number of variable dimensions in this split was reduced to four resp. eleven variables. We implemented this split in order to study effects of social desirability. Because this factor was varied independently it can be neglected for the following analyses. It should be studied in a separate and more detailed way.

the monthly net income (in Euros) for the person described. The respondents are requested to rate on an 11-point-scale whether this amount of income is fair/unfair and whether it is too low or too high. Figure 1 gives an example of a vignette with twelve dimensions. The levels of dimensions are underlined and listed below. The selection of variables was based on theoretical considerations and on empirical knowledge on the relevance of these dimensions. This should ensure that a possible neglect of certain dimensions is not caused by their actual irrelevance but due to a cognitive over-load (see for choice experiments: Hensher 2006: 16).

- Figure 1 here -

In order to minimize the costs and the organizational efforts to implement the random assignment of experimental splits and vignettes to respondents we decided to conduct an online-survey. A further argument supporting this survey mode is the easy documentation of meta-data (e.g. response times) which are useful to analyze respondents' strategies of evaluating the vignettes. Experiments do not necessarily require a representative and randomly chosen sample. In case of smaller number of participants a more homogenous sample is advantageous because the risk of unobserved heterogeneity is dampened (e.g. Diekmann 2007: 337ff.). We chose students of several German universities as respondents because of their relatively high homogeneity and because of the easy channels of communication to them. We used email distribution lists of students' organizations and communities and asked the students to take part in our study.

The used vignettes represent a fractionalized selection out of the universe of vignettes for twelve dimensions. We employed orthogonal main effects („resolution III-Design", Kuhfeld/Randall/Garratt 1994: 546). Given this specification and the variety of dimensions and levels, about 100 vignettes are sufficient for an efficient sample of vignettes.[12] The exclusion of some (really) illogical cases (e.g. persons without occupational experience but long job tenure) reduces the number of different vignettes to 93. Empirically rare cases which might be possible, however, have been kept in the sample due to our research goals. The experimental split with five dimensions uses exactly the same sample of vignettes (practically the seven additional dimensions have been deleted). It would be possible to draw more efficient samples for the vignettes with a smaller number of dimensions but the statistical efficiency should be set constant in order to isolate the methodological effects. For the precision of estimates the statistical efficiency of vignettes sample is after all similarly important as the "cognitive" efficiency of the respondents' evaluations (see for similar arguments regarding choice experiments and conjoint analyses: Melles 2001: 109; Louviere 2001b). Only under consideration of statistical efficiency, differences in the significance of regressions coefficients can be traced back to actual differences in the reactions of respondents and explanations due to differences in statistical power can be ruled out. Finally, the use of identical vignettes samples for both splits of five and twelve

[12] A value of 98.2 is reached for the D-efficiency. Values over 90 are satisficing (Kuhfeld 2005). The efficiency is however reduced by the exclusion of implausible cases.

dimensions makes it possible to strictly separate design effects from the effects of different vignettes. To put it differently: Using our approach we can clearly trace back possible differences in the response behavior to the influence of different design elements (instead of different combinations of vignettes dimensions).

All participants have been assigned to the two major splits (5 vs. 12 dimensions) and the subsets of vignettes by random. Moreover, the sequence of vignettes varied randomly across respondents in order to exclude effects of contrast or sequence. Thereby "extreme" vignettes have been assigned randomly to certain positions in the sequence of vignettes. Detected effects of sequence are thus more directly linked to learning or exhaustion. A random selection per respondent has the additional advantage of a higher methodological variation (e.g. occurrence and frequency of implausible cases) between the respondents.

The online-survey was conducted from December 2007 until March 2008. Vignettes have been integrated into a questionnaire with additional items about political and social attitudes and socio-demographic variables. 558 students clicked the link to the online-survey, 460 students took part in the survey and delivered 3,480 evaluations of vignettes. [13] Table 1 shows the number of cases realized for the different experiments.

- Table 1 about here -

Data analysis has to consider the multilevel structure of the data. Presenting several vignettes to the respondents creates – as mentioned above – a hierarchical data set (for an illustration: Beck/Opp 2001). At the lowest level, we find the evaluations of the vignettes. A second level is given by the respondent and his/her characteristics. This paper takes only dimensions of vignettes into account (at the lowest level). Therefore, and due to a homogenous sample we refer to the data structure by the calculation of robust standard errors (Wooldridge 2003: 258ff.; Wooldridge 2002; for model selection in vignettes studies: Jasso 2006; Auspurg/Abraham/Hinz 2008; Hox/Kreft/Hermkens 1991). The analysis and explanation of variance in the fundamental judgments between subjects is not relevant for our interests. Further and detailed remarks on strategies of analysis and operationalization of concepts (e.g. how to define an implausible case) are made in the following section on results.

5. Results

5.1 Descriptive Results

[13] Due to the employed sampling method no response rates can be reported. Again, it is to take into account that we only aim at a test of methodological hypotheses, and certainly not at a representative survey of justice evaluations. For this reason the lack of a random sample seems to be unproblematic. Repeated participations were prevented technically as good as possible.

Before testing the hypotheses with multivariate models we present some descriptive results on response behavior. 124 from 558 respondents did not complete the questionnaire (22.2% dropouts). Dropouts concentrate on the introductory page and the questionnaire before the vignette part. Only 23 respondents (4.1 percent of all respondents) quitted within the vignette-section. Due to this very low dropout rate a test of hypotheses 1 and 4 is not adequate. In light of these results we can however state that the evaluation task under the condition of twelve dimensions and the existence of implausible cases within the presented vignette sample does not lead to substantial dropouts. This is also true for non-responses within the vignette-part: only 68 vignettes (1.9 percent) were not evaluated.

Our preceding remarks show that a low degree of willingness to cooperate or a cognitive overload within the vignette part may instead affect the response behavior and the quality of collected data: „If tasks are too long or too difficult or lack sufficient realism and credibility, data quality will suffer in the sense of not containing the information sought. Unfortunately, respondents generally answer the questions asked and seldom go out of their way to point out problems with tasks posed" (Carson et al. 1994: 355). A first hint of possible response-sets gives the distribution of vignette evaluations over experimental groups in Table 2. We observe a slightly lower variance of evaluations between and within respondents in the twelve compared to the five dimension condition, meaning that in the twelve dimension condition respondents show more constant response behavior – but this difference is only significant at the 10 and not at the 5 percent level.[14]

- Table 2 about here -

5.2 Multivariate Analyses

A high complexity of the evaluation task may lead to an inconsistent response behavior (H_{1b}) – technically: low explained variances and high error variances –, and a fading out of single dimensions (H_{1c}), i.e. weak effect sizes and less numbers of significant effects in the models. These assumptions are tested by OLS-regression models, as the data have a hierarchical structure we estimate coefficients with robust standard errors. To be able to separate method-effects from third-variable-effects the results for the twelve dimensions condition are reported without (model 2) and with the additional dimensions (model 3).

From the coding of the dependent variable follows that positive (negative) coefficients denote that the income of the respective vignette person is evaluated as unjustly too high (too low). Hence, negative coefficients are synonym with the view that the vignette person should earn more, i.e. his/her income should be higher. For example, in all three models a person with vocational training should earn more than a person with no completed education. For the methodological aims of this paper it is

[14] T-Test for mean differences of the respondent specific evaluation variance for five vs. twelve dimensions: t = 1.478; p = .1400 in a two-tailed test with correction for the violated assumption of unequal variances (Levene's Test).

more important observing differences between the coefficients across the three models. The sign of the coefficients does not change across the models. But for most of the coefficients, their absolute values are lower in the twelve dimension condition than in the five dimension condition. A Chow-test confirms significant differences between model 1 and 2 (F = 4.04, df = 7, N = 459; p= .0003). Testing the interaction effects between number of dimensions and the single vignette characteristics (dimensions) leads to significant differences for „university degree“ and “occupational prestige”. As the effects of these two variables remain stable when controlling for the other dimensions the difference is not a third-variable-effect but may be the result of a diminishing relevance of certain dimension under conditions of high complexity. In contrast the explained variances (R^2-values) as a proxy for consistent evaluation behavior do not differ substantially between models.

- Table 3 about here -

As we find only small evidence for a cognitive over-load our first hypothesis is only partly confirmed. But our results also show that effect sizes from factorial surveys should be interpreted with caution, in particular if there is no way found to control for the complexity of vignettes. There is another implication of our findings: In some studies high R^2-values are used as an indicator for having employed those characteristics which are important for respondents' evaluations (cf. Beck/Opp 2001: 302). Our results show that this may be a false conclusion. As model 3 shows that all variables for additional characteristics have significant effects but they do not enhance the explained variance. Therefore a good model fit in terms of R^2 is a measure of consistent response behavior but not necessarily an indicator for a substantial exhaustive explanation of response behavior. Another important point is that respondents in our study were only confronted with maximal ten different vignettes. Cognitive overstrain may only be observed when respondents have to evaluate a higher number of vignettes.

In hypothesis 2 it is stated that the sequencing has an effect on the consistency of responses, which should be especially true for the twelve dimension condition. To test this assumption we report regression estimations separately for the positions of vignettes. Figure 2 shows for both – five and twelve – conditions the R^2-values depending on the order position of the vignettes (dark lines above). As these results may also be instructive for learning and exhaustion effects we additionally report the mean response time per vignette (light lines beneath).

- Figure 2 about here -

Regarding the explained variances we observe a slight increase during the response sequence. However, the mean response time per vignette is declining – especially after the first vignette – rapidly and with decreasing rate up to the seventh vignette. This shows a learning effect: respondents are more and more able to evaluate the vignettes consistently in less time. Contrary to the assumption in hypothesis H_{2b} this effect is not stronger in the twelve dimension condition, i.e. learning effects are quite the same regardless if respondents are confronted with five or twelve dimensions. But this interpretation holds only if we can show that the increasing consistency is not the result of a fading out of dimensions or a simplified decision heuristic. Therefore separate regressions were estimated with the first, the second and the last third of vignettes. The results of these regressions (not reported) show that the model estimations do not differ significantly, this means that the number of influential dimensions, the effect sizes, and the response patterns are stable over the sequence of vignettes. Despite the high complexity of twelve dimensions, the first vignettes already cause reliable evaluations, this means that the first vignettes do not have to be regarded as mere "exercises" which should be eliminated from the analyses (see Caussade et al. 2005: 632). But as we have a homogenous respondent sample (regarding age and education) evaluating only a fairly small number of vignettes these results might be biased and should be tested in further studies.

In order to test hypotheses 3 we need, first, a clear definition of implausible vignettes and, second, an adequate method. Regarding the definition of implausible vignettes we rely on the results of 60 pretest interviews using the same vignette sample as in the study at hand. In this pretest the oral reactions of respondents on each single vignette were documented. Most of the respondents considered those vignettes as "implausible" where occupations for which a university degree is required (e.g. lawyer) appear in combination with "no educational degree" or "vocational degree". Accordingly we define those vignettes as implausible where the educational degree does not match the requirements of the respective occupation in Germany, which is true for 15.9 percent. Moreover, we select a second type of implausible vignettes using combinations of income and occupation. On the basis of the German Socio-Economic Panel 2007 (GSOEP, see Wagner/Frick/Schupp 2007) we calculated the mean of the actual income for those occupations we are using in our vignette sample. We define those vignettes as implausible, where the absolute difference between the fictitious income (vignette sample) and the real mean income (GSOEP) amounts at least 3,000 Euros (25 percent of the vignette universe are in this respect implausible).

Additionally we have to choose an adequate analytical strategy to identify the effects of the two types of implausible cases (occupation-education, occupation-income) on respondent behavior. The problem is that a straightforward comparison of regression estimates would neglect the fact that the independent variables in the two conditions (plausible vs. implausible vignettes) have different variances. In the case of plausible cases we have per se higher correlations and lower variances of the vignette dimensions. And, the discrepancies in the number of cases will complicate comparisons (implausible cases build only a minority). Moreover, our theoretical assumption is that we only observe changes in the response behavior from the point on an implausible case occurred in the sequence of vignettes. More precisely: Respondents will show a less consistent (H_{4b}) or a more simplified (H_{4c}) response behavior *after* they were confronted with an implausible vignette. To detect these kinds of reactions we need to compare the response behavior up to and from the first appearance of an implausible vignette.[15]

Following this strategy we list regression estimates in table 4 for the two types of implausible vignettes: Model 1 and 2 reports the results for the occupation-education and model 3 and 4 for the occupation-income implausible vignettes. Model 1 and model 3 is based on the responses before the first occurrence of an implausible case, model 2 and 4 lists the estimates for all responses from the point on an implausible vignette appeared. As model 2 and 4 are related to responses in further sequence, we control for the order of vignettes. A Chow test exhibits the model differences as significant for both types of implausible vignettes (F=2.06, p= .046, df = 7/459; resp. F=71.37, p =.000, df = 7/459). In case of the occupation-education type – besides a level effect, the constant in model 2 is smaller compared to model 1 – this is due to the education dimension: the effect of vocational degree differs on the ten, and the effect of the university degree on the five percent level (tested by a joint model with interaction terms). In contrast in case of the occupation-income type only the income dimension is significant (model 3 and 4). For both types of implausible vignettes we observe the same direction of differences: From the occurrence of an implausible case on, the effects become weaker. This is against our assumption, but it is coherent: After they were confronted with an implausible case respondents fade out or exclude the respective dimensions – or in other words: the dimensions causing the implausibility are not taken seriously anymore. If these results should be confirmed in other studies the strong recommendation should be to avoid those implausible cases.

- Table 4 about here -

Our expectations regarding the consistency of responses are also only partly confirmed. After an implausible case the consistency of responses – measured by R^2-values – declines only slightly for the occupation-income type, but in case of the occupation-education type we observe a small increase of R^2. In the light of these contradictory results we seek broader empirical evidence to assess the validity of our hypotheses. Using multivariate models we test the effects of implausible cases on response consistency and response time. As a measure for consistency we use the unexplained variance – more precisely the squared residuals – and regress them on our design variables i.e. sequential order of vignettes, twelve vs. five dimension condition, response time. Negative effects denote a better model fit respectively lower inconsistencies in response behavior. Table 5 shows that there are only two

[15] With this strategy the problem of different variances is attenuated but not dispelled. Under the conditions of implausible cases all the responses are related to vignette characteristics with greater variance and smaller correlations – as a consequence of not restricting any combinations – which leads to a higher precision of the estimation, i.e. more power to detect significant effects. Therefore stronger effect sizes of those dimensions which constitute an implausible case do not necessarily reflect the fact that respondents are more focusing on these dimensions – as Faia (1980) stated – but might as well reflect mathematical and statistical effects (cf. Creyer/Ross 1988). For other explanations of changing response behavior after implausible cases see Ohler/Louviere/Swait 2000; Louviere 2001a; Wittink/Krishnamurthi/Nutter 1982, Wittink/Krishnamurthi/Reibstein 1989; Perrey 1996.

significant effects (at a 10% level, see model 1): From the occurrence of implausible cases on the inconsistency diminishes (which we interpret as an evidence for a simplified response behavior).

- Table 5 about here -

Model 2 shows that the response time decreases with the sequential position of the vignettes meaning that respondents are more and more able to make time efficient judgments. The significant quadratic term points to an „u-shaped"-effect.[16] In addition the more time-consuming processing of vignettes with a higher number of dimensions is reconfirmed. But implausible cases – as we operationalize them by the occupation-education type – do not produce a sketchily response behavior, at least measured in terms of time.[17]

6. Summary and Conclusions

The factorial survey as a method of data collection is well established in sociological research on norms and attitudes. Further sociological research fields (e.g. research on discrimination) use the technique as an innovative survey method. Despite of the increase of the method's application there are only a few methodological studies. In particular practical criteria for the application of the method in survey research are still missing (see Beck/Opp 2001: 283). Moreover, there are substantial doubts on internal and external validity. Artificial judgments might result from cognitive over-loads and/or from the application of simplified heuristics. As long as methodological studies are rare the results from vignettes studies are hardly reliable. The risk of producing methodological artifacts seems rather high.

For a start, this paper aims at a study of stability and consistency of responses dependent on the design of the factorial survey, i.e. complexity, sequence and plausibility of vignettes. An experimental online-survey with about 400 students enables differentiated analyses on the impact of the number of dimensions, learning effects and implausible cases. A general result of our research is that all three aspects matter for response patterns.

Our analyses demonstrate that complexity of vignettes measured by the number of dimensions significantly influences the judgments. The effects of single dimensions are weakened as the number of dimensions increases. This means that the interpretation of absolute effects may not be stretched too far. The effect of single dimension seems to be a function of its "uniqueness" in the vignettes presented. This should be considered if results from different studies are compared. To put it in another way: reliable comparisons of effects from different surveys need at least similarly complex

[16] According the model estimates the turning point would be at the tenth vignette. As we present our respondents only ten vignettes, we can not proof this result.

[17] In the occupation-income type we find a significant decrease of the response time (in average .016 seconds), which shows that a more subtle test of our hypotheses is needed.

vignettes. The heuristics which are observed for a higher level of complexity would be artificial as far as they do not match with real judgments (see Swait/Academowitz 2001:147). Non-significant results should be validated with vignettes using a lower number of dimensions. Good news is that the overall effects on the respondents' behavior are not very dramatic. Respondents (in our case: students) can obviously handle a rather high complexity of twelve dimensions. Because of the interdependence with other design variables (as the number of levels) and with characteristics of the respondents (cognitive ability) more detailed research is needed.

Methodological studies on conjoint analyses and choice experiments support the existence of a complex relationship of learning and exhaustion effects. In order to depict these effects completely the number of vignettes used in our study is too low. A striking result, however, is the clear dominance of learning effects until the last of our vignettes (the tenth vignette). Learning is reflected primarily by a higher response speed at a stable rate of consistent responses (R²). The evaluations made on the very first vignettes are already reliable and confirmed by the later vignettes. This result supports the principal reliability of the first "unpracticed" judgment and the validity of studies which implement a between subject design (in an extreme case: a single vignette per respondent). The question how the trade off between the quantitative gains in judgments and the loss of data quality in a within subject design is defined can be answered only with more vignettes per deck. In further research it should be tested whether our results are confirmed by other forms of complexity and more heterogeneous samples of respondents.

Praised as a specific strength of the method (how to evaluate a deviant situation), critics argue that empirically rare, „virtual" vignettes lead to artificial judgments. This skeptical view is supported by our analysis. Implausible combinations of variable values do not yield a drastic increase in dropouts or non-responses (the quota of dropouts and non-responses is very low in general) but a continued response with a significantly lower relevance of the dimensions causing the implausibility. This is very important concerning the interpretation of coefficients. The lack of or low levels of statistical significance for single variables might be caused in consequence of the lack of plausibility in the variable's values and not by their genuine irrelevance. Should this result be confirmed in further studies the practical advice would be to forego such implausible cases – at least to use them parsimoniously. Thanks to software providing powerful algorithms the loss of efficiency due to the exclusion of implausible cases can be reduced to an acceptable level.

Our analyses additionally demonstrate that the discussion of methodological issues requires multiple criteria. Simplified decision rules may contribute to a high quality of measurement – evaluated by the explained variance, which however do not cover the actual attitudes and decision rules of respondents.[18] The central conclusion is that R² values are usable indicators for the consistency of judgments but not for the extent to which we are successful in identifying all dimensions relevant for the evaluation. Threatened by cognitive over-load the respondents stay with a

[18] If very few dimensions are considered, a high reliability/consistency is no great cognitive achievement. As an indicator for a valid measurement a high R^2 value is therefore not sufficient.

high consistency but judgments become less adequate. Thus, factorial surveys are suitable instruments if the relevance of single variables should be studied (e.g. for testing hypotheses) but they are less adequate if exhausting rules of judgment are to be detected. It is possible to make statements on the relevance of the considered dimensions but it is not possible to draw conclusion on the irrelevance of further dimensions even if the R^2 values are very high. This result again underlines the prominent role of the theoretical selection of dimensions.

Because of the variety of further methodological problems and the expected interdependence with other design variables we achieve only a first impression of the importance of design elements. In addition, there are other methodological issues which we did not discuss as the mode of presentation (paragraph or tabulated description) and the usage of different scales. Moreover, the statistical robustness of the results should be studied. According to mostly used strategies, we defined the judgment on a pseudo-metric scale. In reality, they are measured on an ordinal scale. OLS models are seen as relatively robust against violations of assumptions (Windship/Mare 1984) but there is too little research on possible gains in analysis when models are applied that are better suited to the data structure (multi level analyses).[19] The ongoing project with its broad data collection will help to answer those open questions. According to our first results these questions need further methodological and theoretical attention.

References

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

Auspurg, K. and M. Abraham, 2007: Die Umzugsentscheidung von Paaren als Verhandlungsproblem. Eine quasiexperimentelle Überprüfung des Bargaining-Modells. Kölner Zeitschrift für Soziologie und Sozialpsychologie 59: 271-293.

Auspurg, K., M. Abraham and Th. Hinz, 2008: Wenig Fälle, viele Informationen: Die Methodik des Faktoriellen Surveys als Paarbefragung. In Kriwy, P. und Ch. Groß (Eds.): Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen (pp. 179-210). Wiesbaden: VS Verlag für Sozialwissenschaften.

Beck, M. and K.-D. Opp, 2001: Der faktorielle Survey und die Messung von Normen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 283-306.

Berk, R. A. and P. H. Rossi, 1977: Prison reform and state elites. Cambridge, Mass.: Ballinger.

Buskens, V. und J. Weesie, 2000: An Experiment on the Effects of Embeddedness in Trust Situations: Buying a Used Car. Rationality and Society 12: 227-253.

Bradley, M. and A. Daly, 1994: Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation 21: 167-184.

Carroll, D. J. and P. E. Green, 1995: Psychometric Methods in Marketing Research: Part I, Conjoint Analysis. Journal of Marketing Research 32: 358-391.

[19] At least with close-ended scales the repeated task of judgments may lead to an unintended censoring of the judgments (when there have been already extreme evaluations, the judgments can no longer be differentiated in an adequate way). This may lead to biased results and also suggests employing special regression models (e.g. Tobit regressions). This may be especially the case when many implausible cases occur, since they particularly motivate to extreme answers.

Carson, R ., J. J. Louviere, D. A. Anderson, Ph. Arabie, D. Bunch, D. A. Hensher, R. M. Johnsons, W. F. Kuhfeld, D. Steinberg, J. Swait and H. Timmerman, 1994: Experimental Analysis of Choice. Marketing Letters 5: 351-368.

Caussade, S., J. de D. Ortúzar, L. I. Rizzi and D. A. Hensher, 2005: Assessing the influence of design dimensions on stated choice experiment estimates. Transportation Research Part B 39: 621-640.

Creyer, E. and W. T. Ross, 1988: The Effect of Range-Frequency Manipulations on Conjoint Importance Weight Stability. Advances in Consumer Research 15: 505-509.

DeShazo, J.R. and G. Fermo, 2002: Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. Journal of Environmental Economics and Management 44: 123-143.

Diefenbach, H. and K.-D. Opp, 2007: When and Why do People Think there should be a Divorce? An Application of the Factorial Survey. Rationality and Society 19: 485-517.

Diekmann, A., 2007. Empirische Sozialforschung. Reinbek b. Hamburg: Rowohlt

Dülmer, H., 2001: Bildung und der Einfluss von Argumenten auf das moralische Urteil: Eine empirische Analyse zur moralischen Entwicklungstheorie Kohlbergs. Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 1-27.

Dülmer, H. and M. Klein, 2003: Die Messung gesellschaftlicher Wertorientierungen via Conjoint- und Vignettenanalyse: Ein Ansatz zur adäquaten Operationalisierung von Ingelharts materialistischen und postmaterialistischen Wertorientierungen. Unpublished manuscript.

Dülmer, H., 2007: Experimental Plans in Factorial Surveys: Random or Quota Design? Sociological Methods & Research 35, 382-409.

Eifler, S., 2007: Evaluating the Validity of Self-Reported Deviant Behavior Using Vignette Analyses. Quality & Quantity 41: 303-318.

Faia, M., 1980: The Vagaries of the Vignette World: A Comment on Alves and Rossi. American Journal of Sociology 85: 951-954.

Garrett, K., 1982: Child Abuse: Problems of Definition. In Rossi, P. H. und S. L. Nock (Eds.): Measuring Social Judgements. The Factorial Survey Approach (pp. 177-204). Beverly Hills: Sage.

Greene, W. H., 2003: Econometric Analysis. Prentice Hall: New York.

Groß, J. and Ch. Börensen, 2008: Wie valide sind Verhaltensmessungen mittels Vignetten? Ein methodischer Vergleich von faktoriellem Survey und Verhaltensbeobachtung. In Kriwy, P. und Ch. Groß (Eds.): Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen (pp. 149-178). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hechter, M., J. Ranger-Moore, G. Jasso and Ch. Horne, 1999: Do Values Matter? An Analysis of Advance Directives for Medical Treatment. European Sociological Review 15: 405-430.

Hechter, M., H. Kim and J. Baer, 2005: Prediction Versus Explanation in the Measurement of Values. European Sociological Review 21: 91-108.

Hembroff, L. 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, P. L. J. and F. A. Boerman, 1989: Consensus with respect to the fairness of incomes: Differences between social groups. Social Justice Research 3: 201-215.

Hensher, D. A., 2004: How do Respondents Handle Stated Choice Experiments? Information processing strategies under varying information load. Working Papter DoD#5, University of Sydney: Institute of Transport Studies.

Hensher, D. A., 2006: Revealing Differences in Willingness to Pay due to the Dimensionality of Stated Choice Designs: An Initial Assessment. Environmental & Resource Economics 34: 7-
Hermkens, P. L. J. and F. A. Boerman F. A., 1989: Consensus with respect to the fairness of incomes: Differences between social groups. Social Justice Research 3: 201-215.

Horne, Ch., 2003: The Internal Enforcement of Norms. European Sociological Review 19: 335-343.

Hox, J. J., Kreft, I. and Hermkens, P., 1991: The Analysis of Factorial Surveys. Sociological Methods & Research 19: 493-510.

Jann, B., 2003: Lohngerechtigkeit und Geschlechterdiskriminierung: Experimentelle Evidenz. Unpublished manuscript oft he ETH Zürich.

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

Jasso, G., 1994: Assessing Individual and Group Differences in the Sense of Justice: Framework and Application to Gender Differences in the Justice of Earnings. Social Science Research 23: 368-406.

Jasso, G. and M. Webster, 1997: Double Standards in Just Earnings for Male and Female Workers. Social Psychology Quarterly 60: 66-78.

Jasso, G. and M. Webster, 1999: Assessing the Gender Gap in Just Earnings and Its Underlying Mechanisms. Social Psychology Quarterly 62(4): 367-380.

Jasso, G. and K.-D. Opp, 1997: Probing the Character of Norms: A Factorial Survey Analysis of the Norms of Political Action. American Sociological Review 62: 947-964.

Jasso, G., 2006: Factorial-Survey Methods for Studying Beliefs and Judgments. Sociological Methods and Research 34: 334-423.

Johnson, R. F., 2006: Comment on "Revealing Differences in Willingness to Pay Due to the Dimensionality of Stated Choice Designs: An Initial Assessment". Environmental & Resource Economics (2006) 34: 45–50.

Klein, M., 2002: Die Conjoint-Analyse. Eine Einführung in das Verfahren mit einem Ausblick auf mögliche sozialwissenschaftliche Anwendungen. ZA-Information 50: 7-45.

Kuhfeld, W.F., T. D. Randall and M. Garratt, 1994: Efficient Experimental Design with Marketing Research Applications. Journal of Marketing Research 31, 545-557.

Kuhfeld, W. F., 2005: Marketing Research Methods in SAS. Experimental Design, Choice, Conjoint and Graphical Techniques. Cary: SAS Institute.

Liebig, S. and S. Mau, 2005: Wann ist ein Steuersystem gerecht? Zeitschrift für Soziologie 34, 468-491.

Liebig, S. and S. Mau, 2002: Einstellungen zur sozialen Mindestsicherung. Ein Vorschlag zur differenzierten Erfassung normativer Urteile. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 109-134.

Liebig, S., A. Meyermann, und A. Schulze, 2006: Temporal stability of justice evaluations. Paper presented at the 11th Conference of the International Society for Justice Research. Berlin: Humboldt Universität.

Louviere, J. J., 2001a: What if Consumer Experiments Impact Variances as well as Means? Response Variablility as a Behavioral Phenomen. Journal of Consumer Research 28: 506-511.

Louviere, J. J., 2001b: Choice Experiments: an Overview of Concepts and Issues. InBennet, J. and R. Blamey (Eds.): The Choice Modelling Approach to Environmental Valuation (pp. 13-36). Cheltenham, Northhampton: Edward Elgar.

Mayerl, J., P. Sellke and D. Urban, 2005: Analyzing cognitive processes in CATI-Surveys with response latencies: An empirical evaluation of the consequences of using different baseline speed measures. Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart, SISS No. 2/2005.

Melles, Th., 2001: Framing-Effekte in der Conjoint-Analyse. Ein Beispiel für Probleme der Merkmalsdefinition. Berichte aus der Psychologie 6. Aachen: Shaker.

Meudell, M. B., 1982: Household and Social Standing: Dynamic and Static Dimensions. In Rossi, P. H. und S. L. Nock (Hg.): Measuring Social Judgements. The Factorial Survey Approach (pp. 69-94). Beverly Hills u.a.: Sage.

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

Nisic, N. and K. Auspurg, 2008: Faktorieller Survey und klassische Bevölkerungsumfrage im Vergleich – Validität, Grenzen und Möglichkeiten beider Ansätze. In Kriwy, P. and Ch. Groß (Eds.): Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen (pp. 211-246). Wiesbaden: VS Verlag für Sozialwissenschaften.

Nock, S. L., 1982: Family Social Standing: Consensus on Characteristics. S. 95-118 in: Rossi, P. H. und S. L. Nock (Hg.): Measuring Social Judgements. The Factorial Survey Approach. Beverly Hills u.a.: Sage.

Ohler, T., A. Le, J. Louviere and J. Swait, 2000: Attribute Range Effects in Binary Response Tasks. Marketing Letters 11: 249-260.

Orme, B., 2006: Getting started with Conjoint Analysis: Strategies for Product Design and Pricing Research. Madison, Wisconsin: Research Publishers LLC.

O`Toole, R., S. W. Webster, A. W. O`Toole and B. Lucal, 1999: Teachers´ Recognition and Reporting of Child Abuse: A Factorial Survey. Child Abuse & Neglect 23: 1083-1101.

Perrey, J. 1996: Erhebungsdesign-Effekte bei der Conjoint-Analyse. Marketing – Zeitschrift für Forschung und Praxis 18: 105-116.

Rooks, G., W. Raub, R. Selten and F. Tazelaar, 2000: How Inter-firm Co-operation Depends on Social Embeddedness: A Vignette Study. Acta Sociologica 43: 123-137.

Rossi, P. H., 1979: Vignette Analysis: Uncovering the Normative Structure of Complex Judgments. In Merton, R. K., J. S. Coleman und P.H. Rossi (Eds.): Qualitative and Quantitative Social Research: Papers in Honour of Paul F. Lazarsfeld (pp. 176-186.). New York: Free Press,

Rossi, P. H. and W. M. Alves, 1980: Rejoinder to Faia. The American Journal of Sociology, Vol. 85: 954-955.

Rossi, P. H. and A. B. Anderson, 1982: The Factorial Survey Approach: An Introduction. In Rossi, P. H. S. L. Nock (Eds.): Measuring Social Judgments: The Factorial Survey Approach (pp. 15-67). Beverly Hills: Sage.

Rossi, P. H. and S. L. Nock, 1982: Measuring Social Judgements. The Factorial Survey Approach. Beverly Hills u.a.: Sage.

Seyde, Ch., 2005: Beiträge und Sanktionen in Kollektivgutsituationen: Ein faktorieller Survey. Arbeitsbericht Nr. 42 des Instituts für Soziologie. Leipzig: Universität Leipzig.

Shepelak, N. J. and D. F. Alwin, 1986: Beliefs about Inequality and Perceptions of Distributive Justice. American Sociological Review 51: 30-46.

Smith, T. W., 1986: A Study of Non-Response and Negative Values on the Factorial Vignettes on Welfare. GSS Methodological Report No. 44. Chicago: NORC.

Sniderman, P. M. and D. B. Grob, 1996: Innovations in Experimental Design in Attitude Surveys. Annual Review of Sociology 22: 377-399.

Steiner, P. M. and Ch. Atzmüller, 2006: Experimentelle Vignettendesigns in Faktoriellen Surveys. Kölner Zeitschrift für Soziologie und Sozialpsychologie 58, 117-146.

Struck, O., A. Krause and Ch. Pfeifer, 2008: Entlassungen: Gerechtigkeitsempfinden und Folgewirkungen.Theoretische Konzepte und empirische Ergebnisse. Kölner Zeitschrift für Soziologie und Sozialpsychologie 60: 102-122.

Swait, J. and W. Ademowicz, 2001: The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching. Journal of Consumer Research 28: 135-148.

Urban, D. and J. Mayerl, 2007: Antwortlatenzzeiten in der survey-basierten Verhaltensforschung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 59: 692-713.

Wagner, G.G., J. R. Frick and J. Schupp, 2007: The German Socio-Economic Panel Study (SOEP) - Evolution, Scope and Enhancements. Schmollers Jahrbuch (Journal of Applied Social Science Studies) 127: 139-169.

Wason, K. D., M. J. Polonsky and M. R. Hyman, 2002: Designing Vignette Studies in Marketing. Australasian Marketing Journal 10: 41-58.

Will, J. A., 1993: The Dimensions of Poverty: Public Perceptions of the Deserving Poor. Social Science Research 22: 312-332.

Windship, Ch. and R. D. Mare, 1984: Regression Models with Ordinal Variables.

Wittink D. R., L. Krishnamurthi and J.B. Nutter, 1982: Comparing derived importance weights across attributes. Journal of Consumer Research 8:471-474

Wittink D.R., Krishnamurthi L, and D.J. Reibstein, 1989: The Effect in Differences in the Number of Attribute Levels on Conjoint Results. Marketing Letters 1:113-123

Wooldridge, J. M., 2002: Econometric Analysis of Cross Section and Panel Data. Cambridge, Mass.: MIT Press.

Wooldridge, J. M., 2003: Introductory Econometrics. A Modern Approach. Mahson, Ohio: Thomson.

Zimbardo, Ph. G., 1988: Psychologie. Berlin u.a.: Springer.

Table 1: Number of observations (vignette judgments and respondents) realized for the different experiments[1]

	5 dimensions		12 dimensions		Σ	
	vignettes	respondents	vignettes	respondents	vignettes	respondents
Seven vignettes per resp.	1213	176	1109	162	2322	338
Ten vignettes per resp.	574	59	584	63	1158	122
Σ	1787	235	1693	225	3480	460

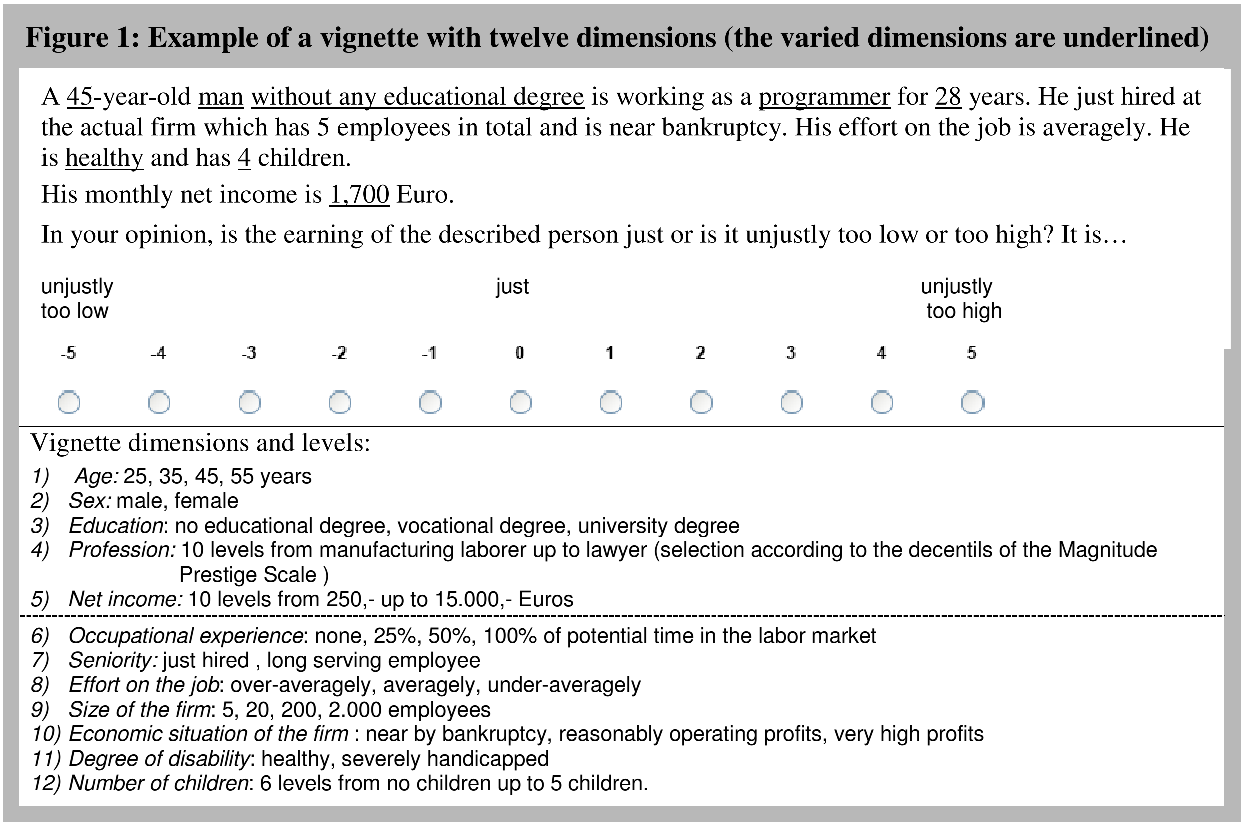
[1] Only respondents who evaluated at least one vignette.

Table 2: Descriptive overview over the vignettes judgments[1]

experimental split	number of cases	mean	standard deviation	average mean per respondent	average standard deviation per respondent
5 dimensions, 7 vignettes	1213	5,21	3,10	5,20	2,94
5 dimensions, 10 vignettes	574	5,44	3,21	5,45	3,12
12 dimensions, 7 vignettes	1109	5,51	2,96	5,51	2,87
12 dimensions, 10 vignettes.	584	5,36	2,98	5,35	2,86

[1] Scale from 1"unjustly too low" up to 11"unjustly too high", value 6 denotes a just earning.

Figure 1: Example of a vignette with twelve dimensions (the varied dimensions are underlined)



A 45-year-old man without any educational degree is working as a programmer for 28 years. He just hired at the actual firm which has 5 employees in total and is near bankruptcy. His effort on the job is averagely. He is healthy and has 4 children.

His monthly net income is 1,700 Euro.

In your opinion, is the earning of the described person just or is it unjustly too low or too high? It is…

unjustly too low					just					unjustly too high
-5	-4	-3	-2	-1	0	1	2	3	4	5
○	○	○	○	○	○	○	○	○	○	○

Vignette dimensions and levels:

1) *Age:* 25, 35, 45, 55 years
2) *Sex:* male, female
3) *Education*: no educational degree, vocational degree, university degree
4) *Profession:* 10 levels from manufacturing laborer up to lawyer (selection according to the decentils of the Magnitude Prestige Scale)
5) *Net income:* 10 levels from 250,- up to 15.000,- Euros

6) *Occupational experience*: none, 25%, 50%, 100% of potential time in the labor market
7) *Seniority:* just hired , long serving employee
8) *Effort on the job*: over-averagely, averagely, under-averagely
9) *Size of the firm*: 5, 20, 200, 2.000 employees
10) *Economic situation of the firm* : near by bankruptcy, reasonably operating profits, very high profits
11) *Degree of disability*: healthy, severely handicapped
12) *Number of children*: 6 levels from no children up to 5 children.

Table 3: OLS-regression of vignette judgements[1] (robuste standard errors in parentheses; significant differences between coefficients of model 1 and 2 shaded)[2]

	Model 1	Model 2	Model 3
	5 dimensions	12 dimensions	12 dimensions
Female vignette person	-0.057	-0.136	-0.105
	(0.122)	(0.115)	(0.113)
Age [years]	-0.021***	-0.029***	-0.020***
	(0.005)	(0.005)	(0.005)
Education (Ref.: no degree)			
- vocational training	-0.654***	-0.472***	-0.429***
	(0.133)	(0.131)	(0.129)
- university degree	-1.126***	-0.623***	-0.830***
	(0.129)	(0.126)	(0.130)
Occupational prestige [10 MPS-Score]	-0.157***	-0.097***	-0.106***
	(0.011)	(0.012)	(0.012)
Net income [in 100,- Euro]	0.060***	0.055***	0.058***
	(0.002)	(0.002)	(0.002)
Number of children			-0.152***
			(0.029)
Severely handicapped (Ref.: healthy)			0.049
			(0.114)
Occupational experience			0.066
			(0.048)
Seniority: long-serving (Ref.: just hired)			-0.645***
			(0.131)
Effort (Ref.: under average)			
- average			-0.813***
			(0.129)
- over average			-0.788***
			(0.138)
Size of the firm (number of employees in 100)			0.028***
			(0.006)
Economic situation of the firm (Ref.: near bankruptcy)			
- reasonable operating profit			-0.037
			(0.130)
- very high profits			-0.292**
			(0.122)
Constant	6.465***	6.274***	6.820***
	(0.280)	(0.236)	(0.272)
Observations:			
- vignettes	1787	1693	1693
- respondents	235	225	225
R²	0.47	0.45	0.49

[1] Scale from 1“unjustly to low“ to 11“unjustly to high “, value 6 denotes a just earning.

[2] Tested with interaction terms (vignette dimension x number of dimensions) in a pooled model, significant level: 5 percent

*** p<0.01, ** p<0.05, * p<0.1 in a two tailed test; estimation with robust standard errors.

Figure 2: R²-values (thick lines above) und processing time per vignette (lower lines) for position of vignette and number of dimensions

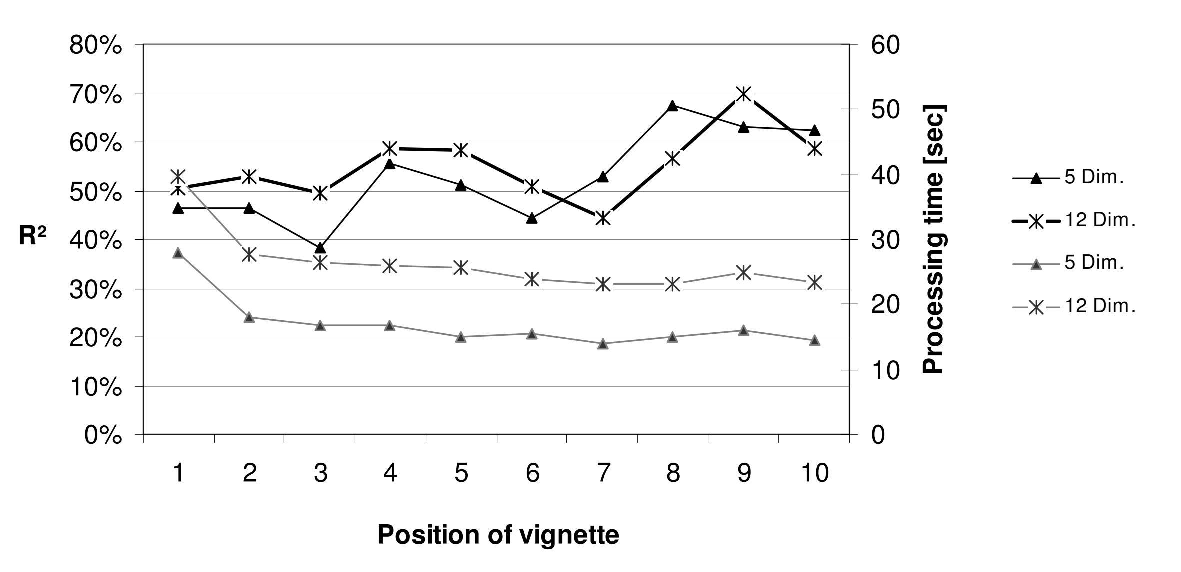


Table 4: OLS-regression of vignette judgments[1] before and after the occurrence of implausible cases (robust standarderrors in parenthesis; sign. differences of coefficients between Modell 1 and 2 resp. 3 and 4 shaded)[2]

	implausible cases			
	occupation-education		occupation-income	
	model 1 before	model 2 after	model 3 before	model 4 after
Female vignette person	-0.244*	-0.023	-0.178	-0.072
	(0.130)	(0.104)	(0.117)	(0.100)
Age [years]	-0.030***	-0.025***	-0.030***	-0.022***
	(0.006)	(0.004)	(0.005)	(0.004)
Education (Ref..: no degree)				
- vocational training	-0.972***	-0.571***	-0.671***	-0.670***
	(0.202)	(0.105)	(0.145)	(0.107)
- university degree	-1.386***	-0.784***	-1.101***	-0.962***
	(0.182)	(0.117)	(0.134)	(0.109)
Occ. prestige [10 MPS-Score]	-0.135***	-0.114***	-0.126***	-0.149***
	(0.015)	(0.010)	(0.013)	(0.010)
Net income [in 100 Euro]	0.058***	0.057***	0.161***	0.055***
	(0.002)	(0.002)	(0.005)	(0.001)
Position of vignette[3]	0.004	0.063***	-0.029	0.071***
	(0.032)	(0.021)	(0.029)	(0.021)
Constant	7.152***	5.813***	5.021***	6.124***
	(0.353)	(0.241)	(0.277)	(0.252)
Observation:				
- vignettes	1301	2179	1197	2283
- respondents	355	400	344	409
R-squared	0.44	0.48	0.56	0.52

[1] Scale from 1"unjustly to low" to 11"unjustly to high ",value 6 denotes a just earning.

[2] Tested with interaction terms (vignette dimension x number of dimensions) in a pooled model, significant level: 5 percent.

[3] Metric variable; testing with dummy-coding yields the same results.

*** p<0.01, ** p<0.05, * p<0.1in a two-tailed test; estimation with robust standard errors.

Table 5: OLS-regression of squared residuals[1] and response time[2] per single vignette (robust standard errors in parenthesis)

	model 1 **squared residuals[1]**	**model 2** **response time [2]**
Sequential order of vignette	-0.106	-0.064***
	(0.197)	(0.004)
Sequential order of vignette, squared	0.006	0.005***
	(0.018)	(0.000)
Twelfe dimensions (Ref.: five)	-0.123	0.119***
	(0.300)	(0.008)
After occurrence of an implausible case[3]	-0.564*	-0.004
	(0.306)	(0.008)
Response time per vignette[1]	-1.461*	
	(0.865)	
Constant	6.102***	0.371***
	(0.554)	(0.010)
Observation:		
- vignettes	3095	3095
- respondents	416	416
R-squared	0.00	0.26

[1] Residuals of an OLS-regression of the vignettes evaluation on the first five dimensions.

[2] Response time in seconds, per respondent weighted with the response time for a evaluation task on a analogous item scale (thereby controlling for the "baseline-speed", see Mayerl/Sellke/Piet 2005; Urban/Mayerl 2007). For both response time measures outliers (the upper five-percent-percentile) were excluded. Therefore the number of cases is lower in the analyses presented here.

[3] Operationalized by an implausible combination of occupation and education.

*** p<0.01, ** p<0.05, * p<0.1 in a two-tailed test; estimation with robust standard errors.