Experimental Vignette Studies in Survey Research

Article *in* Methodology European Journal of Research Methods for the Behavioral and Social Sciences · January 2010

DOI: 10.1027/1614-2241/a000014

Christiane Atzmüller
University of Vienna

Original Article

Experimental Vignette Studies in Survey Research

Christiane Atzmüller[1] and Peter M. Steiner[2,3]

[1]University of Applied Sciences, Vienna, Austria, [2]Institute for Advanced Studies, Vienna, Austria, [3]Northwestern University, Evanston, IL, USA

Abstract. Vignette studies use short descriptions of situations or persons (vignettes) that are usually shown to respondents within surveys in order to elicit their judgments about these scenarios. By systematically varying the levels of theoretically important vignette characteristics a large population of different vignettes is typically available – too large to be presented to each respondent. Therefore, each respondent gets only a subset of vignettes. These subsets may either be randomly selected in following the tradition of the factorial survey or systematically selected according to an experimental design. We show that these strategies in selecting vignette sets have strong implications for the analysis and interpretation of vignette data. Random selection strategies result in a random confounding of effects and heavily rely on the assumption of no interaction effects. In contrast, experimental strategies systematically confound interaction effects with main or set effects, thereby preserving a meaningful interpretation of main and important interaction effects. Using a pilot study on attitudes toward immigrants we demonstrate the implementation and analysis of a confounded factorial design.

Keywords: factorial survey, experimental vignette design, fractional factorial design, confounded factorial design, ANOVA, multilevel analysis

In quantitative research, the combination of the vignette technique with a traditional survey is a promising but too infrequently used research method for investigating respondents' beliefs, attitudes, or judgments. Vignette studies combine ideas from classical experiments and survey methodology to counterbalance each approach's weakness. On the one hand, traditional surveys show a high external validity which is mainly due to their claim of representativeness and their multivariate and multivalent measurements. However, this goes along with a low internal validity caused by the multicollinearity of measured variables and the passive way of taking measurements (i.e., without any experimental intervention or control of explanatory variables). On the other hand, classical experimental designs derive their high internal validity from orthogonal design plans and an active mode of measurement enabled by the controlled intervention. But single experiments have the drawback of low external validity which is mainly due to their nonrepresentativeness and oversimplified setting. Vignette studies try to overcome these limitations by combining the traditional survey with a vignette experiment. For Sniderman and Grob (1996, p. 378) this combination of the traditional representative survey and the vignette analysis with their different strengths in external and internal validity was one of the innovational breakthroughs in the design of public opinion surveys: "the availability of multifactorial, multivalent designs has encouraged a reorientation from narrowly methodological concerns to broader substantive issues."

Vignette Study: Vignette Experiment and Traditional Survey

A quantitative *vignette study* consists of two components: (a) a vignette experiment as the core element, and (b) a traditional survey for the parallel and supplementary measurement of additional respondent-specific characteristics, which are used as covariates in the analysis of vignette data. In this article, we solely discuss the vignette experiment.

A *vignette* is a short, carefully constructed description of a person, object, or situation, representing a systematic combination of characteristics. Figure 1 shows an example of a single vignette on attitudes toward immigration (Steiner & Atzmüller, 2006). This vignette characterizes an applicant for Austrian citizenship by a very specific combination of five factors: (*A*) country of origin (Nigeria, Iran, Hungary), (*B*) marital status (wife from Austria, wife from his country of origin, single), (*C*) official accusation (no complaints, failure of obligatory registration with the police, slight bodily assault), (*D*) language proficiency (very good German, broken German), and (*E*) occupational profession (unskilled worker, employee). The question to be answered by the respondents was: "Should this applicant become an Austrian citizen?" Students and old aged pensioners had to answer this question from their own perspective and the other group's perspective (i.e., what they think that students or pensioners would say).

no. 12

Mr. Miko is from **Hungary**.
He is **single** and speaks **broken German**.
He is currently working in Austria as an **employee**.
Once he was accused **of a slight bodily assault**.

I

strongly agree ○ ○ ○ ○ ○ ○ strongly disagree

Most pensioners

strongly agree ○ ○ ○ ○ ○ ○ strongly disagree

Figure 1. Vignette of an applicant for the Austrian citizenship. The question to be answered by a student was: "Should this applicant become an Austrian citizen?"

Within vignette studies, respondents are typically confronted not only with one single vignette but with a whole population of vignettes in order to elicit their beliefs, attitudes, judgments, knowledge, or intended behavior with respect to the presented vignette scenarios. In our example, the full factorial combination of all five factors with three and two factor levels results in a vignette population of $3 \times 3 \times 3 \times 2 \times 2 = 108$ different vignettes. Finally, the aim of a vignette study is to identify and assess the importance of those vignette factors which causally affect individual responses to the contextualized but hypothetical vignette settings.

In comparison to traditional survey items, the special design of vignettes enables the simultaneous presentation of several explanatory as well as contextual factors which leads to more realistic scenarios presented to respondents. As we discuss in detail below, the experimental variation of all vignette factors allows the estimation of unconfounded and context-dependent effects of explanatory vignette factors. For this reason, vignette studies are a very powerful tool for causal investigations of respondent judgements.

The flexibility of vignette studies also supports the realism of vignette measurements. Depending on the research question, vignettes can be presented to respondents in quite different forms, for instance, as text vignettes in keyword, dialog, or narrative style, or as cartoons, pictures, audio, or video vignettes. Since the mode of stimulus presentation affects individual judgments, the presentation mode must be carefully considered and, if possible, systematically controlled within an experimental setting.

Types of Vignette Experiments

The number of quantitative vignette studies increased over the last years in various fields of application, including psychology (Barrera & Buskens, 2002; Dülmer, 2001; Hunter & McClelland, 1991; Ludwick & Zeller, 2001; Thurman, Lam, & Rossi, 1988), sociology (Alves & Rossi, 1978; Beck & Opp, 2001; Jasso & Webster, 1999; Liebig & Mau, 2005; Mäs, Mühler, & Opp, 2005; Rossi & Nock, 1982; Rossi, Sampson, Bose, Jasso, & Passel, 1974), marketing (Wason, Polonsky, & Hyman, 2002), and education and training (Seguin & Ambrosio, 2002; Veal, 2002). In all these vignette studies, the total vignette population is obtained by a full factorial combination of all factors under investigation. Depending on the number of factors and factor levels, the overall size of vignette experiments varies from very small designs with only four vignettes (two factors with only two factor levels each) to rather huge designs with thousands or more vignettes (e.g., Rossi, Simpson, & Miller, 1985; Shlay, Tran, Weinraub, & Harmon, 2004). While small experimental vignette designs have a long tradition in psychology (e.g., Jones & Aronson, 1973; Walster, 1966), large designs are typical for vignettes used within survey research in social sciences (see references above).

As with experimental designs, three general types of vignette experiments can be distinguished: (a) within-subjects designs, (b) mixed designs, and (c) between-subjects designs. In within-subjects designs each respondent judges exactly the same set of vignettes. The vignette set may either represent the total vignette population, if it is small enough (e.g., Pizarro, Uhlmann, & Bloom, 2003), or a subset of a larger vignette population. In mixed designs different groups of respondents get different vignette sets but within each group each respondent receives the same vignettes for judgment. In between-subjects designs each respondent judges only one single vignette, but they are quite rare since serious measurement problems arise, due to different vignette contexts (e.g., Birnbaum, 1999).

In this article we focus on large vignette experiments, that is, experiments with vignette populations that are too large to be judged by each respondent. Such large vignette populations are typical in survey research and arise when substantive theory does not allow a reduction of the population by dropping some factors or factor levels. Hence, other strategies for restricting the number of vignettes presented to each respondent must be applied. Two strategies are possible: First, the selection of a small subsample of the total vignette population, and second, the partitioning of the vignette population into respondent-specific sets. We discuss these strategies with respect to two main approaches dominant in survey research: (i) the factorial survey and (ii) the experimental vignette design.

Most of the quantitative vignette studies in the social sciences have been undertaken according to the *factorial survey* which is mainly characterized by *randomly* selecting subpopulations or vignette sets. Factorial surveys were first applied by Rossi and his colleagues in 1974 (Rossi, Waite, Bose, & Berk, 1974; Rossi, Sampson, et al., 1974). Rossi himself attributes the origin of the vignette analysis to a comment by Lazarsfeld on his dissertation in 1951 (cf. Rossi, 1979). He called the technique factorial survey since it "combines ideas from balanced multivariate experimental designs with sample survey procedures" (Rossi & Anderson, 1982, p. 15). Important contributions to the factorial vignette survey have been made by Jasso (for an overview, see Jasso, 2006), but also by Auspurg, Hinz, and Liebig (2009) and Dülmer (2007).

In contrast, *experimental vignette designs* use classical experimental designs instead of random sampling for the selection of respondent-specific vignette sets. Main methodological contributions were made in 1977 by Cook (1977, 1979) and 1978 by Alexander and Becker (1978) who explicitly introduced confounded and fractional factorial designs for vignette studies in survey research (see also Atzmüller, 2006; Steiner & Atzmüller, 2006).

The aim of this article is to discuss different strategies for selecting subpopulations or vignette sets in large vignette studies with respect to their implications for the analysis and interpretation of vignette data. We start with a more theoretical discussion of these selection strategies and the statistical methods for analyzing vignette data. Then, the practical implementation of an experimental vignette study according to a confounded factorial design is demonstrated using a pilot study on attitudes toward immigration (Steiner & Atzmüller, 2006). We practically identify the most important steps of a vignette study: (i) construction of the total vignette population, (ii) determination of the number of vignette sets and vignettes per respondent, (iii) construction of vignette sets, (iv) sampling of respondents and collection of data, and (v) analysis of vignette data and interpretation of results.

Strategies for Selecting Vignette Subpopulations and Sets

Constructing the Vignette Population

The first step in designing a vignette study is the construction of a whole population of different vignettes, which should be presented to respondents for eliciting their beliefs, attitudes, or judgments. For this purpose, the relevant factors and appropriate factor levels for the research hypothesis under investigation have to be determined according to a strong substantive theory. If there is only a weak theory, quantitative as well as qualitative preliminary investigations should be undertaken to identify the factors of relevance.

Table 1. Full factorial $2 \times 2 \times 2$ vignette design consisting of eight vignettes

	$C = c_0$		$C = c_1$	
	$B = b_0$	$B = b_1$	$B = b_0$	$B = b_1$
$A = a_0$	$a_0b_0c_0$	$a_0b_1c_0$	$a_0b_0c_1$	$a_0b_1c_1$
$A = a_1$	$a_1b_0c_0$	$a_1b_1c_0$	$a_1b_0c_1$	$a_1b_1c_1$

Once the relevant factors and their levels are determined, the vignette population can be constructed by systematically varying the factor levels for each factor. Suppose that there are k factors $V_1, \ldots, V_k$ with v_i factor levels ($i = 1, \ldots, k$), then the total vignette population is obtained by a full factorial design, that is, a complete combination of all vignette factors $V_1 \times V_2 \times \ldots \times V_k$. Consequently, the total number of possible vignettes is given by $m = v_1 \times v_2 \times \ldots \times v_k$.

Table 1 shows a simple example of a $2 \times 2 \times 2$ full factorial vignette design, with three factors A, B, C and two factor levels each ($a_0, a_1; b_0, b_1; c_0, c_1$). Hence, the total vignette population consists of eight vignettes. If each respondent judges all eight vignettes all main effects as well as all two-way and three-way interaction effects can be estimated. According to the full factorial design, main effect A is given by the difference $a_0 - a_1 = (a_0b_0c_0 + a_0b_1c_0 + a_0b_0c_1 + a_0b_1c_1) - (a_1b_0c_0 + a_1b_1c_0 + a_1b_0c_1 + a_1b_1c_1)$. Similarly, interaction effect AB is given by $(a_0b_0c_0 + a_1b_1c_0 + a_0b_0c_1 + a_1b_1c_1) - (a_1b_0c_0 + a_0b_1c_0 + a_1b_0c_1 + a_0b_1c_1)$ and interaction ABC by $(a_0b_0c_0 + a_1b_1c_0 + a_1b_0c_1 + a_0b_1c_1) - (a_1b_0c_0 + a_0b_1c_0 + a_0b_0c_1 + a_1b_1c_1)$. The remaining main and interaction effects can be analogously determined.

However, in most practical applications in survey research the total vignette population is too large to be presented to each respondent, such that a selection of a subpopulation or subsets of the vignette population is necessary. Two main strategies for selecting subsets from the vignette population exist: First, the selection of a single but much smaller *subpopulation* of the total vignette population, and second, the *partitioning* of the total vignette population into smaller vignette sets.

Selecting a Vignette Subpopulation

For the selection of a vignette subpopulation, which should be small enough that each vignette can be judged by each respondent, two techniques can be applied. Either a specific experimental design, called fractional factorial design, or a random selection strategy. It is important to recognize that in comparison to a full factorial design, which enables the estimation of all main and interaction effects, the selection of a subpopulation always results in a loss of information due to a confounding of effects. For instance, some main effects may be confounded with interaction effects, without any chance to separate effects. However, the selection strategy used – a fractional factorial design or a random selection procedure – determines the confounding structure and interpretation of effects.

Experimental Vignette Selection: Fractional Factorial Design

In a fractional factorial design only a fraction of the total vignette population is used and presented to each respondent (within-subjects design). Depending on the number of factors and factor levels the fraction of the total vignette population can be one half, one third, one fourth, on eighth, and so on (Alexander & Becker, 1978; Cochran & Cox, 1950; Kirk, 1995). The advantage of this design is that solely higher order interaction effects, which are typically negligibly small or equal to zero, are confounded with main effects.

Consider our illustrative example and assume that each respondent only judges the half of all eight vignettes. Table 2 shows the subpopulation of four vignettes selected according to a fractional factorial design. In this design, vignettes were selected such that all levels of each factor are equally often represented (two times) and two-way interaction effects are confounded with main effects as follows. The interaction effect *AB*, that is $(a_0b_0 + a_1b_1) - (a_1b_0 + a_0b_1)$, can be estimated neither within factor level c_0 nor within factor level c_1 of factor *C* since factor level combinations a_0b_0 and a_1b_1 are only measured together with c_0, and a_1b_0 and a_0b_1 only with c_1, However, the estimation of *AB* seems possible across the two factor levels c_0 and c_1. Since not all vignettes are measured the interaction effect reduces to $(a_0b_0c_0 + a_1b_1c_0) - (a_1b_0c_1 + a_0b_1c_1)$. But this is equivalent to the main effect of *C* $(c_0 - c_1)$. Hence, main effect *C* is confounded (aliased) with interaction effect *AB*. It is easy to verify that also *A* is confounded with *BC* and *B* with *AC*. The three-way interaction effect *ABC* cannot be estimated at all.

For larger designs, the confounding structure can be deliberately planned, that is, one can decide which interaction effects remain estimable and which ones will be confounded. In general, higher order interaction effects assumed to be zero are confounded with main effects, while those supposed to be of significance may be excluded from confounding. If nonzero interaction effects are confounded with main effects, a simple interpretation of main effects is misleading. Fortunately, interaction effects of order 3 and higher are often equal to zero or negligibly small such that the confounding with main effects is frequently justified.

Literature on experimental designs provides a set of designs for the implementation of fractional factorial designs with several factors and factor levels (Cochran & Cox, 1950; Kirk, 1995). Note that fractional designs with the same number of factor levels are less complex than designs with different numbers of factor levels. For this reason, vignette designs using an experimental vignette selection should not be too large and complex. They should only include the most important factors with preferably the same number of factor levels. However, D-efficient quota designs might be a viable alternative for more complex designs, though they are not perfectly orthogonal and less efficient than fractional factorial designs (Auspurg et al., 2009; Dülmer, 2007).

Randomized Vignette Selection

Within factorial surveys, vignette subpopulations are selected by a random procedure rather than according to an experimental design (Jasso, 2006; Rossi & Anderson, 1982). In comparison to experimental designs, a deliberately planned confounding of effects is not possible. Random selection of vignettes usually results in a complex random confounding structure: main effects may be completely confounded with lower and higher order interaction effects, and even with other main effects. Which effects are confounded with each other is solely determined by the random selection process.

For our illustrative example Table 3 shows a set of four vignettes which could have been obtained by random selection. In this example, factor *B* only varies with factor *C*, that is, factor level b_0 is measured only together with c_0, and b_1 only with c_1. As a consequence, main effects *B* and *C* are confounded with each other and cannot be separately estimated. Also interaction effect *BC* cannot be determined. Moreover, it can be shown that *AB* is confounded with *AC* and *A* is confounded with *ABC*. In this example, only main effect *A* can be meaningfully estimated, given that the three-way interaction effect *ABC* is negligibly small. Compared to an experimental vignette selection, the loss of information due to a random confounding of effects is obvious.

Though the example presented shows a rather extreme selection, a meaningful interpretation of main effects is only possible if (i) the confounding structure is carefully investigated and (ii) confounded effects can reasonably be assumed to be zero or negligible. Interaction effects should be estimated only if they are unconfounded. Although theory tells us that a random selection of vignettes generally results in approximately orthogonal designs with approximately unconfounded effects (Rossi & Anderson, 1982), there is no guarantee that no severe departure from orthogonality results. Moreover, beside the complete confounding of effects (as described above) also a partial confounding typically occurs. This means that not all factor combinations are equally often represented within the sampled vignette population which results in less efficient and underpowered estimates and – thereby increasing sample size needs. For random selection strategies it is always advisable to check the confounding structure prior to the data collection process. If an unpleasant confound of interesting main and interaction effects is detected, the random selection

Table 2. Fractional factorial $2 \times 2 \times 2$ vignette design consisting of four vignettes

	$C = c_0$		$C = c_1$	
	$B = b_0$	$B = b_1$	$B = b_0$	$B = b_1$
$A = a_0$	$a_0b_0c_0$			$a_0b_1c_1$
$A = a_1$		$a_1b_1c_0$	$a_1b_0c_1$	

Table 3. Four randomly selected vignettes of a $2 \times 2 \times 2$ vignette design

	$C = c_0$		$C = c_1$	
	$B = b_0$	$B = b_1$	$B = b_0$	$B = b_1$
$A = a_0$	$a_0b_0c_0$			$a_0b_1c_1$
$A = a_1$	$a_1b_0c_0$			$a_1b_1c_1$

procedure should be repeated until an acceptable confounding structure results. However, the main advantage of a randomized vignette selection procedure is that it is rather simple and also works for large designs with different numbers of factor levels.

In factorial surveys, a different random selection process as described above is most often used. Instead of randomly selecting a single subpopulation for all respondents (within-subjects design), a separate subpopulation is drawn for each respondent. These respondent-specific sets are usually independently selected for each respondent, that is, with replacement of vignettes after completion of each set.[1] Hence, each respondent receives a different set but with some or all vignettes contained in several other sets (i.e., a kind of mixed design which converges to a between-subjects designs as the vignette population increases). The rational behind this procedure is that each respondent receives its own representative vignette set, and that a large part of the total vignette population will be exhausted by the selected sets (Rossi, 1979; Rossi & Anderson, 1982). However, for large vignette populations the confounding problem remains the same, though more effects can be estimated, but between respondents only.

Partitioning the Vignette Population into Vignette Sets

Partitioning the total vignette population into disjoint vignette sets is a useful alternative whenever a confounding of interaction effects with main effects should be avoided. Depending on the number of factors, factor levels, and the maximum number of vignettes per respondent, the total population of m vignettes is systematically partitioned into s equally sized sets with $m_s = m/s$ vignettes each such that groups of respondents get different vignette sets for judgment (mixed design). This procedure guarantees that the total vignette population is exhausted. If the total population is too large to be partitioned into a small number of sets only a fraction of the total vignette population could be partitioned, but this should be avoided with regard to the restrictions discussed for fractional factorial designs above. As with the selection of vignette subpopulations, the construction of vignette sets can be done according to an experimental design or by a random partitioning procedure.

Experimental Partitioning: Confounded Factorial Designs

The total vignette population can be partitioned into equally sized vignette sets using a confounded factorial design (Cochran & Cox, 1950; Kirk, 1995). As with fractional factorial designs the confounding structure can be deliberately planned. A proper confounding plan ensures that only s-1 higher order interaction effects are confounded, but now with the *set effect* instead of the main effects. The set effect reflects potential differences in judgments from set to set and is usually of no interest, such that its confounding with interaction effects is less problematic than the corresponding confounding with main effects in fractional factorial designs. Consequently, all main and lower order interaction effects can be estimated without any confounding. Only higher order interaction effects of minor interest, which can be assumed to be zero or negligible, are deliberately confounded with the set effect.

Table 4. Confounded factorial $2 \times 2 \times 2$ vignette design with two vignette sets

	$C = c_0$		$C = c_1$	
	$B = b_0$	$B = b_1$	$B = b_0$	$B = b_1$
Set I				
$A = a_0$	$a_0b_0c_0$			$a_0b_1c_1$
$A = a_1$		$a_1b_1c_0$	$a_1b_0c_1$	
Set II				
$A = a_0$		$a_0b_1c_0$	$a_0b_0c_1$	
$A = a_1$	$a_1b_0c_0$			$a_1b_1c_1$

Consider Table 4 where all eight vignettes of the total population have been systematically partitioned into two sets of four vignettes each. In this setting, the three-way interaction effect ABC is confounded with the set effect since the measurement difference between set I and set II $(a_0b_0c_0 + a_1b_1c_0 + a_1b_0c_1 + a_0b_1c_1) - (a_1b_0c_0 + a_0b_1c_0 + a_0b_0c_1 + a_1b_1c_1)$ is identical to the interaction effect ABC of a full factorial design. Given that the estimation of the three-way interaction effect ABC is not of interest, a confounding with the set effect is not problematic. However, the main effects and the two-way interaction effects remain estimable across vignette sets. Elaborated confounding plans can be found in the literature (Cochran & Cox, 1950; Kirk, 1995). As with fractional designs, the number of factors should not be small, and designs with different numbers of factor levels are rather complex.

Randomized Partitioning

The total vignette population can also be partitioned by randomly selecting vignette sets (preferably of equal size) without replacement. However, this procedure is rarely used in factorial vignette surveys. As with random selections of subpopulations random partitions of the vignette population lead to an uncontrollable confounding of set, main, and interaction effects. For a meaningful interpretation of main effects, it is necessary to check the actual confounding structure and to assume that some of the confounded effects are negligibly small. An unpleasant confounding, particularly of main effects with other main effects or with important interaction effects, can be avoided by repeating the random partitioning process until an acceptable confounding structure results.

[1] We classified this selection strategy as a selection of a subpopulation rather than a partitioning of the vignette population because there is no guarantee that the respondent-specific sets do not overlap and that the whole vignette population is completely exhausted.

Compared to confounded factorial designs, the main advantage of random partitioning is its lower complexity in constructing vignette sets. However, restrictions in interpreting main and particularly interaction effects can be quite strong.

Analysis of Vignette Studies

The analysis of fractional or confounded designs has three characteristics which are not common to simple regression analysis: (i) The multiple measurement of respondents (block structure in ANOVA terminology), (ii) the set effect based on the construction of vignette sets, and (iii) the unbalancedness of vignette data, especially concerning unmeasured factor level combinations.

Since each respondent judges a set of vignettes we obtain *multiple measurements* for each respondent. Therefore, vignette data typically show a *two-level structure* with the first level representing the vignette level and the second level the respondent level. Research questions usually focus on both levels, that is, on vignette factors influencing respondents' judgments and on respondent characteristics explaining judgment differences between groups of respondents. Hence, the statistical analysis should take both levels simultaneously into account. Two methods for estimating effects can be used. First, the analysis of (co)variance (ANOVA, ANCOVA) according to the specific experimental design (e.g., randomized block confounded factorial design; Kirk, 1995). Second, a multilevel analysis – also called hierarchical linear model or random coefficient model (Hox, Kreft, & Hermke, 1991; Raudenbush & Bryk, 2002). In comparison to ANCOVA, multilevel analysis is more flexible – not only random intercepts but also random slopes can be easily modeled. In addition, dealing with missing and unbalanced data is facilitated in multilevel analysis.

In a confounded factorial design not all respondents judge the same vignette set, only groups of them receive the same vignette set. Consequently, respondent-specific differences in judgments can be also caused by the difference in their vignette sets. The inclusion of the *set effect* as indicator variables at level two controls for these differences in analyzing vignette data. Note that a set effect can only be separated from differences in respondent judgments if multiple measurements are available for each vignette set. Since factorial surveys typically sample individual vignette sets for each respondent, a potential set effect is intermixed with respondent differences. In contrast, confounded factorial designs take the set effect into account since each vignette set is presented to a group of respondents. The set effect is deliberately confounded with higher order interaction effects such that main effects as well as lower order interaction effects remain estimable without any confounding.

Vignette data are *balanced* if each vignette is equally often measured. Vignettes unmeasured due to a fractional design result in a deliberately planned confounding of effects but not in unbalanced data. Unbalanced vignette data occur with randomly selected vignette sets for each respondent since they do not lead to a uniform distribution of vignettes. For large vignette populations some vignettes may even remain unmeasured. Unbalancedness also occurs if implausible vignettes are excluded from the vignette population or if some respondents refuse to judge single vignettes or the whole vignette set. In any case, unbalancedness results in a complete and partial confounding of estimated main and interaction effects. Depending on the extent and structure of the confounding, estimated effects should be cautiously interpreted, if necessary one should even abstain from their interpretation. Moreover, random selection procedures and the deletion of implausible vignettes may lead to disconnected vignette data such that only subgroups of data can be analyzed (for a detailed discussion on the connectedness of data, see Searle, 1987, p. 139ff). Note, ANCOVA and hierarchical linear models result in identical parameter estimates if vignette data are completely balanced. To summarize, a carefully planned experimental vignette design with the same number of measurements for each vignette facilitates the statistical analysis and a meaningful interpretation of effects in any case.

Practical Implementation of an Experimental Vignette Design

In the following we describe the practical implementation and analysis of a vignette experiment using the already introduced pilot study on attitudes toward immigration. The vignettes on attitudes toward immigrants were constructed as personal descriptions of applicants for Austrian citizenship, based on the five factors described above: *A*: country of origin (3 factor levels), *B*: marital status (3), *C*: official accusation (3), *D*: language proficiency (2), and *E*: profession (2). Hence, the total vignette population consists of $3 \times 3 \times 2 \times 2 \times 3 = 108$ different vignettes. Since 108 vignettes are too many to be judged by each respondent, we decided to partition the vignette population according to a confounded factorial design resulting in a confounding of higher order effects with the set effect (Kirk, 1995, p. 587ff). Substantive considerations and research interests did neither allow the use of less factors nor a simple design with the same small number of factor levels for each factor. Our design consists of three factors *A*, *B*, and *C* (country of origin, marital status, official accusation) with three factor levels each and two factors *D* and *E* (language proficiency, profession) with two factor levels each. This causes additional demands concerning the design and analysis in comparison to a design with the same number of factor levels for each factor (Kirk, 1995, p. 640ff).

Determining the Number of Vignette Sets and Vignettes per Respondent

In a first step, one has to decide into how many equally sized sets the 108 vignettes should be partitioned. If we define a minimum size of six vignettes per respondent, the following set sizes are possible: 18 sets with 6 vignettes each (18 × 6), 12 × 9, 9 × 12, 6 × 18, 4 × 27, 3 × 36, or 2 × 54. We considered the sets with 54 and 36 vignettes as too large to be judged by respondents – mainly due to fatigue effects. On the other hand, six vignettes per respondent are too few, because this would have led to a complete confounding of five-way and several four-way interaction effects, and to a partial confounding of two-way interaction effects with the set effect. A partial confounding is unavoidable, if the set size is not an integral multiple of the possible factor level combinations. For 18 sets with 6 vignettes each, a partial confounding of the two-way interaction effects *AB*, *AC*, *BC*, and *DE* arises because all of the 3 × 3 and 2 × 2 factor level combinations are not completely covered by each single set. However, main effects can be estimated without any confounding because 6 is an integral multiple of 3 and 2 (the number of factor levels for *A*, *B*, *C*, and *D*, *E*, respectively). This means that each factor level is equally often measured in each vignette set – 2 or 3 times, respectively. In order to avoid a too strong partial and complete confounding, the number of vignettes per set should be as high as possible. If we do not accept a partial confounding of two-way interaction effects with the set effect, it would be necessary to use at least 36 vignettes per set (36 is an integral multiple of all two-way factor combinations: 3 × 3, 3 × 2, 2 × 2). However, we considered 36 vignettes per respondent as too many. We also rejected the design with 27 vignettes per set because main effects *D* and *E* (27 is no integral multiple of 2) and the interaction effects among the 3 × 2 and 2 × 2 factor level combinations would have been partially confounded. Finally, we decided to choose 6 sets with 18 vignettes each because all main effects and two-way interaction effects, with the exception of *DE* (2 × 2), remain completely unconfounded. While the three-way interaction effects *ABC*, *ADE*, *BDE*, and *CDE* are partially confounded with the set effect, the other three-way interaction effects can be estimated without any confounding since 18 vignettes are able to cover each of the 3 × 3 × 2 factor level combinations exactly once (integral multiple). For all four-way interaction effects a partial confounding results. Finally, the five-way interaction effect *AB-CDE* is completely confounded with the set effect. For our substantive research question this design was sufficient, since we were primarily interested in the main and two-way interaction effects. We did not expect any significant effects for the higher order interaction effects.

Construction of Vignette Sets

Once the size of vignette sets is determined, vignettes are systematically assigned to sets. This has to be done in a way such that the factor level combinations of the unconfounded main and interaction effects occur equally often in each vignette set. For the partially confounded interaction effects, the factor level combinations should be distributed as balanced as possible. Since designs with different factor levels are rarely used, it is difficult to find corresponding design plans in the literature. Therefore, we started with an initial systematic assignment and then slightly varied it until a satisfying solution with regard to the confounding structure was achieved. The confounding structure can be investigated by cross-tabulating the factors for each set separately. Finally, we obtained a design where the five-way and some four-way interactions (*ABCDE*, *ABCE*) were completely confounded with the set effect, while the interaction effects *BC*, *CE*, *DE*, *ABC*, *ACE*, *BCD*, *BCE*, *ADE*, *BDE*, and *CDE* and all four-way interaction effect were only partially confounded – resulting in less efficient estimates. This design was satisfying because we were particularly interested in the two-way interaction effects between the country of origin (*A*) and the other factors, which are in no way confounded. Note that the construction of respondent-specific vignette sets by the much simpler procedure of randomly drawing vignette sets without replacement would have resulted in a random confounding of set, main, and interaction effects. Even the most interesting main effects would have been partially and probably also completely confounded with the set effect and interaction effects. Hence, their substantive interpretation would have been restricted due to the necessary assumption of inexistent set and interaction effects (complete confounding) and the less powered significance tests (partial confounding).

Respondent Sample and Data Collection

In June 2003 we contacted 36 respondents in Vienna (18 students and 18 old aged pensioners). In total, 36 × 18 = 648 vignettes were presented to respondents such that 6 measurements were obtained for each vignette. The balanced respondent sample enabled an efficient estimation of differences between the two respondent groups. Within each respondent group, vignette sets were randomly allocated to respondents, and the vignette order was randomized by shuffling the vignette cards. No data were missing for the first scale (self-assessment, that is, respondents had to answer, whether the presented applicant should become an Austrian citizen from their *own* perspective). For the second scale (i.e., respondents had to answer the same question from the perspective of the other group – students or pensioners) one complete nonresponse of one respondent occurred. This case was omitted in analyzing the second scale only.

Analysis and Results

We analyzed our vignette data by using a multilevel model allowing only for random intercepts. We did not include random slopes because of the small sample size of respondents. However, for large sample sizes random slopes and cross-level interactions, particularly with the set effect, can be taken into account. Separate analyses of student and pensioner judgments were undertaken for the first and second scale. Table 5 shows the corresponding parameter estimates.

Since the parameter estimates are based on non-orthogonal dummy coding their confounding structure differs from the one of orthogonal effects as discussed above. All esti-

Table 5. Multilevel estimates for the first and second scale of old aged pensioners and students (random-intercept model)

	First scale (own perspective)		Second scale (others perspective)	
	Pensioners	Students	Pensioners	Students
	Coef. (s.e.)	Coef. (s.e.)	Coef. (s.e.)	Coef. (s.e.)
Constant	3.99 (.63)**	5.05 (.65)**	3.35 (.45)**	4.84 (.65)**
Set				
Set 1 (ref.)				
Set 2	1.37 (.85)	.21 (.86)	−.31 (.60)	−.04 (.88)
Set 3	1.48 (.85)	.83 (.87)	.39 (.60)	.59 (.88)
Set 4	1.67 (.85)	.59 (.86)	.37 (.60)	.74 (.88)
Set 5	1.28 (.85)	−.64 (.86)	.13 (.60)	−.56 (.88)
Set 6	1.07 (.85)	−.62 (.87)	.19 (.60)	−.07 (.99)
Country				
Nigeria (ref.)				
Iran	.15 (.18)	−.35 (.19)	.17 (.13)	.18 (.15)
Hungary	.37 (.18)*	−.13 (.19)	.24 (.13)	.44 (.15)**
Marital status				
Austrian wife (ref.)				
Foreign wife	−.58 (.18)**	−.23 (.24)	−.73 (.13)**	−.23 (.15)
Single, not married	−.45 (.18)**	−.47 (.23)*	−.64 (.13)**	−.11 (.15)
Accusation				
No complaints (ref.)				
Failure of registration	−1.08 (.18)**	.02 (.24)	−.30 (.13)*	−.36 (.15)*
Slight bodily assault	−2.15 (.18)**	−1.30 (.23)**	−.80 (.13)**	−1.15 (.15)**
Language				
Good German (ref.)				
Broken German	−.70 (.14)**	−.51 (.11)**	−.60 (.11)**	−.57 (.12)**
Profession				
Unskilled worker (ref.)				
Employee	.25 (.14)	−.13 (.19)	.23 (.11)*	.18 (.12)
Country:Profession				
Iran:Employee		.67 (.27)*		
Hungary:Employee		.54 (.27)*		
Marital status:Accusation				
Foreign wife:Registration		−.79 (.36)*		
Single:Registration		−.08 (.34)		
Foreign wife:Bodily assault		.12 (.34)		
Single:Bodily assault		.13 (.34)		

Note. The coefficients (coef.) relate to the 6-point scale from 1 (no agreement) to 6 (full agreement). Standard errors (s.e.) are given in parentheses. Reference categories are indicated by (ref.).
Significance of the coefficients (two-sided *t* tests): $*p < .05$, $**p < .01$.

mated parameters in our model are confounded with four- and five-way interaction effects in a different way.[2] Under the assumption that the confounded four-way and five-way interaction effects are zero, a cautious interpretation of coefficients is possible. With the exception of the self-assessment model for students (first scale: own perspective) no model showed significant interaction effects.

In both respondent groups, official accusation – particularly a slight bodily assault – shows the biggest effect on the first scale, followed by the language proficiency and the marital status. Among the pensioners the degree for granting the Austrian citizenship to an applicant accused for slight bodily assault decreases by 2.15 points on the 6-point scale. Among students this effect is 1.30 points, but only if the

2 The hierarchical models were estimated in the statistical programming language R using the function *lme()* of the *nlme*-package (Pinheiro, Bates, DebRoy, & Sarkar, 2007). The confounding structure of parameter estimates, which depends on the contrast chosen, can be investigated via the *alias()* function (R Development Core Team, 2007).

applicant is married to an Austrian wife (however, interaction effects are very small). Bad language skills reduce the agreement to the Austrian citizenship by .70 and .51 points, respectively. Regarding marital status, applicants married to an Austrian wife have the best prospects for getting the Austrian citizenship. The profession of the applicant shows only an effect among students. The country of origin does not show any effect, either among students or among old aged pensioners. Among students there are two significant interaction effects. For instance, the interaction between the country of origin and the profession shows that the professional qualification has a positive effect only for applicants from Iran and Hungary but not for applicants of Nigeria. However, due to the low power of the pilot study some interaction effects might not have been detected.

For the second scale – the rating from the perspective of the other group – effects are very similar to those of the first scale, though without significant interaction effects. For student judgments rated by pensioners, the effect of the country of origin is more important. Taking both scales together three conclusions can be drawn. First, students are more liberal in their own judgments than pensioners (constant of 5.05 in comparison to 3.99 for pensioners, but no significant difference, $p = .21$). Second, average judgments of students and pensioners differ in the factor "official accusation". Officially accused applicants are much more strongly sanctioned by pensioners. Third, pensioners' thoughts about how students judge applicants (second scale) largely correspond to students' self-assessment (first scale). In contrast, students assume that pensioners judge much more conservatively than they really do. The difference in the two corresponding constants, 3.35 and 3.99, is significant ($p < .0001$).

Conclusions

When the number of factors and factor levels necessary for testing some substantive theory cannot be kept small, the total vignette population is typically too large to be judged by each respondent. Hence, each respondent gets only a subset which is either a random or experimentally planned selection of the vignette population. In survey research, random selection strategies have been much more popular than experimental vignette designs because the lower design complexity of randomization facilitates the practical implementation of a vignette study. But the resulting gain in practicability may cause a severe loss of information due to an uncontrolled confounding of effects. However, a controlled confounding and a clear interpretation of effects are maintained if experimental vignette designs are applied. Such designs should be used whenever an efficient confirmatory analysis of causal research questions is the main research concern. The number of factors and factor levels is usually kept as low as possible in order to avoid complications in tailoring the total vignette population into vignette sets. Experimental vignette designs typically include two to seven factors with only a small and preferably the same number of factor levels for each factor. Vignette populations rarely exceed more than 200 vignettes.

In contrast, factorial surveys with randomly selected vignette sets are most useful for exploring very large vignette populations where the interest is primarily on the estimation of main effects rather than interaction effects which are assumed to be zero or negligibly small. Random selection facilitates the handling of designs with a large number of factors and different numbers of factor levels. Factorial surveys usually include 5–10 factors with 2–10 factor levels. But there are also studies with up to 20 factors and 57 factor levels such that the resulting hypothetical vignette populations comprise trillions of unique vignettes (e.g., Rossi et al., 1985; Shlay et al., 2004). However, the usefulness of such large vignette populations is questionable because the confounding of effects necessarily increases as the size of the vignette population increases and the portion of actually sampled vignettes decreases.

We are convinced that small vignette experiments are more efficient for most practical applications. The implementation according to an experimental design plan with a deliberate and clear-cut confounding structure allows an accurate and efficient estimation of main and also interaction effects (at least two-way interactions) which are very common for individual judgments and beliefs. For small vignette designs with only a few factors and factor levels (with the same number of factor levels for each factor), experimental vignette designs can be easily constructed – corresponding design plans can be found in the experimental design literature (e.g., Kirk, 1995). However, if substantive considerations do not allow a restriction of the vignette design to a few factors and factor levels two study designs still preserve the main advantages of experimental designs: (1) the two-step vignette study design and (2) the multiple vignette study design.

The two-step design decomposes the overall study into an exploratory and a confirmatory step. In the exploratory step, a factorial vignette study with randomly selected vignette sets can be implemented to identify the most relevant factors of the topic under investigation. Alternatively, a fractional factorial design with a deliberate confounding of main effects with higher order interaction effects may be used. In the confirmatory step, the identified relevant factors and factor levels are used to conduct a much smaller experimental vignette study for a more thorough investigation of main and interaction effects.

In a multiple vignette study design a carefully designed series of small vignette experiments is implemented. A multitude of vignette experiments allows the variation of several vignette elements. For instance, the vignette context, single vignette factors, or factor levels can be systematically varied across experiments. With regard to our data the context "application for the Austrian citizenship" could be changed into "application for a job", which introduces a different contextual embedding and thereby increases the generalizability of results. If single factor levels or factors are changed, one could choose other countries of origin or replace a specific factor with another one (e.g., country of origin with religious conviction). Since mode and form of the vignette presentation could influence respondent judgments, the systematic change of the presentation mode – such as video vignettes in comparison to text vignettes – can also be taken into account. We think that a well-planned multiple vignette design with a series of internally valid experiments is able to cover a broad range of relevant factors and thereby increases external validity as well as the generalizability of context-specific results.

Finally, the combination of the vignette technique with a representative sampling of respondents and a traditional questionnaire allows the investigation of conditional vignette effects (conditional on respondent characteristics) and the generalization to a well-defined target population. In many fields of application, traditional survey research could substantially benefit from the flexibility and methodological strength of vignette studies.

Acknowledgments

This work was partially supported by the Jubilee Fund Grant 12320 from the Austrian National Bank (OeNB).

References

Alexander, C. S., & Becker, H. J. (1978). The use of vignettes in survey research. *Public Opinion Quarterly, 42*, 93–104.

Alves, W. M., & Rossi, P. H. (1978). Who should get what? Fairness judgments of the distribution of earnings. *American Journal of Sociology, 84*, 541–564.

Atzmüller, C. (2006). *Die Vignettenanalyse als empirisches Messmodell: Methodische und methodologische Neuerungen gezeigt am Beispiel der Untersuchung normativer (Vor-) Urteile unter Jugendlichen zum Thema Integration von MigrantInnen* [The vignette analysis as empirical model: Methodological improvements shown as an example for a normative prejudice/judgement survey of adolecents with the topic "integration of immigrants"]. University of Vienna, Vienna, Unpublished doctorial dissertation.

Auspurg, K., Hinz, T., & Liebig, S. (2009). Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey [Complexity, learning effects and plausibility of vignettes in the factorial survey design]. *Methoden – Daten – Analysen, 1*, 59–96.

Barrera, D., & Buskens, V. (2002). *Imitation and learning under uncertainty: A vignette experiment*. ICS/Department of Sociology, Utrecht University.

Beck, M., & Opp, K. (2001). Der faktorielle Survey und die Messung von Normen [The factorial survey and the measurement of norms]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 53*, 283–306.

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods, 4*, 243–249.

Cochran, W. G., & Cox, G. M. (1950). *Experimental designs* (2nd ed.). New York: Wiley.

Cook, F. L. (1977). *Differences in public support for seven social welfare groups: Description and explanation.* Doctoral dissertation, University of Chicago.

Cook, F. L. (1979). *Who should be helped? Public support for social services*. London: Sage.

Dülmer, H. (2001). Bildung und der Einfluss von Argumenten auf das moralische Urteil. Eine empirische Analyse zur moralischen Entwicklungstheorie Kohlbergs [Education and the influence of arguments for moral judgement: An empirical analysis of Kohlberg's moral development theory]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 53*, 1–27.

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

Hox, J. J., Kreft, I. G. G., & Hermkens, P. L. J. (1991). The analysis of factorial Surveys. *Sociological Methods & Research, 19*, 493–510.

Hunter, C., & McClelland, K. (1991). Honoring accounts for sexual harassment: A factorial survey analysis. *Sex Roles, 24*, 725–752.

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

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

Jones, C., & Aronson, E. (1973). Attribution of fault to a rape victim as a function of respectability of the victim. *Journal of Personality and Social Psychology, 26*, 415–419.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Liebig, S., & Mau, S. (2005). Wann ist ein Steuersystem gerecht? Einstellungen zu allgemeinen Prinzipien der Besteuerung und zur Gerechtigkeit der eigenen Steuerlast [When is a system of taxation fair? Attitudes towards general principles of taxation and fairness of the own tax burden]. *Duisburger Beiträge zur Soziologischen Forschung, 1.*

Ludwick, R., & Zeller, R. A. (2001). The factorial survey: An experimental method to replicate real world problems. *Nursing Research, 50*, 129–133.

Mäs, M., Mühler, K., & Opp, K. (2005). Wann ist man Deutsch? Empirische Ergebnisse eines faktoriellen Surveys [When are you German? Empirical results of a factorial survey]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 57*, 112–134.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2007). nlme: Linear and nonlinear mixed effects models. *R package version, 3*, 1–83.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution for moral responsibility. *Journal of experimental social psychology, 39*, 653–660.

R Development Core Team. (2007). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved, April 15, 2007, from http://www.R-project.org.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd ed.). London: Sage.

Rossi, P. H. (1979). Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman, & P. H. Rossi (Eds.), *Qualitative and quantitative social research. Papers in honor of Paul F. Larzarsfeld* (pp. 176–186). New York: Free Press.

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

Rossi, P. H., & Nock, S. L. (Eds.). (1982). *Measuring social judgments: The factorial survey approach*. Beverly Hills, CA: Sage.

Rossi, P. H., Sampson, W. A., Bose, C. E., Jasso, G., & Passel, J. (1974). Measuring household social standing. *Social Science Research, 3*, 169–190.

Rossi, P. H., Simpson, J. E., & Miller, J. L. (1985). Beyond crime seriousness: Fitting the punishment to the crime. *Journal of Quantitative Criminology, 1*, 59–90.

Rossi, P. H., Waite, E., Bose, C. E., & Berk, R. E. (1974). The seriousness of crimes: Normative structures and individual differences. *American Sociological Review, 39*, 224–237.

Searle, S. R. (1987). *Linear models for unbalanced data*. New York: Wiley.

Seguin, C. A., & Ambrosio, A. L. (2002). Multicultural vignettes for teacher preparation. *Multicultural Perspectives, 4*, 10–16.

Shlay, A. B., Tran, H., Weinraub, M., & Harmon, M. (2004). *How parents evaluate child care: A factorial survey analysis of perceptions of child care quality, fair market price and willingness to pay by low-income, African American mothers*. Philadelphia: Temple University.

Sniderman, P. M., & Grob, D. B. (1996). Innovations in experimental design in attitude surveys. *Annual Review of Sociology, 22*, 377–399.

Steiner, P. M., & Atzmüller, C. (2006). Experimentelle Vignettendesigns in faktoriellen Surveys [Experimental vignette designs in factorial surveys]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 58*, 117–146.

Thurman, Q. C., Lam, J. A., & Rossi, P. H. (1988). Sorting out the cuckoo's nest: A factorial survey approach to the study of popular conceptions of mental illness. *Sociological Quarterly, 29*, 565–588.

Veal, W. R. (2002). Content specific vignettes as tools for research and teaching. *Electronic Journal of Science Education, 6.*

Walster, E. (1966). Assignment of responsibility for an accident. *Journal of Personality and Social Psychology, 3*, 73–79.

Wason, K. D., Polonsky, M. J., & Hyman, M. R. (2002). Designing vignette studies in marketing. *Australasian Marketing Journal, 10*, 41–58.

Christiane Atzmüller

Competence Centre for Social Work
University of Applied Sciences
FH Campus Wien
Favoritenstraße 226
1100 Vienna
Austria
Tel. +43 1 606 68 77 8308
E-mail christiane.atzmueller@fh-campuswien.ac.at

