methods, data, analyses | Vol. 16(2), 2022, pp. 315-334 DOI: 10.12758/mda.2021.13

New Methodical Findings on D-Efficient Factorial Survey Designs: Impacts of Design Resolution on Aliasing and Sample Size

Julia Kleinewiese

Mannheim Centre for European Social Research (MZES), University of Mannheim

Abstract

In empirical surveys, finding a sufficient number of respondents can be challenging. For factorial survey experiments, drawing a vignette-sample ("fraction") from a vignette-universe can reduce the minimum number of respondents required. Vignette-samples can be drawn by applying D-efficient designs. Theoretically, D-efficient resolution V designs are ideal. Due to reasons of practicability, however, resolution IV designs have usually been applied in empirical social research and are considered to be sufficient when it is clear up front, which two-way interactions are likely to have an effect. Against this backdrop, this article focusses on two research questions: (1) In resolution IV designs, are those two-way interactions that are not orthogonalized truly not aliased with any main effects? (2) How does design resolution affect the minimum size of the vignette-sample that is necessary for achieving an adequate level of D-efficiency? These questions are examined by applying SAS-macros for computing D-efficient samples, pre-construction assessment and post-construction evaluation. The resulting aliasing structures indicate a discrepancy between previous definitions of design resolutions and the aliasing structures of designs resulting from the SAS-macros. Additionally, they suggest taking a second look at the assumption that higher resolutions or larger vignette universes will always necessitate designs with larger vignette-samples (and thus larger sets or more respondents).

Keywords: D-efficiency, design resolution, sample size, factorial survey, aliasing, confounding

When collecting quantitative data, a major issue is finding a sufficient number of respondents (cf. Engel & Schmidt, 2019). Factorial surveys offer some unique opportunities based on design, such as reducing the number of required respondents, but also challenges that need to be considered carefully – from the design-stage onwards. Factorial survey vignettes are an established method in quantitative social science research measuring attitudes or behaviour. Methodical research implies that vignettes can have a high external validity, i.e. are suitable for measuring real-life attitudes and behaviour (Hainmueller et al., 2015) but may sometimes run into issues, such as social desirability bias (Eifler, 2007; Eifler et al., 2014). Nevertheless, they are especially important for research that cannot be conducted in real-life, due to practical or ethical considerations (such as research on crime and violence; e.g. Verneuer, 2020).

This article aims to contribute towards the growing methodical literature on factorial survey designs in a way that makes it easier for researchers without extensive expertise in this area to clearly understand, implement and reflect on design decisions and their consequences for analyses. Because specific (e.g. D-efficient) designs are becoming increasingly popular due to allowing for the practical implementation of vignettes with a large number of dimensions (and/or levels), i.e. a large overall vignette-universe, it is important that clear design categories (for important features such as orthogonality) are offered – to prevent avoidable mistakes during the design-stage.

The findings in this article are new in the sense that they are not intuitively drawn from previous literature, although experts sometimes take them into account automatically. This article exemplifies "new" important aspects that foreground both the "benefits" and "dangers" of D-efficient designs to help researchers to (I.) optimize their designs (e.g. reduce sample sizes, avoid misspecifications) as well as (II.) reflect on their designs and possible implications for their analyses as well as interpretation of results.

In factorial surveys, it is common to divide the vignettes into sets and present each respondent with such a set, thus, gaining more than one response per person for the dependent variable(s). Hence, a smaller number of respondents is required (cf. Atzmüller & Steiner, 2010). In combination with this or on its own, a smaller vignette-sample, a "fraction", is sometimes drawn from the overall universe. This can be extremely useful for reducing the number of respondents that are required but it also implicates new aspects that must be addressed in depth to ensure that e.g. the internal validity of the experiment is upheld and estimates in the analysis are not biased as a result of the design (Auspurg & Hinz, 2015). To date, random proce-

Direct correspondence to
Julia Kleinewiese, Mannheim Centre for European Social Research (MZES),
University of Mannheim, 68131 Mannheim, Germany.
E-mail: kleinewiese@uni-mannheim.de

dures have been the 'go to' method for allocating vignettes to sets and for drawing vignette-samples.

Nevertheless, drawing random vignette-samples comes with some drawbacks, which can be especially meaningful when the sample size is relatively small (see e.g. Auspurg & Hinz, 2015); with random designs, we have no means of controlling two of the important properties of the experimental design: Orthogonality and level balance. To tackle this issue, some researchers have been examining and applying quota designs, including D-efficient designs. Such methods have been widely investigated and applied in a related type of survey experiment, in discrete choice experiments (DCEs), which are applied mostly by economists (e.g. Louviere et al., 2000; see e.g. Gundlach et al. [2018] for a sociological DCE). Unlike the case of DCEs, for factorial surveys, the research on this method of selecting, for example, a sample of vignettes from the universe is very limited (but see e.g. Auspurg & Hinz, 2015; Dülmer, 2015, 2007; Kuhfeld, 2003). Much of this research has focussed on comparing D-efficient designs with random sample designs. Such research is highly valuable for illuminating what the advantages and drawbacks of each procedure are. This article seeks to take this research as a point of departure from which to present accessible methodological information on D-efficient designs.

Currently, there is only a small amount of research that focuses specifically on D-efficiency in factorial surveys. The examination of the method as well as its procedures has focused primarily on comparing different features (including D-efficiency itself) of D-efficient and random vignette-samples (e.g. Auspurg & Hinz, 2015; Dülmer, 2007). The current article builds on such previous research. The two research aims are specifically oriented towards the concept of design resolution: (1) To discuss discrepancies between conceptualizations of resolution IV designs and their implementability with SAS[1]; (2) to examine how the resolution – III, IV (with 1, 3 or 5 two-way interactions orthogonalized) and V – affects the minimum size of the vignette-sample that is necessary to still achieve an adequate level of D-efficiency (over 90).

It is for the aforementioned reasons that the current focus is placed on design resolution of D-efficient factorial surveys. The theory section gives a general overview of the research field on factorial survey D-efficiency while the application section exemplifies issues regarding design resolutions of D-efficient vignette-samples and their implications for confounding as well as sample size (e.g. Kuhfeld, 2003). The steps are as follows: First, a brief overview of factorial survey designs is presented, introducing factorial surveys and describing different methods for drawing samples of vignettes as well as the concomitant (dis-)advantages. This is followed,

1 The SAS-version that I am referring to in this article is "SAS OnDemand for Academics – SAS Studio" which is a cloud/online version. Since it is free for all academics it is used very frequently. I refer to it as "SAS" but the assessments and comments made may not hold true for other versions of the software.

by a "state of the art" section on D-efficiency in factorial surveys that includes the theoretical premises that are of importance for the subsequent section: The application (using SAS-macros). This section focuses on the discrepancies between conceptualization and SAS-implementation regarding confounding (aliasing) structures of "resolution IV" designs and on how design resolution impacts the size of the smallest possible vignette-sample that can be constructed from a given full factorial. This section is succeeded by a discussion of the results and a brief conclusion that presents the general implications and recommendations for future research on and with D-efficient factorial survey designs.

Factorial Survey Designs

Factorial Survey Methodology

A factorial survey systematically varies dimensions in scenarios and presents the resulting vignettes to respondents (e.g. Auspurg et al., 2015; Wallander, 2009; Steiner & Atzmüller, 2006; Beck & Opp, 2001; Alexander & Becker, 1978). A parallel between factorial surveys and experiments lies in the condition that the researcher controls the "treatments" (dimensions), so that they can be measured independent of each other (cf. Auspurg et al., 2009; Rossi & Anderson, 1982). By means of varying the dimensions' levels, the factorial survey allows direct deductions concerning the dependent variable's variations, as the effects of unobserved variables are eliminated (Dickel & Graeff, 2016). All levels of a dimension need to be clearly distinct from each other. The vignette universe (vignette population/full factorial) is made up of all the vignettes resulting from each possible combination of the dimensions' levels (Auspurg & Hinz, 2015; Atzmüller & Steiner, 2010; Rossi & Anderson, 1982). In order to avoid dimensions that are composites of a number of attributes, high numbers of dimensions must be selected for some factorial surveys (see Hainmueller et al., 2014). Furthermore, some studies need a high number of levels for one or more dimensions (e.g. due to content-related or analyses requirements). A large number of dimensions and/or levels quickly leads to a very large vignette universe.

The dependent variable is frequently measured on a scale, as a rating score in response to a question (Dickel & Graeff, 2016) regarding the vignette. Frequently, 11-point scales are used (Dülmer, 2014; Wallander, 2009). Usually, additional respondent-specific data are collected and can be included in the analysis of the vignette evaluations (Steiner & Atzmüller, 2006). The aim of statistical analyses is determining the effect of each dimension (and often some interactions) in regard to the respondents' judgments as well as identifying and explaining the differences

between respondents or groups of respondents (Auspurg & Hinz, 2015; Auspurg et al., 2015; Steiner & Atzmüller, 2006; Beck & Opp, 2001).

If sufficient numbers of respondents are available, every vignette from a vignette universe should be judged by at least five respondents (because if e.g. only one person rates a vignette it is completely confounded with their personal features) (Auspurg & Hinz, 2015). However, with a rising number of vignettes, it becomes increasingly difficult to recruit the necessary number of respondents. There are two solutions which have been prioritized in factorial survey applications, separately or in combination: (1) dividing the overall number of vignettes into sets of equal size or (2) selecting only a sample of the vignettes from the universe (cf. Steiner & Atzmüller, 2006).

Forming sets (decks/blocks) with a specific number of vignettes has the advantage that one can greatly reduce the number of respondents required. With this proceeding, each respondent only answers one set of vignettes. Vignettes can be assigned to sets through experimental variation or random allocation (with or without replacement) (cf. Steiner & Atzmüller, 2006). For optimal distribution, the vignette universe should be a whole multiple of the set size (number of vignettes per set) (Atzmüller & Steiner, 2017; Auspurg & Hinz, 2015). As respondents presumably differ in their assessment tendencies, the measurements are not independent across all vignette-responses. The equivalent variance component is incorporated in the statistical analysis through the modelling of a set effect. Auspurg and Hinz (2015) state: "[…] some parameters become confounded with deck effects [… but] When all decks are rated by several respondents […] these parameters remain identifiable in estimations across respondents" (p. 39).

There are several methods for selecting a sample of vignettes from the universe. Steiner and Atzmüller (2006) argue that in the case of randomly drawn (sets of) vignettes, a very complex interaction structure is formed, which may lead to considerable interpretation problems in regard to the estimable effects; they declare that the common implicit assumption that the interaction effects mixed in the effects of interest are equal to zero, is generally an unsatisfactory solution.

This brief introduction to the current state of factorial survey research – particularly its design – provides a basis for understanding the particular methodical design-aspects that are of relevance to the goals of this article. In order to provide insights into other design options and what (not) to do, a number of aspects regarding the design of vignette studies will, subsequently, be described in more detail. All of this shows why the current aims are relevant and provides sufficient knowledge for comprehension of the applied section. The following section gives a brief overview of the different proceedings for drawing a sample of vignettes from the overall vignette universe.

Methods for Drawing Vignette-Samples

There are two important properties of the experimental design regarding the vignette universe as well as vignette-samples: The first property is *orthogonality.* A matrix is orthogonal when the single columns are not correlated with one another. This enables independent (from each other) estimation of the effects of the factors (cf. Auspurg & Hinz, 2015). Thus, for a factorial survey design, orthogonality means that the dimensions (and their interactions) do not correlate with each other (Auspurg & Hinz, 2015; Atzmüller & Steiner, 2010; Taylor, 2006; Rossi & Anderson, 1982); "[…] it enables the researcher to estimate the influence of single dimensions independently of each other" (Auspurg & Hinz, 2015, p. 25). A vignette universe (full factorial) is always orthogonal. The second property is *level balance*. Level balance means that all levels (of every dimension) occur with equal frequencies. Level balance indicates that maximum variance (of the levels) can be used to estimate the effect of each dimension, which leads to the lowest standard errors and, therefore, maximizes the precision of the parameter estimates (cf. Auspurg & Hinz, 2015).

The Cartesian product of dimensions and levels equals the size of the vignette universe. If each respondent judges all vignettes from a universe, the factors are orthogonal to one another in their composition (Dülmer, 2007). In factorial surveys, each person usually only responds to a selected number of vignettes from the universe. This can be achieved through dividing the universe into vignette-sets of equal size ("blocking") and presenting each respondent with one set only, otherwise, by selecting a sample of vignettes from the universe (or both of the aforementioned).

This section focusses on drawing samples from a vignette universe – presenting methods for drawing such vignette-samples. There are two categories into which techniques for attaining samples fall: (1) *Random samples* are predominantly used to attain a vignette-sample from the universe (the aim is to represent its possible level combinations as closely as can be achieved), however, (2) *quota designs* can also be applied (Dülmer, 2007).

(1) *Random samples* can be drawn once (in sets) and then judged by several respondents (*clustered random design*) or they can be drawn uniquely for each respondent (s*imple random design with or without replacement).* The former procedure ensures – given a sufficient number of respondents – several ratings of each included vignette (cf. Jasso, 2006). Each of these strategies has its advantages and its disadvantages. Drawing only once and presenting the resulting sets to several respondents is advisable when one is interested in respondent-specific variation in the vignette-judgements. However, a wider overall portion of the vignette universe is very likely to be achieved when a unique deck is drawn for each respondent (Jasso, 2006).

(2) *Quota samples* are commonly used in conjoint analysis and discrete choice experiments (cf. Dülmer, 2007). There are two types that have been applied frequently: *Fractional factorial designs* (e.g. Marshall & Bradlow, 2002) and *D-efficient designs* (e.g. Kuhfeld et al., 1994). In both variants, the vignette-sample is drawn only once (and then usually divided into sets). Quota sampling utilizes the available knowledge on the statistical properties of the universe in order to select the vignette-sample (of a given size) that most closely/ideally upholds these properties (cf. Dülmer, 2007).

A fractional factorial design is a *symmetrical orthogonal design* when the vignette universe properties of equal level frequencies (symmetrical/balanced) and orthogonality of all factors (e.g. dimensions, interactions) are upheld. It is an *asymmetrical orthogonal design* when it does not have absolute level frequency but preserves orthogonality because one dimension's levels occur with proportional frequency to the other dimensions' levels (Dülmer, 2007).

D-efficient designs relax the rule that a (sample-)design must be perfectly orthogonal. Symmetrical orthogonal designs (perfect level balance as well as orthogonality) represent the vignette universe most closely and minimize parameter estimates' variance. D-efficiency chooses symmetrical orthogonal designs as a point of reference and is, thus, "a standard measure of goodness" (Dülmer, 2007, p. 387) for jointly assessing both orthogonality and level balance, which increases the precision of estimates of the parameters in statistical analyses (Auspurg & Hinz, 2015).

Designs that have a D-efficiency of 100 are also (fractional factorial) symmetrical orthogonal designs (Dülmer, 2007) because they are orthogonal and exhibit level balance. When this is not the case, the best "compromise" between the aims of orthogonality and level balance is searched for (D-efficiency will then be lower than 100). When orthogonal coding has been applied to the vignettes, the range of D-efficiency is 0-100 (see e.g. Dülmer, 2007; Kuhfeld, 1997; Kuhfeld et al., 1994).

There are a number of 'pros and cons' regarding the methods that can be used for selecting a subsample from a vignette universe. From the statistical perspective, reasons why quota designs should be favoured over random designs are their higher efficiency, reliability and power. However, these arguments are primarily applicable for studies that use a fairly low set size, where the selected design is highly D-efficient and quite a high unexplained inter-respondent heterogeneity is to be expected. On the other hand, quota designs can be less valid than random designs; this is most likely when using designs with a low resolution (Dülmer, 2007). In consequence, what type of design is the most expedient for a study can vary – depending on, for instance, the respondent sample and the amount of resources available for implementing the survey. In the past, a majority of factorial survey studies used random designs (Wallander, 2009). However, increasingly D-efficient designs are becoming

more popular. The remaining sections of this article, therefore, focus exclusively on D-efficient designs.

D-Efficiency

Taking the preceding overview as a point of departure, this section provides a more in-depth elaboration of D-efficiency. It begins with a general section on D-efficient designs that is followed by subsections on sample size as well as design resolution. This constitutes the final theoretical building block for assessing the implications in the applied section.

D-Efficient Designs

When one applies (D-)efficiency-maximizing methods for finding a suitable vignette-sample, one should be able to reach the same amount of precision as with random sampling but with fewer respondents and/or vignettes per set. Moreover, it can be easier to reach and asses the goal that all parameters of interest can be identified. Against this background, an objective is to find a fraction of the vignette universe with maximal gain of information, about all parameters that are of relevance for the research aim(s).

The previously described combination of considering both orthogonality and level balance can be specified in regard to optimizing D-efficiency: The goal is maximizing the variance of the dimensions' levels while simultaneously minimizing the correlations between the factors (e.g. dimensions, interactions). The equivalent optimums are level balance and orthogonality.

D-efficiency contains (is reliant on) the Fisher Information Matrix (FIM) [X'X], where X indicates a vector (of vignette variables) (Auspurg & Hinz, 2015; Kuhfeld et al., 1994). There are other measures of efficiency (such as A-efficiency; a function of the arithmetic mean of the X'X matrix) than D-efficiency, which is based on the geometric mean of the matrix. However, these efficiency measures are usually highly correlated with each other and D-efficiency is used most frequently (Auspurg & Hinz, 2015; Kuhfeld, 1997). The formula for D-efficiency is as follows [p=parameters to be estimated (including the intercept); n_s=number of vignettes in the fraction; |X'X|=FIM]:

$$\text{D-efficiency} = 100 \times \frac{1}{n_s \times \left|(X'X)^{-1}\right|^{\frac{1}{p}}} = 100 \times \left(\frac{1}{n_s} \times \left|X'X\right|^{\frac{1}{p}}\right)$$

(Auspurg & Hinz, 2015; Dülmer, 2007; Kuhfeld et al., 1994)

Fewer dimensions (or other estimated parameters e.g. 2-way-interactions) reduce the correlation of parameters with each other. Larger vignette-samples, from a vignette universe of a given size, (sample-sizes prescribe the degrees of freedom for parameter estimates) decrease covariation (and, therefore, correlations) between dimensions, increasing precision of parameter estimates. The FIM reflects how high the information is for parameter estimates. The information matrix is the inverse of the variance-covariance matrix.

As stated in subsection *Methods for drawing vignette-samples*, when the dimensions' levels from a universe are in orthogonal coding, the maximum D-efficiency that can be reached is 100. To elaborate upon this: Methodical literature states that a D-efficiency over 90 should be sufficient for experimental survey designs in the social sciences (Auspurg & Hinz, 2015).

The more efficient a design is, the fewer vignette-judgements one requires to achieve the same (level of) statistical power:

> Efficiencies are typically stated in relative terms, as in design A is 80% as efficient as design B. In practical terms this means you will need 25% more (the reciprocal of 80%) design A observations (respondents, choice sets per respondent or a combination of both) to get the same standard errors and significances as with the more efficient design B. *(Chrzan & Orme, 2000, p. 169)*

Sample Size

When the size of a vignette-sample from a given universe increases it becomes more likely that one can reach a high D-efficiency. Auspurg and Hinz (2015) state that this leads to a trade-off because – given a fixed number of respondents and set size – the number of respondents per set decreases. However, an additional option is that one could increase the set size (even if this can increase the design effect; for more information on the design effect see e.g. Auspurg & Hinz, 2015, pp. 50-55).

The smallest possible vignette-sample is the number of parameters that are to be estimated plus one. The smallest sample is normally very inefficient and does not fulfill the criteria of a D-efficiency over 90 (cf. Auspurg & Hinz, 2015).

In Auspurg and Hinz' (2015) comparison of random vignette-samples and D-efficient vignette-samples, using two different vignette universes, the D-efficient designs are always more efficient. The differences are especially high for small vignette-samples and decrease as the sample size increases. The maximum correlations of the random samples are much higher than those of the D-efficient samples, meaning that the random samples' dimensions (experimental factors) loose much of their independency, threatening internal validity. D-efficient samples usually exhibit higher variance (of levels within each dimension), which means higher sta-

tistical power for correctly identifying the effects of the dimensions. Due to hardly any randomness in the selection of the vignette-sample, the variation in the D-efficiency of (same-sized) D-efficient vignette-samples over several "tries" is very low in comparison to the variation exhibited by random samples.

Design Resolution

While small vignette sample size with a D-efficiency of 100 ensures that the dimensions (main effects) are orthogonal to each other and have level balance (in estimation: standard errors are minimized; statistical efficiency is maximized), this is still likely to lead to biased estimates if relevant two-way (or higher) interactions are not negligible. If such interaction effects are not specified in a design, but do have an effect, this leads to confounding of main and interaction effects. This can bias the estimations of the main effects and rules out the estimation of the interaction effects. If main effects are biased, this leads to biased (in some cases entirely false) interpretations of the data (Auspurg, 2018). For this reason, it is important to consider, which effects have been orthogonalized in a D-efficient design. Commonly, this has been approximated by applying the categorization of designs into "resolutions".

> Resolution identifies which effects are estimable. For resolution III designs, all main effects are estimable free of each other, but some of them are confounded with two-factor interactions. For resolution IV designs, all main effects are estimable free of each other and free of all two-factor interactions, but some two-factor interactions are confounded with other two-factor interactions. For resolution V designs, all main effects and two-factor interactions are estimable free of each other. *(Kuhfeld, 2003, p. 237)*

D-efficiency is always measured relatively to the selected design resolution. When the orthogonally coded levels of a dimension (in a given vignette-fraction) are completely identical with those of a 2-/3-/4-way interaction then, statistically, they are entirely correlated and their effects cannot be separated in the analysis i.e. they are completely "confounded" or "aliased". This can also include the intercept. The coefficients of main effects that are aliased with interaction effects are only estimable (unbiased) if those interaction effects have no effect (effect = 0) on the dependent variable. If the wrong assumptions are made this results in biased estimates of the (main) effects (cf. Auspurg & Hinz, 2015).

In marketing research, resolution III designs (also termed "orthogonal arrays") are mostly used (Kuhfeld, 2003). However, in the social sciences, one should always consider possible two-way interactions that might have an effect (e.g. Auspurg, 2018). While, therefore, it seems advisable to use resolution V designs in sociological research, to date, resolution IV designs have usually been applied. This

can be sufficient, however, when using resolution IV designs the researcher should be aware that this might cause biased results if they err in the assumption that the confounded interactions are negligible.

The rules for which level of factors can be estimated independently from one another (or part of the factors from a level; e.g. resolution IV designs) differ, depending on whether or not the number of the resolution (r) is (1) odd – e.g. resolutions III and V – or (2) even – e.g. resolution IV. The general rule for the former (1) is that all effects of order $e = (r - 1)/2$ or below are estimable independently from one another but at least some of the effects of order e are aliased with interactions of order $e + 1$. In the latter case (2) the rule is slightly different: Effects of order $e = (r - 2)/2$ are estimable independently from one another and also from interactions of order $e + 1$ (Kuhfeld, 2005).

Much previous research has been conducted under the assumption that higher resolutions will always necessitate designs with larger vignette-samples (e.g. Kuhfeld, 2003, 2005). Moreover, research has increasingly questioned the primacy given to maximizing the efficiency of designs, arguing that unbiased estimation of effects should be the superior goal (Auspurg, 2018; Czymara & Schmidt-Catran, 2018). Minimizing (possible) bias in estimation of effects requires using higher design resolutions.

Application with SAS-Macros

This section turns towards a practical application, examining examples of implementations of D-efficient factorial survey designs using SAS-macros. D-efficient factorial survey designs in the social sciences are normally constructed by means of computer algorithms. In sociology, the SAS-macros written by Warren F. Kuhfeld (for more details see e.g. Kuhfeld, 2003) are commonly used (see e.g. Auspurg & Hinz, 2015; Dülmer, 2007). These macros enable the computation of D-efficient samples and sets as well as pre-construction assessment and post-construction evaluation. A number of details regarding the design can also be evaluated in varying detail (e.g. correlations, aliasing structure).

Proceedings: The Design and the Macros

I used the SAS-macros %mktruns and %mktex to test my propositions. My first aim was to use the SAS macros to try and construct resolution IV designs that fulfil the conceptual requirements that Kuhfeld (2003, p. 237) defined. I selected a $2^8 = 256$ vignettes universe because I presume that this simple structure is very useful for assessing aliasing structures. I used the macros to construct a fraction with a D-efficiency of 100, thus, 0 violations (of orthogonality and level balance)

and a sample size of n=16 vignettes. I included one two-way interaction-effect to be orthogonalized (x1*x2). I documented the aliasing scheme of the design, in order be able to assess whether or not the properties postulated in literature are present. I then repeated this procedure for two more resolution IV designs – one design with 3 two-way interactions (x1*x2 x2*x3 x3*x4) and one with 5 two-way interactions (x1*x2 x2*x3 x3*x4 x4*x5 x5*x6). Both of these designs also had a D-efficiency of 100, 0 violations and a vignette-sample size of n=16.

For my second research aim, which focusses on the relationship of design resolution and sample size, I selected three vignette universes: $4^4 = 256$, $4^4 2^1 = 512$ and $4^4 2^2 = 1024$. The number of dimensions and their levels for the first universe were selected due to a specific research interest in this structure and the others each add one more two-level dimension, causing each universe to be twice as large as the previous one. This, of course, could have been done differently. I searched for the smallest possible vignette-sample with a D-efficiency as close as possible to 100 for each universe. I constructed five designs for each universe, documenting the sample size, the D-efficiency and the number of violations. The first fraction for each universe was a resolution III design. The second, third and fourth designs were always of resolution IV (according to SAS). Three designs were selected from resolution IV because this resolution can be used to describe the inclusion of various numbers of two-way interactions, from merely one (more than resolution III) to all but one (less than resolution V). The objective was to see if there is a large difference between the minimum sample sizes of resolution IV designs, depending on how many interactions are fixed as orthogonalized in a design. The first of these designs orthogonalizes one two-way interaction (x1*x2), the next design three two-way interactions (x1*x2 x1*x3 x1*x4) followed by a design with five two-way interactions (x1*x2 x1*x3 x1*x4 x2*x3 x2*x4). These three steps were chosen because five is the highest number of two-way interactions possible in the first vignette universe as a resolution IV design (since that is one below 6 two-way interactions, which would be a resolution V design for the first vignette universe). As a final step, a resolution V design was computed and documented for each vignette universe.

Results

Resolution IV Aliasing

Regarding the three resolution IV designs that were supposed to be computed for the first research aim, I find that no main effects are aliased with one another. Furthermore, orthogonalized interactions are not aliased with main effects or other orthogonalized interactions. However, some other two-way interactions are aliased with main effects, orthogonalized interactions and interactions that were not specified to be orthogonal. This, in effect, does not qualify the designs to be of resolu-

tion IV (for this, none of the two-way interactions should be aliased with any main effects) but rather only to be of resolution III. The result is that for this vignette universe, it would have only been possible to use SAS to compute a resolution III or a resolution V design, in accordance with Kuhfeld's definition (2003, p. 237).

Resolutions and Sample Sizes

Table 1 depicts the smallest sample sizes, the D-efficiencies of the samples as well as the violations for each universe when a resolution III design is chosen. For each of the three vignette universes, the sample size is n=16 for the smallest possible size with an adequate D-efficiency (over 90 and as close as possible to 100). The fractions in Table 1 all have a D-efficiency of 100 and, therefore, have 0 violations (of orthogonality and level balance).

Table 2 gives an overview of the sample sizes, the D-efficiencies of the vignette-samples and the violations for the three designs that fall into the category "resolution IV". As shown below, the first and second design-types (1 and 3 two-way interactions) are the same across all vignette universes and in regard to each other. The smallest possible sample size always consists of 64 vignettes, has a D-efficiency of 100 with 0 violations. The final resolution IV design-type (with 5 two-way interactions) has a larger number of vignettes for the smallest sample sizes possible than the first two types. However, it remains the same across the vignette universes. The size is n=128 with a D-efficiency of 96 (with a slight variation in the second decimal place) and 1 violation for each vignette universe.

Table 3 presents the smallest possible vignette-sample sizes, with their D-efficiencies and violations for resolution V designs of each of the three vignette universes. The sample sizes of the first two universes are n=128 with a D-efficiency of 95 (with some variation in the first and second decimal places) and 1 violation. For the last and largest universe, it was not possible to compute a sample of that size with a D-efficiency over 90. The smallest sample size that is possible and fulfils this criterion is n=256. It has a D-efficiency of 100 with 0 violations.

Table 1 Smallest vignette-sample sizes with resolution III design

Universe (from dimensions & levels	Number of dimensions	No. of 2-way interaction-factors	Resolution III		
			Sample size (n)	D-efficiency	Violations
$4^4 = 256$	4	6	16	100	0
$4^4 2^1 = 512$	5	10	16	100	0
$4^4 2^2 = 1024$	6	15	16	100	0

Table 2 Smallest vignette-sample sizes with resolution IV (1/3/5 two-way interactions) designs

Universe (from dimensions & levels	Number of dimensions	No. of 2-way interaction-factors	Resolution IV								
			1 two-way interaction-factor			3 two-way interaction-factors			5 two-way interaction-factors		
			Sample size (n)	D-efficiency	Violations	Sample size (n)	D-efficiency	Violations	Sample size (n)	D-efficiency	Violations
$4^4 = 256$	4	6	64	100	0	64	100	0	128	96.12	1
$4^4 2^1 = 512$	5	10	64	100	0	64	100	0	128	96.13	1
$4^4 2^2 = 1024$	6	15	64	100	0	64	100	0	128	96.16	1

Table 3 Smallest vignette-sample sizes with resolution V design

Universe (from dimensions & levels	Number of dimensions	No. of 2-way interaction-factors	Resolution V		
			Sample size (n)	D-efficiency	Violations
$4^4 = 256$	4	6	128	95.21	1
$4^4 2^1 = 512$	5	10	128	95.68	1
$4^4 2^2 = 1024$	6	15	256	100	0

Discussion of the Results

It is commonly assumed that when designs are of resolution IV, all main effects are estimable independently from one another and from all of the two-way interactions, while two-way interactions may be aliased with each other (e.g. Kuhfeld, 2003). My results offer new insights into the computational issues (SAS) of constructing resolution IV designs. The computed designs concur with the definition in previous literature in that no main effects are aliased with one another and that the orthogonalized two-way interactions are not aliased with the main effects. Also, some two-way interactions that are not orthogonalized are confounded with other two-way interactions. However, a discrepancy between conceptualization and implementation arises: Some non-orthogonalized two-way interactions are aliased with main effects, which may cause estimates of the main effects to be biased. This means that looking at the aliasing structures of the aspiring resolution IV designs shows that they do not fulfil all theoretical requirements and must, instead, be defined as resolution III designs. This suggests that the "catch all" category (resolution IV) between the clearly defined resolutions III and V needs to be treated with caution in implementation. If possible, I suggest that researchers select a resolution V design. If not, aliasing schemes must be carefully monitored (and reported as supplementary material to publications – for reasons of transparency).

Regarding the results on how resolutions impact the smallest possible vignette samples with an adequate D-efficiency, first some general observations: Within each resolution (or subcategory in the case of resolution IV) the smallest sample size is the same for all vignette universes (except for the largest universe in resolution V); even though universes two and three each have one dimension more and are twice as large as the directly preceding (smaller) universe. This is an interesting finding because the same sample size is relatively a smaller fraction when the universe is larger, for example, n=16 is (relatively) a smaller fraction of the vignette universe for design three (1/64) than for the second largest universe (1/32) and the smallest universe (1/16). It is also noteworthy that all sample sizes are whole multiples of each other, of the dimensions as well as their levels and that the universes are whole multiples of the samples. This is due to the structure of the full factorial and may not be so clear cut in the case of, for example, samples of vignettes from universes that are made up of dimensions whose numbers of levels are not multiples of one another. Violations are always equal to 0 when D-efficiency is at 100. This is because that amount of D-efficiency requires perfect orthogonality and level balance. There is a noticeable difference in sample sizes across the resolutions (and subcategories). Resolution III has a 16-vignette D-efficient sample. If one interaction is included (resolution IV, category 1) then the minimum-sample is four times larger (n=64) than in the resolution III designs. For resolution IV with 3 two-way interactions, category 2, the sample size remains at n=64. However, for resolution IV, category 3, with 5 interactions, the sample size (n=128) is twice as

large as for the first two categories in the resolution and eight times as large as in the resolution III designs. Comparing designs of the three categories of resolution IV fractions one can, therefore, claim that there are substantial differences between some (but not all) differing numbers of orthogonalized interactions in regard to the minimum sample size within the resolution. For the third universe with resolution V there is no subsample of vignettes that is smaller than 256 and has a D-efficiency of over 90. A difference in the sample size between the universes is present only for this resolution. The results suggest that a larger vignette universe does not have to increase the smallest possible sample size. Moreover, a higher resolution does not have to increase the smallest possible sample size. Interestingly, the resolutions do not necessarily determine the boundaries at which the minimum sample sizes increase.

Conclusions

The current application provides an added value for vignette-design methodology: As a first step, it examines structural properties of computed (SAS-macros) designs and, for two important issues pertaining design resolutions, compares the results shown in the computed designs with the assertions from the literature. The conclusions provide a basis for future computational (SAS-macros) or mathematically-driven research on design resolution and sample sizes of D-efficient designs. Although the results can lead only to tentative conclusions they should lead to further extensive exploration of this topic.

Some central deductions drawn from the conducted research are: (1) The examined aliasing structures indicate a discrepancy between previous definitions and the aliasing structures of designs resulting from SAS-macros. (2) For the selection of a small sample size, the overall size of the vignette universe does not necessarily play a fundamental role, rather the dimensions' level-combinations. It should be considered from the early stages of design onwards that smallest sample sizes with an adequate D-efficiency can vary strongly depending on the combinations of numbers of dimension-levels that are chosen. When all dimensions have the same number of levels or the level-number of a part of the dimensions is a whole multiple of the other dimensions' level-number smaller vignette-samples can reach an adequate D-efficiency than with more irregular combinations. (3) There is a trade-off between a minimal vignette-sample size and number of orthogonalized factors (not necessarily resolutions). (4) Resolution V designs with an implementable sample size are often possible. Therefore, it is highly recommendable to apply resolution V designs. Sometimes, however, this may not be implementable in research practice, leading researchers to apply resolution IV designs. When implementing resolution IV designs, one should always state precisely, which interactions have been orthog-

onalized. Furthermore, especially when using computer algorithms (e.g. SAS-macros), one must assess the aliasing structures of the design in order to determine if the output design fulfils all of the theoretically presumed orthogonalizations.

Of course, these suggestions are more implementable for some vignette studies than others. Studies, for example, on situational, deviant actions (e.g. Kleinewiese & Graeff, 2020; Wikström et al., 2012) often have more flexibility when it comes to selecting the exact numbers of dimensions' levels. Studies on other topics, such as the gender-pay-gap (e.g. Auspurg, Hinz & Sauer, 2017), may include dimensions (e.g. gender) in which the number of levels is not so easily alterable.

Put in a nutshell, this article clearly shows that D-efficient designs are suitable and expedient for a majority of factorial survey studies – even for researchers without prior "expert knowledge" on experimental survey methodology. It exemplifies, how small changes in design can have large implementation-advantages regarding sample sizes and aliasing. At its core, it reflects upon previous common usage of "resolution IV designs", showing the potential drawbacks of this approach. Based on the conceptual and applied sections, it advises making the usage of resolution V designs a standard in social science research. It supports the necessity of improving transparency regarding research designs. This is important because researchers, reviewers, publishers and readers should have a clear comprehension of the design and its implications for the analyses and the interpretation of the results.

Taking this as a point of departure, future studies should systematically examine the proposed examples (e.g. via comparisons with random samples) to provide further support for the suggested proceedings. Another interesting design-aspect requiring further examination is the interrelation of vignette sampling and blocking. While previous research shows that D-efficient blocking of vignettes to sets leads to less biases in effect estimates than random blocking (Su & Steiner, 2020), as a next step, it would be important to further examine the interrelations of sampling and blocking (both D-efficient and random), especially regarding implementation and possible issues.

References

Alexander, C. S., & Becker, H. J. (1978). The use of vignettes in survey research. *Public Opinion Quarterly, 42*(1), 93-104.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(3), 128-138.

Atzmüller, C., & Steiner P. M. (2017). Was ist ein faktorieller Survey?. In M. W. Schnell, C. Schulz, C. Atzmüller, & C. Dunger (Eds.), *Ärztliche Wertehaltungen gegenüber nichteinwilligungsfähigen Patienten* (pp.29-52). Wiesbaden: Springer.

Auspurg, K. (2018). Konfundierte Ergebnisse durch ein zu stark beschränktes Design?. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *70*(1), 87-92.

Auspurg, K., Hinz, T., & Liebig, S. (2009). Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey. *Methoden – Daten – Analysen, 3*(1), 59-96.

Auspurg, K., & Hinz, T. (2014). *Factorial survey experiments* (Vol. 175). Sage Publications.

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

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

Beck, M., & Opp, K. (2001). Der faktorielle Survey und die Messung von Normen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 53*(2), 283-306.

Chrzan, K., & Orme, B. (2000). An overview and comparison of design strategies for choice-based conjoint analysis. *Sawtooth software research paper series*, *98382*.

Czymara, C. S., & Schmidt-Catran, A. W. (2018). Konfundierungen in Vignettenanalysen mit einzelnen deffizienten Vignettenstichproben. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *70*(1), 93-103.

Dickel, P., & Graeff, P. (2016). Applying factorial surveys for analyzing complex, morally challenging and sensitive topics in entrepreneurship research: The case of entrepreneurial ethics. In E.S.C Berger, & A. Kuckertz (Eds.), *Complexity in entrepreneurship, innovation and technology research* (pp. 199-217). Cham: Springer.

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

Dülmer, H. (2014). Vignetten. In N. Baur, & J. Blasius (Eds.), *Handbuch Methoden der empirischen Sozialforschung* (pp. 721-732). Springer VS, Wiesbaden.

Dülmer, H. (2015). The factorial survey: Design selection and its impact on reliability and internal validity. *Sociological Methods & Research*, *45*(2), 304-347.

Eifler, S. (2007). Evaluating the validity of self-reported deviant behavior using vignette analyses. *Quality & Quantity*, *41*(2), 303-318.

Eifler, S., Pollich, D., & Reinecke, J. (2014). Die Identifikation von sozialer Erwünschtheit bei der Anwendung von Vignetten mit Mischverteilungsmodellen. In S. Eifler, & D. Pollich (Eds.), *Empirische Forschung über Kriminalität* (pp. 217-247). Springer VS, Wiesbaden.

Engel, U., & Schmidt, B. O. (2019). Unit- und Item-Nonresponse. In N. Baur, & J. Blasius (Eds.), *Handbuch Methoden der empirischen Sozialforschung* (pp. 385-404). Springer VS, Wiesbaden.

Gundlach, A., Ehrlinspiel, M., Kirsch, S., Koschker, A., & Sagebiel, J. (2018). Investigating people's preferences for car-free city centers: A discrete choice experiment. *Transportation Research Part D: Transport and Environment*, *63*, 677-688. https://doi.org/10.1016/j.trd.2018.07.004

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis*, *22*(1), 1-30.

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

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

Kleinewiese, J., & Graeff, P. (2020). Ethical decisions between the conflicting priorities of legality and group loyalty: Scrutinizing the "code of silence" among volunteer firefighters with a vignette-based factorial survey. *Deviant Behavior*, *4*(6), 1–14. https://doi.org/10.1080/01639625.2020.1738640

Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research*, *31*(4), 545-557.

Kuhfeld, W. F. (1997). Efficient experimental designs using computerized searches. *Research Paper Series, SAS Institute, Inc*. Cary, NC: SAS Institute Inc.

Kuhfeld, W. F. (2003). *Marketing research methods in SAS: Experimental design, choice, conjoint, and graphical techniques*. Cary, NC: SAS Institute Inc.

Kuhfeld, W. F. (2005). Experimental design, efficiency, coding, and choice designs. *Marketing research methods in SAS: Experimental design, choice, conjoint, and graphical techniques*, 47-97.

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: analysis and applications*. Cambridge: Cambridge university 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.

Steiner, P. M., & Atzmüller, C. (2006). Experimentelle Vignettendesigns in Faktoriellen Surveys. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 58*(1), 117-146.

Su, D., & Steiner, P. M. (2020). An evaluation of experimental designs for constructing vignette sets in factorial surveys. *Sociological Methods & Research*, *49*(2), 455-497

Taylor, B. J. (2006). Factorial surveys: Using vignettes to study professional judgement. *British Journal of Social Work, 36*(7), 1187-1207.

Verneuer, L. M. (2020). Selbstbericht und Vignette als Instrumente zur empirischen Abbildung von Gewalt als Sanktionshandlung. In I. Krumpal, & R. Berger (Eds.), *Devianz und Subkulturen* (pp. 241-277). Springer VS, Wiesbaden.

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

Wikström, P.-O. H., Oberwittler, D., Treiber, K., & Hardie, B. (2012). *Breaking rules: The social and situational dynamics of young people's urban crime*. Oxford University Press.