Script: An Introduction to Factorial Survey Experiments (FSE), Part II

NCRM Online learning resources

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

Hello and welcome to Part II of this introductory course to factorial survey experiments. In the first part, we saw some examples of how factorial surveys are applied in the literature and learned about the main features of this method. In the second part of this course, we will learn about the most important steps in designing and conducting FSEs, and get an idea of how to analyse the data obtained from FSEs.

Before we begin, I would like to draw your attention once again to the reading list in the supplementary material. I mostly rely on the book by Auspurg and Hinz, but you will find other methodological research on some of the issues we will talk about in this course in the supplementary reading list.

By now, you know that FSEs are comprised of hypothetical descriptions of real-world scenarios, which are called the vignettes. For example, a vignette may describe a hypothetical family or a job applicant. The characteristics that comprise the vignettes are called the dimensions. The dimensions are the independent variables in your theoretical model, for example, a person‘s marital status or gender. The levels are the values of these dimensions, for example, married/not married or male/female. The levels of the dimensions are randomly combined by the researcher to construct complete descriptions of hypothetical scenarios (i.e., vignettes). Another important term is the vignette universe, also called the full factorial. The vignette universe is composed of all vignettes that represent all possible combinations of the dimensions’ levels. The number of vignettes in the universe is obtained by the Cartesian product of all levels of all dimensions. For example, let us consider a simple design involving 3 dimensions with 2 levels each. Fully crossing all levels of all dimensions would result in 8 vignettes or, in other words, 8 possible combinations of levels. In this example, we would therefore have 8 different vignettes in the vignette universe. Each vignette represents an experimental condition (i.e., a specific combination of the levels of all dimensions) that would be randomly assigned to respondents in the survey.

Let us have a closer look at the vignette universe of our example. Say our three dimensions are gender, employment status, and marital status. The corresponding two levels of each dimension are male/female, unemployed/employed, and married/not married. As we know from the previous slide, the corresponding vignette universe is composed of 8 vignettes because there are 8 possible combinations of the levels of all dimensions. The table on the right describes all of these 8 vignettes in our example. A useful feature of the vignette universe is that the dimensions and their interactions are mutually uncorrelated or, to use a more technical term, they are orthogonalised. Thus, whether a vignette person is male of female does not influence whether the person is unemployed or employed/married or not married. Also, the vignette universe is characterised by level balance, which means that the levels of the vignette dimensions occur with equal frequency. Orthogonality and level balance are desirable features. Level balance ensures that the impact of the dimensions on the outcome of interest is estimated with maximum efficiency. Orthogonality and the random assignment of vignettes to respondent is required to be able to interpret the results in a causal way.

In many applications, however, the vignette universe can quickly become quite large and include hundreds if not thousands of vignettes or more, as shown in this example from a study on unemployment and the willingness to accept job offers. This example shows 7 dimensions with three levels and one dimension with 5 levels, resulting in over 10,000 vignettes in the vignette universe.

Thus, researchers typically rely on samples of vignettes from the universe and respondents are only shown a subset of these vignettes. However, sampling vignettes from the universe is always accompanied with a loss of orthogonality and level balance. Some dimensions or their interactions might be correlated in the vignette sample.

In fact, researchers have to make several methodological decisions when designing factorial surveys. First, you need to decide on the number of dimensions and levels, which determine the size of the vignette universe. Researchers must also decide which characteristics need to be held constant to avoid confounding. Moreover, researchers have to decide on the number of vignettes to be sampled from the vignette universe. Also, respondents are typically not shown all the vignettes but only a subset of the vignettes. Researchers have to decide on how many vignette sets should be created. For example, if you have 100 vignettes, you could divide them into 10 vignette sets of 10 vignettes each. Researchers must also decide on the method to be used for composing the vignette sample and vignette sets. Finally, a restricting factor for setting up the experiment may be the size of the respondent sample. Often, you might have to try different combinations of dimensions and number of vignettes to achieve the most efficient design that is feasible in your research context. Next, we will talk about these issues in more detail.

Let us start with the number of dimensions. This slide shows a comparison between two vignettes, one with five dimensions and one with 12 dimensions. As you can see in the example on the right, the vignette text is much longer and respondents have to consider many characteristics simultaneously, whereas the example on the left seems much simpler and easier to evaluate.

It is important to note that dimensions should be selected based on theory and the research question. Regardless, however, there are several methodological issues to consider regarding the number of dimensions. This decision may also depend on the number of vignettes shown to each respondent, which is why I will discuss these issues together. A higher number of dimensions may lead to cognitive overload, particularly if combined with a large number of vignettes per respondent. Respondents might further engage in satisficing behaviour if confronted with too much information in the vignettes or a high number of complex vignettes. Greater numbers of dimensions also increase the risk of fatigue or learning effects. Respondents might get tired, resulting in less valid responses or learn about the researchers‘ intentions. Finally, the researcher might risk order effects with greater numbers of dimensions. Order effects mean that the order of the vignettes presented to respondents becomes confounded with the dimensions in the vignettes. In turn, a low number of dimensions might not sufficiently represent the object of the vignettes and be perceived as unrealistic by the respondents. In this case, respondents might find it too difficult to make judgments. Moreover, researchers should generally make sure that all theoretically relevant dimensions in a given research context are included in the design (either as experimental variables or as constants). Otherwise, the problem of omitted variable bias may arise if respondents make up the missing information when evaluating the vignettes. Finally, a low number of dimensions might lead to fatigue effects if there is not enough variation between the vignettes. This might be particularly the case in designs involving a low number of dimensions but a higher number of vignettes per respondent. Whether a given design is perceived as too complex or too simple also depends on the type of sample used, for example whether the sample is lower-educated or higher-educated on average. Thus, researchers should also keep in mind their target population when deciding on the number of dimensions and vignettes per respondent.

Let us turn to the number of levels. In general, only the levels necessary to test specific research hypotheses should be included. You should select levels that are of equal relevance. Respondents might otherwise focus on extreme values and ignore other levels leading to an underestimation of the impact of these levels on the outcome of interest (called range effects). Also, single dimensions comprising more levels than others might attract more attention from respondents as they are more likely to vary between the vignettes, leading to a higher impact of these dimensions on the outcome of interest (called the number-of-levels effect). Finally, illogical or implausible combinations of levels should be avoided. For example, someone who is employed cannot be unemployed for 2 years at the same time. Such combinations should be removed from the experimental design before creating the vignette universe. Removing illogical or implausible vignettes afterwards results in a loss of orthogonality and level balance.

There are several methodological studies, which provide guidelines on how to construct the experimental design in terms of the number of dimensions and levels as well as the number of vignettes per respondent. Some of this research are listed here, more research can be found in the supplementary reading list. The general recommendation from this literature is to use between five or nine dimensions, although a lower or higher number might be used if appropriate in the respective research context. The number of levels should be multiples of each other to avoid some of the methodological issues we have discussed. A balanced number of levels across vignettes also helps to create more efficient vignette samples, which we will talk about later. The literature also recommends not to use more than 10 vignettes per respondent. The order of vignettes should be randomised across respondents to avoid order effects. This means, for example, that Vignette 1 in your experimental design is not always evaluated first across respondents. If possible, also randomise the order in which information is presented within the vignettes, particularly in more complex designs. However, this might lead to confusion or unrealistic representations of a given scenario and you should carefully consider the appropriateness of such a design in your particular research context.

Once the vignette universe has been created, it must be decided whether a respondent should answer only one vignette or several. Designs with only vignette are called between-subjects designs. In within-subjects designs, each respondent evaluates all vignettes. In mixed designs, only different groups of respondents receive the same vignettes. Between-subjects designs minimize learning or fatigue effects but require a larger respondent sample size to achieve enough statistical power than within-subjects or mixed designs. For large vignette samples, a within-subject design is not advisable as respondents could be overburdened. As I said, we will focus on mixed designs, where a vignette sample is drawn from the universe and the vignettes are divided into different sets that are evaluated by respondents.

Three methods are mostly discussed in the literature for sampling vignettes from the vignette universe and for creating different vignette sets: random sampling, fractional factorial designs, and D-efficient sampling. Of course, in cases where you employ the full vignette universe, you would only apply these methods to create vignette sets.
Random sampling is relatively easy to implement as vignettes are randomly sampled from the universe. However, random sampling has been criticised to be less efficient, or only asymptotically efficient in large samples. Random sampling might further induce spurious correlations in the vignette sample or single vignette sets as there is no control over the structure of correlations between vignette dimensions.
In fractional factorial designs only a fraction of the vignette universe is used. The main effects of vignette dimensions always remain uncorrelated and typically only higher order interactions are correlated. However, the vignette sample might be balanced or unbalanced. Thus, fractional factorial designs are less suited for complex designs involving different numbers of levels per dimensions, as it gets more difficult to find orthogonal solutions. This also applies to creating single vignette sets based on this method.

D-efficient sampling relaxes the requirement of orthogonality and optimizes both orthogonality and level balance. In contrast to fractional factorial designs, it is therefore better suited for more complex designs. D-efficient sampling allows full control over which vignettes dimensions and interactions of dimensions are correlated in the vignette sample. Similarly, D-efficient blocking techniques can be used to allocate vignettes to different vignette sets while optimising the vignette sets for orthogonality and level balance.

Efficient sampling methods should be preferred over random sampling. The D-efficiency method in particular is recommended and widely used in the literature as almost a standard way to sample vignettes and to create vignette sets. This is because D-efficient designs allow full control over the correlation structure in the vignette sample and vignette sets, even in more complex designs. A D-efficiency score of 100 signifies perfect orthogonality and level balance, which can only be achieved in the vignette universe. Regarding the vignette sample, it is generally recommended that designs with a D-efficiency score over 90 should be selected. As I said, a D-efficient blocking technique is also recommended for building the vignette sets, which are also called decks. D-efficient blocking of vignettes to vignette sets ensures that orthogonality and level balance are optimised within each vignette set. Unfortunately, however, I am not aware of any software that allows to compute D-efficiency scores for vignette sets. Luckily, there is software available that can do all of these things for us, such as the SAS %Mktex makro or the AlgDesin package in R. The software draws the vignette sample for you and creates the vignette sets that you can then further prepare in other software (e.g. Stata) to build the actual vignettes.

Once you have created your vignettes, you need to decide how to measure your outcome of interest. Remember that the evaluations of the vignettes comprise your dependent variable. Thus, the type of response scale you select influences your possibilities for data analysis. As far as I can judge, rating scales are most often applied in vignette experiments. For example, a rating scale may range from unfair to fair. If using ratings, an 11-point scale is generally recommended. However, you may also choose other response scales, such as ranking or unordered categories. You can also create continuous measures by, for example, asking respondents to allocate money between vignette persons. In this case, the dependent variable would be the amount the respondents have allocated to each vignette person.

Once you have completed all the previous steps, your vignette experiment is ready for data collection. This step involves integrating your vignettes in a standard survey, such as an online survey or paper-pencil questionnaire. Regardless of which type of survey you choose, the survey must be set up in way such that each respondent is randomly assigned one questionnaire version, which in the case of factorial surveys equals one vignette set with a random order of the vignettes within the set. The randomisation can be done manually by the researcher before the data collection (e.g., in case of paper pencil questionnaires or if the respondent sample is known), or the randomisation can be programmed into the online survey using, for example LimeSurvey, for which you typically have to create questions for each vignette. I cannot go into detail on how to do this here, but you can find more information and practical examples in the book by Auspurg and Hinz. No matter what you do, it is imperative that you are able to identify which respondent has received which vignettes in which order.

In general, two data sets are important for data analysis once you have collected your data: the so-called experimental set up data and the respondent data.
The set-up data comprises the vignette sample or, in cases in which all vignettes in the universe were used, the vignette universe. The experimental set up data does not contain the vignette evaluations of respondents, only the experimental conditions.

In contrast, the respondent data, that is, the survey data that you collected contains the vignette evaluations and indicates for each respondent which vignettes they have evaluated in which order. However, it does not include the vignette dimensions. Of course, the respondent data may include other information that you collected such as respondent age.

In cases where unique vignette sets were created for each respondent before the data collection, the experimental set-up data would already include the information on the order of vignettes shown to each respondent. We will see examples of both types of set up data in the following slides.

In any case, both data sets need to be merged for the analysis. Therefore, it is crucial that the respondent data includes information on which respondent has received which vignettes.

Let us assume you have conducted an online factorial survey. As explained on the previous slide, your experimental set up data contains all the vignettes in your sample, which might look like this example. Say we have 36 vignettes in our sample, which were allocated to 6 different vignette sets each containing 6 vignettes. Say we have three experimental variables, gender, employment status, and nationality with a varying number of levels. The set-up data would then have 36 rows, one for each vignette showing the combination of levels of the three dimensions.

In contrast, the respondent data will look like the upper table on this slide. The respondent data will be organised in wide format; thus, each row corresponds to one respondent. This slide shows an example where each respondent has received four vignettes so that the table fits on the slide. However, it could also be six vignettes. For each vignette there is one column comprising the evaluations of each vignette (measured, for example, on a rating scale from 0 to 10). You would also have a column identifying the vignette set the respondent has received. You may have additional columns for some respondent characteristics.

To be able to merge the respondent data with the set-up data, the data need to be reshaped from wide into long format. In the long format, there are as many rows per respondent as they are vignettes that the respondent has evaluated. There is now one variable for the vignette ratings, which can later be used as dependent variable in a regression model. You have also a new variable for the vignette position, that is, the order of the vignettes shown to respondents. Using the variable indicating the vignette set, you can now merge the two data sets.

This slide shows one example of how a set up data would look if the randomisation was done before data collection (e.g., in paper questionnaires). The set-up data would then also have a respondent identifier and a variable indicating the order in which the vignettes are represented to respondents. In this example, the first respondent to participate in your survey would be assigned vignette set 6, in which vignette 32 would be shown first.

Once you have merged your data, you can start analysing it. Except for between-subject designs, your statistical model needs to account for the hierarchical data structure. Since each respondent rated multiple vignettes, the vignette evaluations are nested within respondents. Therefore, the vignette ratings within respondents are not independent from each other. To account for this clustering, you may apply ordinary least squared regression with robust clustered standard errors at the respondent level, or you may apply multilevel modelling. We will learn how to do this in Part III.

To sum up, let‘s review the different steps in conducting a factorial survey experiment. First, you need to decide on the number of dimensions and their levels, and which characteristics should be held constant to avoid confounding. Step 2 involves deciding on whether all the vignettes in the universe can be used or whether a sample of vignettes should be drawn. At this step, you would also check whether they might be any implausible or illogical cases. If yes, then you should consider going back to Step 1 to assess whether such cases can be avoided. Step 3 and 4 may be considered simultaneously. Step 3 involves deciding on the number of vignettes to be sampled from the universe and the method to be used to sample them. At the same time, you need to decide how many vignettes should be shown to respondents, which is Step 4. As discussed, D-efficient designs are recommended in both cases. These steps can be part of an iterative process, where you would go back and forth between different steps to find a feasible solution. Lastly, you need to decide on the presentation format of the vignettes and the response scale.

In Part III, I will show some practical examples of how to analyse the data using Stata. However, the analyses can easily be implemented in other statistical software (e.g., R) that allows to compute descriptive statistics and to perform regression analysis.

Thank you very much for your attention and have a good day.