The Past, Present and Future of Factorial Survey Experiments: A Review for the Social Sciences

Edgar Treischl & Tobias Wolbring
Friedrich-Alexander-University

Abstract
Factorial survey experiments (FSEs) are increasingly used in the social sciences. This paper provides a review about the use of FSEs and aims to answer three research questions. (1) How has this specific research field developed over time? (2) Which methodological advances have been made in FSE research and to what degree are they applied in empirical studies? (3) Which questions remain unresolved and should be addressed in future research? Using the Web of Science and Scopus databases, we conducted a literature review of FSEs published between 1982 and 2018. Our findings show that the field is developing quickly and that FSEs are becoming increasingly accepted in different research areas. Thereby, FSEs are being widely used not only to study attitudes, but also to explore the determinants of behaviour. Most research applies state-of-the-art techniques in terms of statistical analysis; however, to a lesser extent, studies rely on more sophisticated sampling procedures to draw samples from a large vignette universe. Finally, several methodological questions remain unresolved concerning the realism and complexity of vignettes, social desirability, and the predictive validity of FSEs regarding behaviour due to their hypothetical nature. Against this background, we call for more methodological research to assess the general applicability of FSEs for different research areas. Further, our review suggests the need for better documentation and reporting standards to evaluate methodological aspects of FSEs.

Keywords: factorial survey experiments, methodological advances and pitfalls, predictive validity, realism of vignettes, vignette design
In 2009, Lisa Wallander published a highly cited review article about factorial survey experiments (FSEs). As she pointed out, many scholars were not familiar with them or had substantial reservations against them at the time, even though they had been introduced over three decades prior (see Jasso & Rossi, 1977; Rossi et al., 1974; Sampson & Rossi, 1975). As a result, empirical studies using FSEs were scarce. Figure 1 displays the number of articles published between 1982 and 2018 that refer to an FSE and have been identified in our review. As Figure 1 shows, only a few papers using FSEs were published every year until 2006, which was the last year covered in Wallander’s review. Further, FSEs were virtually absent in leading social science journals.

A decade later, the situation has changed: FSEs have been introduced into survey methodological handbooks (see Aviram, 2012), textbooks are available that explain how to design and conduct FSEs in detail (see Auspurg & Hinz, 2015a; Mutz, 2011), and multifactorial survey experiments are becoming increasingly popular in the social sciences (see Atzmüller & Steiner, 2010; Auspurg & Hinz, 2015b; Jasso, 2006). In accordance with this trend, the number of publications using FSEs has risen markedly since 2006, as Figure 1 indicates. Several of these studies were published in leading journals, such as the American Sociological Review and the European Sociological Review (e.g. Auspurg et al., 2017; Graeff et al., 2014; Wouters & Walgrave, 2017), which further illustrates the increasing use and acceptance of FSEs.

Given the popularity of FSEs and the increasing number of empirical applications since Wallander published her influential review article, we think it is time for an update. Hence, we focus on how the field has developed, which methodological advances have been made, and which challenges of the approach remain unresolved. For this reason, we conducted a literature review covering all articles from Wallander’s review article (1982–2006), as well as more recent applications involving FSEs (2007–2018). For each publication, we collected information about the study topic, research design, outcome measures, and statistical analysis. This gives us the opportunity to make three contributions to the current literature on FSEs. First, we provide an overview of the past and current use of FSE in the social

Acknowledgements
This paper could not have been realized without the support of Janette Buchmann, Sarah Glaab, Nora Spielmann, and Philipp Überall during the data collection process. For helpful feedback, we want to thank Johanna Gereke, Eva Zschirnt, the anonymous reviewers, the editors as well as all participants at the European Sociological Association midterm conference 2018 in Cracow.

Direct correspondence to
Edgar Treischl, Friedrich-Alexander-University, Erlangen-Nürnberg (FAU), Findelgasse 7/9, 90402 Nürnberg, Germany
E-mail: edgar.treischl@fau.de
Second, since some recommendations on how to design and analyse an FSE have been published in recent decades, we briefly introduce readers to these methodological advances, and we examine to what extent these techniques have entered applied research. Methodological advances can help to improve both the internal validity of inferences and the statistical power. Thus, one goal of our review is to

---

1 Choice experiments are another kind of multifactorial survey experiment. In choice experiments, participants are directly confronted with varying trade-offs between two or more alternatives, and are asked to choose between the proposed alternatives. Choice experiments appear to be especially well-suited to studying human decision-making since they are theoretically grounded in the characteristics theory of value (Lancaster, 1966) and random utility models (McFadden, 1974; Manski, 1977). In this review, we focus on FSEs as the most widely used type of multifactorial survey experiment in the social sciences, while choice experiments are more frequently employed in business studies and economics (for the potentials and challenges of choice experiments in the social sciences see Liebe & Meyerhoff, 2021).
give researchers some general background on how to design state-of-the-art FSEs and to provide references for more detailed follow-up.

Finally, we also aim to provide guidance for future methodological research by highlighting unresolved questions in the growing, but small, methodological literature about FSEs. We focus on three partly interrelated issues that have caused controversial discussions within the scientific community, as they may have far-reaching consequences for the validity of FSEs: the realism and complexity of vignettes, concerns regarding the hypothetical nature of the outcome measures in FSEs, and the risk of social desirability bias. While methodological research on these topics is still scarce, we underscore some findings from recent research about the design of FSEs, and contrast these methodological recommendations and insights with current research practices as identified by our literature review.

The remainder of the paper is structured as follows: First, we introduce the FSE approach and provide some methodological background on it. Next, we outline in more detail how we conducted the literature review and describe the dataset. Furthermore, we explain some recent methodological advances and discuss unresolved questions such as the required degree of realism and complexity of vignettes, as well as the link between stated and actual behaviour. Finally, we emphasise key insights and opportunities for future research to deepen our methodological knowledge of FSEs.

The Basic Idea Behind Factorial Survey Experiments

This section outlines the basic idea behind FSEs. Respondents encounter textual descriptions or visual stimuli of a hypothetical situation (vignette or scenario) in an FSE and are asked to rate the scenario. Each vignette contains one or several characteristics (dimensions/factors) that systematically vary across vignettes. Survey participants are randomly assigned to one (between-subjects design) or several (within-subjects design) vignettes, and are asked for their opinion on a certain situation or the intended behaviour in the described scenario.

Figure 2 displays two examples of vignettes. Example A is a vignette by Opp (2002). He examined under which circumstances an anti-smoking norm emerges by eliciting normative judgements. Single dimensions that may have a causal impact on respondents’ opinions are in italics to illustrate the experimental variation across the vignettes. Example B comes from a study by Teti et al. (2016). They

2 Several excellent textbooks about FSEs have been published (see Auspurg & Hinz, 2015a; Mutz, 2011) since the approach was first introduced to the social sciences by Rossi et al. in 1974. This section relies heavily on these textbooks, which provide a more detailed discussion about the fundamentals of FSEs.
asked elderly respondents to make hypothetical relocation decisions and investigated whether FSEs can be applied in housing research.

An FSE combines the methodological rigour of an experimental design with the advantages of survey research by including an experimental research module in a survey and assigning participants randomly to one or several hypothetical descriptions of a situation. This facilitates inferences from experimental results to a target population (see Auspurg & Hinz, 2015a, p. 12-13; Mutz, 2011, p. 10).

Observational studies may suffer from various methodological pitfalls—such as confounding by self-selection of participants and unobserved heterogeneity—thus impairing the identification of causal effects (Rosenbaum, 2010; Shadish et al., 2002). As is well-known, experimental designs have advantages regarding causal inference and, at least in theory, can outperform non- or quasi-experimental designs in regard to issues of internal validity (for the principles of experimental design, see Imbens & Rubin, 2015; Jackson & Cox, 2013). An FSE offers the possibility of estimating the causal effect of a varied dimension on the outcome variable. Random assignment helps to avoid threats to internal validity such as confounding and selection bias. Direct manipulation of treatments (in the FSE, the varied dimension) secures causal ordering, and including a control group avoids biases due to matura-

**Example A (Opp, 2002):**

*Mr. Müller* goes to a restaurant. This is a *top class restaurant* in which smoking is prohibited. There *is nobody* in the restaurant who smokes. *Mr. Müller stays only for a short time* to drink a beer. He smokes most of the time, *more than a package of cigarettes per day.*

**Example B (Teti et al., 2016):**

Imagine that the apartment offered is in *your current district*. It is located very *centrally*, a *2 minute walk* from the nearest bus/train station and *far away* from the home of your daughter/son. The apartment is in the *3rd floor*, *has no elevator*, and has a *large bathtub* (no shower) and a *balcony without steps.*

**Figure 2** Two examples of vignettes
can be avoided by using textual descriptions of hypothetical scenarios instead of actual interventions in the ‘real’ world.

It is easier to sample non-students and ‘non-weird’ people for a survey than to recruit them for lab experiments. Including an experimental module in the survey allows scholars to conduct a population-based FSE, promising broader generalisability beyond potentially selective subgroups such as students. Variations are much easier to implement in an FSE due to the manipulation of textual descriptions. Ethical concerns and practical restrictions do not apply to the same degree as in the lab or in the field. Accordingly, treatments that are hard to implement in the field can be investigated in an FSE. For this reason, an FSE can also help to inform policy about hypothetical worlds and potential interventions discussed in the public discourse without taking the risk and covering the costs of an actual implementation. As the following review shows, FSEs are increasingly being used in the social sciences, even though FSEs also face methodological pitfalls and challenges.

**Literature Review**

After this short introduction to FSE, this section informs about the literature review in two sub-sections. In the first sub-section, we provide details about the data collection process, search strategy, and inclusion restrictions for the literature review. In the second sub-section, we update Wallander’s review by describing our analytical sample in terms of research areas (e.g. topics, the respondents’ countries of origin) and methodological choices (e.g. sampling strategies, between- vs. within-subjects designs).

**Data Collection**

Our literature review is based on a combination of three different approaches to secure broad coverage of FSE publications. First, we covered all 106 publications that Wallander (2009) identified. Second, we made use of the popularity of the first review paper and collected publications citing it. In 2019, Wallander had over 400 citations according to Google Scholar, including many recent FSE applications. Third, we searched for empirical applications of FSEs using the Web of Science and the Scopus database for the time period covered by Wallander (1982–2006), as well as more recent years (2007–2018).³

³ For identifying relevant publications among papers citing Wallander (2009), we applied the same search strategy and criteria as outlined in the third search strategy. However, among those publications many appeared as monographs or were grey literature (such as working papers, project reports, and presentation slides). Consistent with the third search strategy, we did not include them in the review.
We applied the following restrictions to identify publications relevant for our review. The review included publications that refer to ‘factorial survey (experiments)’, ‘vignette study’ and ‘vignette experiment’ in the title, abstract, or keywords. Hence, the review covers FSEs, but not publications with related but different survey experimental research designs, such as conjoint analysis and discrete choice experiments. To identify core articles for the social sciences, we considered publications published in journals listed in the Social Sciences Citation Index (SSCI) of the Web of Science and the category ‘Social Sciences’ of Scopus. We did not cover other document types such as monographs, or conference articles. Further, we only took publications written in English into account. The following search string was used to identify FSEs using the Web of Science:

\[ \text{TOPIC: ("Factorial survey") OR TOPIC: ("Factorial survey experiment") OR TOPIC: ("Vignette study") OR TOPIC: ("Vignette experiment") Refined by: excluding \text{PUBLICATION YEARS: (1974 – 1981 OR 2019 OR 2020 OR 2021) AND DOCUMENT TYPES: (ARTICLE) AND LANGUAGES: (ENGLISH) AND WEB OF SCIENCE INDEX: (WOS.SSCI)}} \]

\text{Timespan: All years. Indexes: SCI-EXPANDED.}

Similarly, the following search string was used to identify FSEs with Scopus:

\[ \text{(TITLE-ABS-KEY (("factorial survey experiment") ) OR TITLE-ABS-KEY (("factorial survey") ) OR TITLE-ABS-KEY (("vignette study") ) OR TITLE-ABS-KEY (("vignette experiment") ) AND PUBYEAR > 1981 AND PUBYEAR < 2019 AND (LIMIT-TO (SUBJAREA, "SOCI") ) AND (LIMIT-TO (DOCTYPE, "ar") ) AND (LIMIT-TO (SRCTYPE, "j") ) AND (LIMIT-TO (LANGUAGE, "English") )} \]

Based on those search strings, we identified 148 publications in the Web of Science and 301 publications in Scopus for the entire time period (1982–2018; last search date: 26 March 2021), with substantial overlap between the two databases. After taking into account the overlap, 353 publications remain in the sample from the Web of Science and Scopus.

Upon closer inspection it turned out that not all of these 353 publications meet the scope condition of our review to report on empirical applications of FSE in the social sciences. Different reasons lead to the exclusion of some publications. The applied exclusion restrictions were as explained in the following (for the excluded number of publications by criteria see Table A1 in the online appendix). Further

---

4 We used the displayed search string to collect data from the Web of Science last time in March 2021. After the last search, the Web of Science database received various substantial updates and extensions in 2021, including a fundamental update of the search tool and a new code structure to search for publications. As a result, the reported search string of our review is not working with the recent version of the Web of Science.
inspection of the publications showed, that some publications did not employ an
FSE but introduce the FSE methodology (e.g. Taylor, 2005). In a similar vein, some
publications report only results of a pilot study to introduce FSEs or discuss them in
light of a specific research area (e.g. Liebig et al., 2015). In some instances, qualita-
tive researchers use vignettes and many health-related research use case scenar-
ios (Kiesewetter et al., 2018) to describe a scenario, but without applying typical
aspects of FSEs (e.g. random assignment, varying dimensions). In addition, past
research sometimes confounds FSEs with factorial experiments (e.g. Baker, 1983).
We did not include these in total 85 studies in our review.

In addition, a small but growing number of publications address methodologi-
cal research questions on FSEs (e.g. for varying the number of vignette dimensions,
see Auspurg & Jäckle, 2017). In the following, we will only provide statistics on
substantive research using FSEs, and exclude review articles, as well as method-
ological research on FSEs. However, these contributions are part of our discussion
on methodological advances. In this part of the review, we also discuss insights
from more recent methodological contributions.

Finally, we decided to focus on studies with textual vignette descriptions
(including tables), but did not include studies using visual stimuli, such as pictures
and video vignettes in our review (e.g. see Oberoi et al., 2016; Wouters & Walgrave,
2017). The main reason was that these studies are often not completely comparable
with research using text vignettes: Studies using video or photo vignettes often vary
a lower number of dimensions due to the effort involved in manipulating visual
stimuli, and the cognitive processes of the respondents when seeing visual stimuli
might be fundamentally different from those when reading text.

Hence, our final dataset based on Scopus and the Web of Science contains 261
publications that met the described criteria. Moreover, the final data considers all
106 publications identified by Wallander (2009) and 74 publications that cite Wal-
lander (2009) and were not listed in the Web of Science or in Scopus. Overall, our
final analytical sample contains 441 publications. A list of included publications
can be found in Table A2 in the online appendix.

We then created a dataset containing detailed information on each publica-
tion. We retrieved most information from the ‘data and methods’ section and, in
some instances, from the ‘appendix’. To summarise how the field has developed,
we collected information about the research topic and classified the outcome mea-
surement of each publication to indicate whether respondents were asked to make
a hypothetical judgement (e.g. fairness of earnings) or to state a behavioural inten-
tion (e.g. willingness to pay for a service). In addition, the data contain information
about the survey sampling strategy, the number of vignette dimensions, the mea-
surement of the outcome, the vignette sampling strategy, and the applied statistical
analysis.
Unfortunately, some publications did not report details about the FSEs in terms of design. In particular, several recent publications did not contain information about how vignettes have been sampled from the vignette universe. The fact that we could not retrieve this information, even after an extensive search, is alarming and calls for establishing standards of how to document and report design aspects of FSEs.

All publications were classified by three different coders. The interrater reliability between the three raters was sufficient ($r=0.89$) for numerical indicators such as the number of dimensions, the number of ratings, or the outcome measure. In contrast, we found the lowest interrater reliability ($r=0.54$) for the binary indicator for sensitive research topics. We have not given detailed statistics on sensitivity in the paper, but cover the topic in our discussion on the predictive validity of FSEs.

### Description of the Sample: Research Areas

Before focusing on methodological advances and unresolved concerns, we update the work of Wallander (2009). Figure 3 plots the top 10 FSE publication topics before and since 2007 in terms of absolute and relative frequency. We identified research areas in which FSEs have been increasingly used since Wallander’s review, which is why we centre on 2007 as the cut-off point and examine the development of FSE publications before and since 2007. As classification is sometimes not straightforward (e.g. research on school-to-work transitions), the categories of the classification are not disjunct. An article could fall in two or more categories.  

As Figure 3 shows, most studies (N=58) have used FSEs to study crime and justice topics (Lyons, 2008; Tolsma et al., 2012). This research area was the most prevalent topic among FSE publications until 2006. The number of published FSEs about justice has decreased to 43 since 2007, but FSEs are still often applied in this field. In contrast, FSEs are increasingly applied in other areas over time. The categories health and care and work display the highest increase in absolute numbers, with 68 and 42 applications since 2007. Overall, 37% of all published FSEs that we identified examine health and care-related topics, such as care planning and needs (Baughman et al., 2019; Jörg et al., 2006), or work-related topics, such as hiring intentions (Di Stasio & Gëruxhani, 2015; van Belle et al., 2018). However, research is not restricted to these topics. FSEs are used to study diverse aspects, and, as a result, we did not classify a certain number of studies under a separate category, but rather as other topics (overall 9%). This includes research about sport behaviour (Chatfield et al., 2018), corruption (Graeff et al., 2014), and the willingness to

---

5 For instance, Haase et al. (2016) examined the male breadwinning model, a topic that might be included in the work or family category. In such an instance, we included the article in both categories.
provide (para) data (Couper & Singer, 2012). In addition to the top 10 topics, FSEs have less frequently been used to study conflict behaviour (see Baron et al., 2001; Bell & Forde, 1999), consumption (Moynihan, 2013; Cahan, 1996), and mobility behaviour (see Abraham et al., 2010; Teti et al., 2016). We classified, but excluded these topics in Figure 3 due to the small number of publications since 2007.

With regard to respondents’ countries of origin, Wallander (2009) reported studies from seven different countries, but over 80% were based on populations from the US. For our review, we observed 294 articles from 41 different countries since 2007. Many studies rely on US populations (31%), but a substantial number of publications come from other countries, chief among them the Netherlands (14%), Germany (13%), and the UK (7%). The growing number of respondents’ countries of origin also illustrates the increased popularity.

---

6 In recent years, the amount of methodological research has grown as well, but methodological research remains rare in comparison to hot topics and the overall amount of FSE studies.
Description of the Sample: Applied Methods

In addition to the diversity of FSE regarding research topics and respondent’s’ country of origin, Wallander (2009) reported that almost every second study aims to make inferences from experimental results to a general population. However, in the first review, it remained unclear which sampling strategy most researchers used for such inferences from the sample to apply to the target population. In the case of a non-probability sample, experiments still provide internally valid estimates of a treatment effect due to the random assignment of subjects to treatment conditions. However, the effect estimates of a convenience sample cannot necessarily be generalised beyond the specific subgroup under investigation in the case of effect heterogeneity across individuals. As our data show, approximately 47% use probability and 53% non-probability samples. Within the group of studies using non-probability samples, most authors have relied on convenience samples, in particular from the student body. However, in a few cases, researchers used referral (4%) and purposive samples (2%). Reflecting most recent studies, another 9% of non-probability samples have used samples from the crowdsourcing website Amazon Mechanical Turk (M-Turk). Hence, while the use of non-probability sampling might often be unproblematic for generalising experimental results if one is willing to assume the absence of effect heterogeneity, most FSEs do not fully utilise the potential of FSEs to generate ‘representative’ samples. As a consequence, in the most extreme case, the findings from a part of the literature might not be generalisable to the target population. In addition, the use of non-probability samples raises questions about the adequacy of inferential statistics, which is frequently applied with such data.

As Wallander revealed in her review, many studies have not applied methods for clustered data, even though a substantial number of vignette studies depend on a within-subjects design with two or more vignette ratings per respondent. In our review, 86% of the studies relied on a within-subjects design. The average number of vignettes per person is nine, with a maximum of 110, and 50% of the studies ask for five or more ratings. Given the broad use of within-subjects designs, the hierarchical structure of the data needs to be considered: Each person provides several ratings. Consequently, single observations are clustered and are no longer independent from each other. Clustered data violate the assumption of regression analysis that residuals are independent and identically distributed. Without adjustments for the clustered data structure, standard errors from a regression analysis are biased. The two most common ways to address this problem are (a) multilevel models with random or fixed effects and (b) robust standard errors clustered around individuals (see Maas & Hox, 2004; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Figure 4 contains a proportional stacked area chart to display the proportions of statistical methods used to analyse FSEs with a within-design over time. Forty-six percent of published articles in 2000–2004 presented the results of a regres-
sion analysis without taking clustered data structure into account. This proportion fell considerably over time. Most studies published since the first review paper no longer ignore the issue. In the period of 2015–2018, the majority of most recent publications used methods for hierarchical data (69%) or relied on cluster-robust standard errors (26%), while only 5% did not take clustering into account.

**Methodological Advances and Unresolved Questions**

This section discusses more recent methodological insights and advances in the design of FSEs. First, we introduce ways to improve the efficiency of the vignette design in the face of a large vignette universe. Second, we provide recent methodological findings regarding the consequences of the vignette design for the validity of the results. We place particular emphasis on the complexity and realism of vignettes, but also cover issues of presentation style and choice of response scales. Finally, recent scholarly publications have examined the relationship between behavioural intentions stated in FSEs and actual behaviour. We discuss why FSEs may or may not help to provide insights into human behaviour, and provide an over-
view of existing studies examining the predictive validity of FSEs. While not all these methodological discussions will lead to clear recommendations, we believe it is important to draw attention to these topics, both for a reflected use of FSEs by applied researchers, and to provide motivation for future methodological research.

**Design Efficiency**

The **vignette universe** contains all vignettes, which result from the combination of all levels of each dimension (**full factorial**) in an FSE. For example, the vignette universe of an FSE with seven dimensions and four levels of each dimension contains $4^7=16,384$ unique vignettes as a result of the Cartesian product. Researchers can use the full factorial in the case of a small universe. However, the vignette universe quickly becomes very large. In such an instance, scholars must construct an experimental design such as **random sampling**, **randomised block confounded factorial (RBCF) designs**, and **D-optimal designs** to draw a smaller, more manageable subset from the vignette universe.

Random sampling techniques reduce the number of vignettes by drawing for each respondent a random set of vignettes from the universe. Random sampling techniques generate an orthogonal (FSE dimensions are uncorrelated) and completely unconfounded vignette set if the vignette sample approaches infinity, but not necessarily for smaller vignette sample sizes (Jasso, 2006; Su & Steiner, 2020). Both other fractional factorial designs try to actively increase the efficiency of the experimental plan compared to random sampling techniques. For instance, RBCF designs use experimental plans to split the vignettes into several vignette sets of equal size, such that only higher-order interaction effects are confounded with the sets. In a similar manner, a D-optimal design is the outcome of a computational optimisation process. A D-optimal design tries to maximise the precision of parameter estimates by searching for an orthogonal and balanced (levels have

---

7 This illustrates another advantage of FSEs. The levels of different dimensions are often highly correlated with each other in surveys and other observational studies (see Auspurg & Hinz, 2015a, p. 10). By using the full factorial, all considered dimensions in an FSE are uncorrelated by design (orthogonal) due to the combination of all dimensions and levels. Orthogonality is a main strength of FSEs since the causal effect of each dimension is identified in such a design.

8 Design efficiency refers to the statistical power of the vignette sample (experimental design) to estimate parameters for main dimensions and interaction effects with a high degree of precision. For more information about design efficiency and recent developments, see Dülmer (2016) or Su and Steiner (2020).

9 As Su and Steiner (2020, p. 36) denoted: “RBCF designs are typically restricted to simple designs with a few factors and ideally the same number of factor levels. For more complex designs that involve large vignette populations, generated from a large number of factors (i.e. five or more factors) with unequal numbers of factor levels (i.e. 2–10 or more levels), adequate RBCF designs might not exist or be challenging to construct”.

equal frequencies) vignette sample of the universe (see Dülmer, 2016). A computer algorithm searches iteratively for combinations of each dimension to optimise the precision of parameter estimates for all main effects and may—depending on specifications—also optimise precision for two-way or higher-order interactions (see Kuhfeld et al., 1994).

Thus, such fractional factorial designs have advantages compared to random samples. Researchers do not have to rely on chance that the random sample is the most efficient design and that key assumptions such as orthogonality hold. For instance, research indicates that D-optimal designs outperform random sampling techniques and a full factorial in the case of a small random sample from the vignette universe due to higher statistical power to estimate interaction effects (Dülmer, 2007). Especially in the social sciences, interaction effects are often of major interest. A random sample is not ideal to estimate these effects, and sometimes interactions are not identified by the design at all.

However, most past research has used random samples of the vignette universe. Only one study (Buskens & Weesie, 2000) out of 44 articles that relied on a vignette sample also used a fractional factorial design until 2006 (see Wallander, 2009, p. 512). In addition to the fact that those designs were not very well-known back then, most statistical software packages had not implemented packages at that time to draw a D-optimal design and to calculate a design’s efficiency. Although fractional factorial designs are now broadly accepted as a useful sampling technique, the unfortunate situation on the software side remains almost unchanged (for an implementation in SAS, see Kuhfeld et al., 1994). Hence, we suspect that an increasing, but still minor share of recent studies uses a D-efficient design.

Our literature review corroborates this apprehension. Figure 5 depicts the use of different vignette sampling techniques over time based on a proportional stacked area chart. Random vignette samples were the most common technique in the period of 2000-2004. Seventy-one percent of all FSE articles report using a random sample of the vignette universe. As Figure 5 shows, there is a clear time trend. While the large majority of studies published during 2000–2004 used random samples of vignettes, only 25% of the identified publications after 2014 did so. However, random sampling techniques are far from being fully replaced by fractional factorial or full factorial designs, although we find an increasing amount of both, especially for full factorial designs since 2000. Unfortunately, a small, but growing amount of research does not provide any information about the process of vignette sampling.

In sum, random sampling techniques are easy to implement and might be a sufficient choice in the case of a small vignette universe and large sample sizes. However, random sampling techniques come with the risk of potentially confounding main and interaction effects, and require untestable assumptions about the absence of certain interactions. RBCF and D-optimal designs help to avoid
such threats to internal validity by securing the orthogonality of main and interaction effects, and often increasing statistical power. Even if more time investment is needed to determine and implement these designs, we especially recommend using them in the case of small samples and a large vignette universe to avoid confounding and underpowered FSEs.

The Realism and Complexity of Vignette Designs

Decisions about the design of a vignette can have far-reaching consequences in terms of internal and external validity. In terms of complexity, a very simple scenario with only a few dimensions and rather low variation across several presented vignettes may lead, on the one hand, to boredom and fatigue effects in within-subjects designs. On the other hand, very detailed scenarios with many dimensions may seem more realistic, but providing too much information may cause cognitive overload, especially if the number of ratings is high. Participants may no longer be able or willing to pay attention to the vignette or to all provided dimensions in the case of information overload. Instead, participants may switch to response sets, use cues and heuristics to come to a decision without too much cognitive effort. Such satisficing behaviour is well-known for conventional survey items (see Krosnick, 1991) and can also occur in different forms in FSEs (see Shamon et al., 2019). For
example, the findings of Auspurg and Jäckle (2017) imply that a large number of dimensions (e.g., 12 dimensions) can lead to order effects.\footnote{Auspurg and Jäckle (2017) further found that respondents’ degree of uncertainty about a topic influences the likelihood of order effects, while other studies discovered no (Robbins & Kiser, 2018)—or at least no strong—evidence for order effects of FSE dimensions (Düval & Hinz, 2020).}

As a consequence, researchers should avoid both too simple and too complex vignettes as well as unrealistic, implausible, and illogical scenarios (e.g., a professor without a school degree). Research shows that the use of such scenarios reduces the internal validity of inferences, because respondents no longer pay attention to the dimensions or, in the worst case, do not take the survey seriously (see Auspurg & Hinz, 2015a, pp. 40–42). That said, how many dimensions should approximately be provided to prevent boredom effects and cognitive overload among participants? The current state of research recommends seven dimensions to provide a good balance between simplicity and complexity (see Auspurg & Hinz, 2015a, pp. 18–22; Sauer et al., 2011). However, this is just a rough rule of thumb, since the choice should be guided by theory and depends on other factors such as research topic, survey length, respondents’ motivation and cognitive skills as well as other FSE design aspects. As Wallander (2009) reported, the number of dimensions in FSE studies published until 2006 has varied greatly between two (Steen & Cohen, 2004) and 25 (Thurman et al., 1988), with a median of six dimensions (see Wallander, 2009, p. 512). As Figure 6 indicates, this finding still holds for recent studies, which frequently deviate from this rough seven dimensions rule of thumb. While the average number of dimensions used in prior research since 2006 is 5.7, a substantial amount of research provides more than nine or less than five dimensions. Overall, 57% fall into the range of seven dimensions, while 38% of the publications provide fewer and 5% more vignette dimensions. Even after restricting the sample to FSE studies that are (a) more recently published and (b) have several ratings per person, we found that 42% of the studies use more or less vignette dimensions than suggested by the methodological literature.

Another important aspect concerning complexity is the presentation style of the vignette. Most researchers use text vignettes, while other forms of multifactorial survey experiments, such as conjoint analysis and choice experiments, often present FSE dimensions in tabular format. The cognitive load of reading a table is likely lower than reading a text with or without highlighted dimensions, which might affect response behaviour. Only recently has the first research about the differences between both presentation styles in FSE been published. Based on a student sample, Sauer et al. (2020) found no significant differences between presentation styles in relation to vignette rating and non-response. In contrast, Shamon et al. (2019) reported less non-response, in particular refusals, for a tabular presentation style.
than for textual scenario descriptions with and without underlining varied information, especially to the less-well educated people. Thus, given the results of these two studies that rely on different samples, the presentation style may affect response behaviour and data quality especially for less educated respondents but may matter less for other participants. Thus, keeping the limited number of studies in mind, one may cautiously conclude that using a tabular format may not hurt in some contexts, but may be beneficial depending on respondents’ background. Since these are just first preliminary conclusions, more research needs to address under which circumstances—including the realism of the vignette and the complexity of the examined topic—the presentation style may affect the quality of the data.

Finally, the realism and complexity of a scenario also depend on the information provided and omitted. FSEs rely on the important yet underappreciated assumption of information equivalence (Dafoe et al., 2018). Participants need all relevant information necessary to assess a situation and to provide a meaningful answer. If a vignette lacks important aspects, respondents may update their beliefs and fill in the missing pieces in accordance with their expectations or stereotypes. 

Note: The histogram shows the number of dimensions in FSE applications (2007–2018) with at least two ratings per person. The red dotted line displays the mean value of all included dimensions, a grey solid line displays the rule of thumb, and grey dashed lines denote the threshold of the rule of thumb.

Figure 6  Number of dimensions in FSE publications 2007-2018
Given that respondents’ expectations and stereotypes might not be exogenous to the individual background and the presented treatments, the lack of relevant information may lead to biased inferences about effects and causal mechanisms at work, which violates the assumption of information equivalence since individuals base their response on different information. For this reason, Dafoe et al. (2018, p. 406) proposed and evaluated three strategies for achieving information equivalence. The first strategy—encouraging respondents to think of an abstract instead of a real-world scenario—turned out to be ineffective. In contrast, the second strategy—using covariate control by specifying background details to prevent respondents from updating their beliefs—helped at least to reduce imbalance for the specified variables. The third strategy relies on framing the vignette scenario as the outcome of a random assignment process. Respondents are told that the treatment is the outcome of a random process (e.g. lottery, natural experiment) to make respondents believe that the treatment is not correlated with other, omitted dimensions, which may have an impact on the respondent’s vignette rating. The third strategy turned out to be most effective in the study, while it remains open to future research to examine how effective this strategy is in other contexts. Irrespective of the findings of follow-up research, it becomes clear that design choices determine how realistic respondents perceive the described scenario to be, and how internally and externally valid inferences from the FSE will be. As our discussion underlines, this task is not just a technical exercise, but requires theoretical guidance and in-depth knowledge of the research topic under investigation.

**Predictive Validity**

Figure 7 depicts the number of articles in our analytical sample across time based on a stacked area chart. As Figure 7 shows, FSEs are increasingly used not only to study attitudes and hypothetical judgements (dark grey area), but also to explore the determinants of behavioural intentions (light grey area). On average, in each time period, approximately 45% of all FSE studies focus on behavioural intentions, with the strongest boost since 2010. For instance, FSEs are currently used to study willingness to pay (see Bekkers, 2010; Bridoux et al., 2016), hiring and job decisions (see Di Stasio & Gërxhani, 2015; van Belle et al., 2018), mobility behaviour (see Abraham et al., 2010; Teti et al., 2016), or medical and care decisions (see Drewniak et al., 2016; Shlay, 2010).

Obviously, an FSE does not measure actual behaviour, but asks participants to assess a hypothetical scenario based on the information provided. Hence, FSEs gauge self-reported behavioural intentions in a hypothetical situation. Thus, an important question regards the predictive validity of such measures (see Eifler & Petzold, 2019; Petzold & Wolbring, 2019): To what extent do hypothetical intentions correspond with real-world behaviour?
A stated intention does not always correspond very well with real-world behaviour (see Barabas & Jerit, 2010; Collett & Childs, 2011). As the theory of planned behaviour (Fishbein & Ajzen, 2010) suggests, and as outlined by Petzold and Wolbring (2019) for FSEs, intentions may only translate into actual behaviour under certain conditions. For instance, actors might plan to act in a certain way, but in reality lack behavioural control or face the high costs of an action. For this reason, FSEs are sometimes criticised for lacking ‘psychological realism’ and predictive validity. In contrast, one could argue that although behavioural intentions do not perfectly predict real-world behaviour, they are important determinants of actual decision-making. Thus, showing what influences behavioural intentions might provide insights into the determinants of human action.

Unfortunately, the current state of research is small and inconclusive about the predictive validity of FSEs. Some evidence suggests low behavioural validity, with substantial differences between hypothetical decision-making and a behavioural benchmark regarding the distribution of the outcomes and their determinants (e.g. Pager & Quillian, 2005; Findley et al., 2017). In contrast, another group of studies concluded that FSEs have high predictive validity, and that the dimensions of an FSE sufficiently correspond with a behavioural benchmark (see Drasch, 2017; Hainmueller et al., 2015; Nisic & Auspurg, 2009; Raub & Buskens, 2008). Finally, a third group of studies offer results that are both partly in line with the first and the second position (e.g. Barabas & Jerit, 2010; Eifler, 2010; Petzold & Wolbring, 2019). They document that distributions of intended and actual behaviour clearly deviate from each other, indicating that other factors (such as social desirability and the costs of an action) co-determine decision-making in the real world. Despite the
reported differences in levels, these studies found that FSEs seem to provide correct estimates of behavioural determinants regarding direction and relative effect sizes (see, however, Barabas & Jerit, 2010).11

Thus, the current state of research does not justify generally rejecting FSEs as a way of generating insights into determinants of behaviour, or using FSEs uncritically for all research questions in terms of behaviour. However, this ambiguous state of research raises several questions for future research that should be kept in mind when deciding whether to use an FSE to answer a specific research question and how to design it. In particular, the state of research raises the question: Under which conditions does an FSE have higher or lower predictive validity regarding human behaviour? Different factors must play a role in such theoretical considerations, including methodological aspects that affect the realism of the vignette, respondents’ experience with the decision situation, and the sensitivity of the topic under investigation.

Concerning sensitivity, Wallander (2009) reported numerous FSEs on a wide range of topics that might be potentially affected by social desirability. Previous research has used FSEs to study racial prejudice (Shlay, 1986; St. John & Healdmoore, 1995), sexual harassment (Hunter & McClelland, 1991; Weber-Burdin & Rossi, 1982), and drinking and driving behaviour (Applegate et al., 1996; Thurman et al., 1993). Past research on sensitive questions and social desirability bias suggests that FSEs are better suited for studying sensitive questions than direct questions (see Alexander & Becker, 1978; Auspurg et al., 2015), and seem to outperform specific survey techniques such as the randomised response technique, which has been developed to attenuate social desirability bias in surveys (Armacost et al., 1991). Being better suited than other methods to attenuate social desirability bias does not imply that FSEs cannot suffer from social desirability bias and are adequate for examining sensitive topics. Social desirability might still be substantial and undermine causal inferences, especially if respondents become aware of the research topic. Respondents may quickly realise what the actual focus of the FSE is if the number dimensions is low, if the variation of the vignette is highlighted, or if several vignettes are rated sequentially in a within-subjects design. To our knowledge, only two studies have experimentally compared within-subjects with between-subjects designs with inconclusive results regarding the impact of

11 One important reason for this inconclusive state of research regarding the predictive validity FSEs might be that the reported validation studies rely on very different research designs, including within-person comparisons (e.g. Pager & Quillian, 2005), natural experiments (e.g. Hainmueller et al., 2015) and experimental designs (e.g. Petzold & Wolbring, 2019). Obviously, the limitations of a design for the estimation of a behavioural benchmark can result in biased estimates and undermine the validation strategy. Further, the measurement of the outcome, the sampling strategy of the FSE, and the behavioural benchmark differ in many of these validation studies, which might undermine comparability (see Petzold & Wolbring, 2019).
within-subjects designs in terms of social desirability (see Auspurg et al., 2015; Walzenbach, 2019). Given the small body of methodological research, we recommend conducting pre-tests to assess the sensitivity of a research topic or of an FSE dimension instead of relying on the general claim that an FSE is better equipped to address sensitive questions.

In addition, one might also suspect that FSEs have higher predictive power if a respondent is familiar with the described situation and if the scenario resembles the actual decision-making process in real life (Hainmueller et al., 2015). As a consequence, one might assume that moving decisions are well evaluated by respondents, who seriously consider or prepare an actual move. Planned behaviour may correspond well with actual behaviour in such an instance. In contrast, some survey participants might not even have in mind in which situation they perceive jaywalking to be acceptable. This may explain why past research has concluded that the same dimensions for the intention to move predict actual moving behaviour (e.g. Nisic & Auspurg, 2009), while predictive power is rather low in the case of jaywalking (Eifler, 2007). These considerations are speculative, but they illustrate that future research should focus more on the predictive validity of FSEs and the development of a theory specifying the conditions under which FSEs are informative about determinants of actual behaviour. To this end, more theory-driven validation studies appear promising, with systematic variation of the discussed factors.

Conclusions

This paper provides a literature review about the use of FSEs in the social sciences (1982–2018). Our literature review shows that the field of FSEs has developed rapidly since the mid-2000s. They are increasingly being applied in different research areas such as crime, care and health, work, and among scholars from different countries, in particular from the US, Germany, the Netherlands, and the UK. Approximately half of recent studies have relied on non-probability samples (such as convenience, referral, and purposive samples; and samples from the crowdsourcing platform Amazon Mechanical Turk), raising questions about both the generalisability of results and the use of significance testing. Most recent studies have depended on within-subjects designs, and almost all have used state-of-the-art techniques to analyse such clustered data. In contrast, more recent advances in procedures for sampling vignette sets from a large vignette universe, such as D-optimal and RBCF designs, have not entered applied research to the same extent. While these techniques help to design FSEs in an order to avoid the confounding of main and interaction effects, and to optimise statistical power, they require additional expertise, specialised software, and time investment. Nonetheless, we especially recommend making extra investments in the case of small samples and
a large vignette universe, while the use of random sampling techniques still leads to inefficiencies and untestable assumptions, but might be acceptable in the case of very large sample from the vignette universe.

Several methodological questions remain unresolved concerning the realism and complexity of vignettes, social desirability, and the predictive validity of FSEs with respect to human behaviour. Regarding the complexity and realism of vignettes, we focused on the number of dimensions in an FSE and highlighted that simple scenarios may lead to boredom and fatigue effects, while very detailed scenarios may seem realistic but cause cognitive overload among respondents. However, a ‘one-size-fits-all’ rule regarding the complexity of vignettes and the number of vignette dimensions does not exist and is unlikely to emerge in the future. Thus, the design of factorial surveys should rely on theoretical considerations, and not just on considerations related to technical aspects of the experimental design. For example, researchers need to take into account the individual background, motivation and cognitive skills of their respondents as well as peculiarities of their research topic. Furthermore, the complexity of a vignette design and the related cognitive load depend not only on the number of dimensions (ratings), but also on other design elements, such as the measurement of the outcome and the vignette presentation style (e.g. Sauer et al., 2020).

In a similar vein, some researchers have used video vignettes to present scenarios. Audio-visual stimuli seem very promising and have the potential to increase the realism of vignettes substantially. Nevertheless, researchers should be aware that conducting video vignettes is demanding and may introduce new methodological pitfalls, such as the confounding of vignette dimensions with the (non)verbal expressions of the actors. Video vignettes may also be prone to other well-known methodological aspects (such as social desirability) due to the salience of certain vignette dimensions (see Ceuterick et al., 2020). Hence, researchers should carefully consider which presentation style seems most adequate. Instead of a ‘one-size-fits-all’ rule, we recommend relying on theoretical considerations and pre-tests to assess the various FSE design aspects. For example, theory can help to identify potential interactions between the research topic, the number of ratings per person, and the number of dimensions. Moreover, participants might be differently affected depending upon their motivation to take part in the survey, their cognitive skills, previous experience, and familiarity with the described situation (see Sauer et al., 2011; Teti et al., 2016).

Further, our review shows that FSEs are not only increasingly being used to study attitudes, but also to explore the determinants of behaviour, and in each observed time period, approximately 45% of all FSE studies focus on behaviour as an outcome. We indicated that the current state of research is inconclusive and raises several methodological and theoretical challenges for future validation studies. These questions illustrate that future research should aim to integrate previous
research and to formulate a theory that specifies the conditions under which FSEs are informative about the determinants of actual behaviour. As long as such theory does not exist, it appears neither warranted to reject FSEs to generate insights into determinants of behaviour, nor to apply them uncritically. When making inferences from stated intentions in FSEs to actual behaviour in the real world, potential differences need to be considered, such as the possibility of an intention-behaviour gap, respondents’ lack of familiarity with the decision situation, and biases due to social desirability.

Finally, there is a need for better documentation and reporting standards to assess the methodological aspects of FSEs. In a substantial number of publications, key information about the FSE design was hard to find, buried in online appendices, or not reported at all. In particular, it was alarming that an increasing number of recent publications did not contain information about how the vignettes were sampled from the universe. The fact that we could not retrieve this information, even after an extensive search, is alarming and indicates the need to establish clear documentation and reporting standards for FSEs. This includes all methodological aspects that are necessary to assess the quality of an instrument and to conduct replications and follow-up studies.

References


Ceuterick, M., Bracke, P., van Canegem, T., & Buffel, V. (2020). Assessing Provider Bias in General Practitioners’ Assessment and Referral of Depressive Patients with Different


Drewniak, D., Krones, T., Sauer, C., & Wild, V. (2016). The Influence of Patients’ Immigration Background and Residence Permit Status on Treatment Decisions in Health Care: Results of a Factorial Survey among General Practitioners in Switzerland. *Social Science & Medicine, 161*, 64–73. https://doi.org/10.1016/j.socscimed.2016.05.039


