sciendo
Swiss Journal of Sociology, 48, 2022, 47–75

Factorial Survey Experiments in the Sociology of Education. Potentials, Pitfalls, Evaluation

Knut Petzold*

Abstract: The potentials and pitfalls of factorial survey experiments (FSE) are discussed for empirical tests of theoretical explanations in the sociology of education. The possibilities and limits of FSE are outlined in relation to the internal validity, construct validity, and external validity of the obtained results and illustrated using an example experiment on the decision of university students to study abroad. It is demonstrated that FSE are an enriching complement to laboratory and field experiments, and observational studies.

Keywords: Factorial survey experiment, validity, educational decisions, study abroad

Der Faktorielle Survey in der Bildungssoziologie. Potenziale, Fallstricke, Evaluation

Zusammenfassung: Die Potenziale und Fallstricke von faktoriellen Surveyexperimenten (FSE) werden für empirische Tests von theoretischen Erklärungen in der Bildungssoziologie diskutiert. Die Möglichkeiten und Grenzen von FSE werden in Bezug auf die interne Validität, die externe Validität und die Konstruktvalidität der Ergebnisse erörtert und anhand eines Beispielexperiments zur Entscheidung von Universitätsstudierenden zum Auslandsstudium illustriert. Es wird gezeigt, dass FSE eine bereichernde Ergänzung zu Labor- und Feldexperimenten, s owie zu Beobachtungsstudien (observational studies) sind.

Schlüsselwörter: Faktorielles Surveyexperiment, Validität, Bildungsentscheidungen, Auslandsstudium

L'expérience d'enquête factorielle en sociologie de l'éducation. Potentiels, écueils, évaluation

Resumé: Les potentiels et les pièges des expériences d'enquête factorielle (FSE) sont discutés pour des tests empiriques d'explications théoriques en sociologie de l'éducation. Les possibilités et les limites du FSE sont discutées en relation avec la validité interne, la validité externe et la validité de construction des résultats et illustrées à l'aide d'un exemple d'expérience sur la décision d'étudiants universitaires d'étudier à l'étranger. Il est démontré que les FSE sont un ajout enrichissant aux expériences en laboratoire et sur le terrain ainsi qu'aux études d'observation.

Mot-clés : Expérience d'enquête factorielle, validité, décisions éducatives, étudier à l'étranger

* Hochschule Zittau/Görlitz – University of Applied Sciences, D-02826 Görlitz, knut.petzold@hszg.de

Seismo
DOI 10.2478/sjs-2022-0001

1 Introduction[1]

In contemporary sociology of education, structures on the macro-level of a society, e. g., educational inequality, are explained through individual educational decisions on the micro-level (Becker and Solga 2012). When actors make educational decisions, causal effects of subjective goals, individual resources, and situational restrictions are usually assumed. Empirical tests on the mechanisms underlying educational decision-making are typically based on conventional survey data. When using such "observational data" (Rosenbaum 2010), endogeneity problems lead to uncertainty in causal conclusions (Rubin 2008). Accordingly, observational data often impede providing resilient answers to causal questions in the sociology of education (Zangger and Becker 2019).

In contrast, randomised experiments are considered the gold standard for identifying causal effects and are increasingly prominent in the social sciences (Jackson and Cox 2013). Due to the random assignment of subjects to systematically manipulated treatments, differences in measured outcomes can be causally attributed to the treatment status (Campbell and Stanley 1963; Shadish et al. 2002; Elwert and Winship 2014). Accordingly, experiments are ideal for empirical tests of theoretical models in the sociology of education. However, laboratory or field experiments were often considered too costly, ethically questionable, and practically impossible in the past (Cook 2001) so that experiments are rarely used in the sociology of education so far (Zangger and Becker 2019).

At this point, survey experiments offer great potential as the experimental approach is implemented in survey research. In sociology, especially factorial survey experiments (FSE) are particularly popular (Rossi and Andersen 1982; Jasso 2006; Auspurg and Hinz 2015). In FSE, unique hypothetical descriptions (vignettes) of decision-making problems are varied along a multi-factorial experimental design (dimensions with levels) and randomly presented to the respondents for assessment. FSE permit detailed and rigorous empirical tests of the predicted effects in educational decision-making and therefore represent an attractive complement to field and laboratory experiments on the one hand and conventional survey data on the other.

Accordingly, applications of FSE in the sociology of education have increased sharply in recent years. Decision-making on school choices (Thelin and Niedomysl 2015; Keller 2018), a desired apprenticeship (Möser et al. 2019), or on starting a course of studies (Finger 2016) has been examined. FSE are especially used to investigate employer preferences regarding the educational characteristics of potential employees (McDonald 2019), while the match between job requirements and applicants' skills (De Wolf and Van Der Velden 2001), formal national and international

1 I woud like to thank three anonymous reviewers and the editors for their valuable comments on an earlier version of this paper. Parts of this work were supported by the Federal Ministry of Education and Research, Germany (grant number 01PW11013).

certificates (Di Stasio 2014), periods of unemployment (Shi et al. 2018), or higher education dropout (Daniel et al. 2019) were of particular interest.

Given this rapidly increasing attention, the method is discussed regarding the potentials, pitfalls, and evaluation strategies in empirical tests in the sociology of education in this article. To this end, some general considerations on theoretical models in the sociology of education are outlined, the potentials and pitfalls of empirical tests with FSE are discussed, and are then illustrated using an application from tertiary education. An FSE on decision-making to study abroad was conducted at a German university and replicated at a Chinese university. The article closes with a brief conclusion on possible directions of future research with FSE in the sociology of education.

2 Theory and Empirical Tests in the Sociology of Education

Numerous approaches of modern educational sociology can be added to methodological individualism (Coleman 1990; Esser 1999). According to this, the contextual factors at the collective level of a society affect incentives in decision-making and behaviour at the individual level. Individual behaviour at the micro-level is then aggregated to the macro level.[2] In this process, collective structures such as educational inequality are reproduced or changed (Grusky 1994; Becker and Solga 2012). Yet, primarily, the central determinants of individual actions in a population are in focus (Hedström and Swedberg 1996; Buskens et al. 2014).

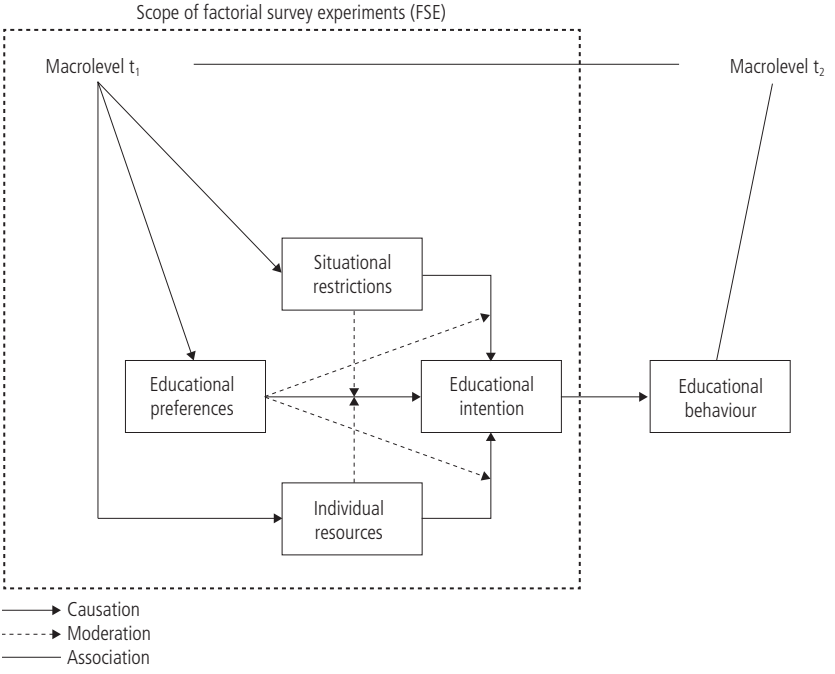
In the sociology of education, micro-theories are often based on variants of the rational choice theory (Boudon 1974; Erikson and Jonsson 1996; Breen and Goldthorpe 1997; Esser 1999: Becker 2000; Becker and Hecken 2009; Stocké 2010), which can be seen as the core theory of methodological individualism (Opp 1999). In this approach, actors have *preferences* (subjective motives, goals) and can choose to perform certain educational behaviours. The educational alternative that best satisfies their preferences will be realised, while considering the present opportunities and constraints.

Constraints are sometimes further distinguished (Rössel 2009; Erlinghagen and Hank 2018). Individual *resources* can be acquired over time and have a more general character across different situations, for instance, income, education, or social contacts. Situational *restrictions,* on the other hand, represent incentives that an actor finds in a situation that cannot be directly controlled. These include, for example, prices, time, or legal restrictions. In applications of FSE, this distinction is helpful, because resources can be measured and restrictions manipulated. Further, since real

2 The process of aggregation by no means only includes the simple summation of the individual actions, but can be based on diverse, for example non-linear, transformational mechanisms that are to be determined depending on the object under consideration (Hedström and Ylikoski 2010).

education-related behaviour cannot be captured in FSE, it is helpful to refer to the theory of planned behaviour (Ajzen 1991). In this approach, an *intention* is formed, which can result in behaviour. Due to restrictions, however, not every intention is translated into *behaviour*. In this perspective, educational decision-making occurs in situations where educational goals, individual resources, and perceived situational restrictions shape an intention that results in educational behaviours (see Figure 1).

Figure 1 General Model of Explanations in Sociology of Education



Empirical tests of area-specific theories of education require additional auxiliary assumptions (Trafimow 2012) in terms of measurement and causality. The better empirical translations correspond with the theoretical concepts and assumed causality, the more valid are the conclusions. Yet, any aspect can be challenged when testing a theory with regard to the extent to which alternative explanations are excluded. As a result, no final judgment on validity is possible. Evaluations of validity must therefore account for the specific research design that was applied (Goldthorpe 2001).

Originally introduced by Campbell (1957) and Campbell and Stanley (1963), the concept of "internal validity" refers to the question whether observed covariation between two variables (i. e. the presumed treatment and the presumed outcome)

reflects a causal relationship, taking the data generating process into account. The assessment of auxiliary assumptions further links to the concept of "construct validity" that refers to the degree to which inferences are warranted from observed study particulars (i. e. persons, settings, relationships) to the underlying constructs that are to be represented. "External validity", finally, addresses the question whether a revealed cause-effect relationship holds over different persons, settings, treatment variables, and outcomes (Shadish et al. 2002, 38).

Ideally, empirical tests ensure high internal validity, high construct validity, and high external validity. In this regard, FSE provide several potentials but also pitfalls, compared to other methodological approaches. We discuss this in the following section, the pros and cons of FSE, in comparison with laboratory experiments (Webster and Sell 2007), field experiments (Gerber and Green 2012), natural experiments (Dunning 2012), and observational studies (Rosenbaum 2010).

Overall, like laboratory and field experiments, empirical tests with FSE guarantee a higher internal validity than tests with observational studies. Potential replications and heterogeneous samples also facilitate the assessment of the external validity. At the same time, the theory-based data collection permits a direct test strategy and reduced social desirability bias ensures a high degree of construct validity. Accordingly, FSE are a useful addition to laboratory, field, and natural experiments on the one hand and observational data on the other. Table 1 summarises the methodological considerations that are further elaborated in the next sections.

Table 1 Comparison of Methods

	Factorial survey experiment	Laboratory experiment	Field experiment	Natural experiment	Observational studies
Measurement					
Subjective preferences	Yes	Yes	No	(Yes)	Yes
Personal resources	Yes	Yes	Only observables	(Yes)	Yes
Situational restrictions	Yes	Yes	Yes	(No)	(No)
Educational behaviour	No	No	Yes	Yes	Yes
Design					
Treatment manipulation	Yes	Yes	Yes	No	No
Random assignment	Yes	Yes	Weak	No	No
Controlled (random) sampling	Simple	Difficult	No / Difficult	No	Typical (cross sec.) / Initial (panel)
Simple replicability	Yes	Yes	(Yes)	No	No

2.1 Potentials of Factorial Survey Experiments in Empirical Tests

2.1.1 Direct Test Strategy

When testing hypotheses about (educational) decision-making, data collection is ideally structured by the theoretical model. The more detailed elements of a theoretical model are tested the less untested additional assumptions must be made. According to this, a "direct" and an "indirect" test strategy is sometimes distinguished (Brüderl 2004). An indirect test strategy is an instrumentalist approach, testing only behavioural implications, while determinants are typically approximated through the social context (Lindenberg 1996). However, correct conclusions (e. g. observed educational behaviour) can be deduced also from false premises (e. g. underlying goals, resources and restrictions) and many different correct premises can imply a conclusion. Thus, one will not know what the correct premises are when a conclusion is confirmed. By contrast, in the direct test strategy, not only are behavioural implications tested but also underlying preferences and restrictions (Becker and Hecken 2009).

When using FSE, data collection is theory-guided. All conceivable restrictions in a situation can be manipulated experimentally, while the respondents' subjective goals, individual resources, and context variables can be measured. Situational treatments and the respondents' characteristics are not confounded and even effects on different outcomes for the same vignette can be compared, i.e., intentions regarding educational alternatives. In this way, theoretical models in decision-making can be tested directly, so that fewer assumptions have to be made. FSE are therefore predestined to apply a direct test strategy (Brüderl 2004, 178).

The high level of flexibility in construction is shared with laboratory experiments. However, it is usually not possible to measure individual goals and unobservable resources in field experiments. Since researchers have no control over the treatment, outcome, and setting, natural experiments are generally less suitable for strictly theory-based empirical tests. Observational data are often collected routinely in standardised surveys without specific theoretical guidance, often leading to missing data problems. Therefore, observational data permit almost inevitably only an indirect test strategy, while there is a particular risk of "variable sociology", in which micro-theoretical assumptions are tested using only social context variables (Esser 1996; Goldthorpe 2001). In field experiments and observational studies, realised educational behaviour is under study so that it remains unclear which alternative opportunities actors have considered. In these regards, FSE can be seen as superior to field experiments, natural experiments, and observational studies.

2.1.2 Causality

Referring to the counterfactual approach to causality (Pearl 2010; Morgan and Winship 2015), the difference between two outcomes for one unit under investigation, one in the treatment state and one in the non-treatment state, reflects the individual causal effect. The "fundamental problem of causal inference" (Holland 1986, 947)

is, however, that outcomes of one unit cannot be observed in both treatment states at one point in time. Therefore, potential outcomes must be estimated using comparative cases from a control group, while the units of both groups are identical except in their treatment status. If this conditional independence assumption (CIA) regarding the treatment state is fulfilled, the variance of the outcome can be traced back to the exogenous treatment (Jackson and Cox 2013; Elwert and Winship 2014).

FSE ensure a high internal validity when causal hypotheses are tested (Mutz 2011; Auspurg and Hinz 2015). Multi-factorial designs show orthogonal distributions across all levels so that the treatments are not correlated with one another (Dülmer 2007; 2016). The random assignment of vignettes ensures that the treatments are not confounded with the respondents' characteristics. Manipulation and randomisation satisfy the CIA. The concept of causation implies a process in time (Goldthorpe 2001; Shadish et al. 2002), which is reflected by the sequence of treatment and outcome.

Randomisation and manipulation are also used in laboratory and field experiments, but not in natural experiments. However, relevant treatments of theoretical interest (e. g. social origin) can often not be manipulated and alternative educational opportunities are difficult to investigate in field experiments (Cook 2001). For practical or ethical reasons, randomisation can often not be carried out in educational research (e. g. pupils to schools) or is undermined by self-selective processes of subjects. Particularly in field experiments, there is often no full control over the randomised assignment, so that the CIA may be at risk and actual treatment exposure must be considered (Zangger and Becker 2019). In contrast, FSE permit randomisation, manipulation, and examining all treatments and outcomes regardless of real-world restrictions.

Observational studies are even more limited when it comes to internal validity, as there is neither manipulation nor randomisation. The sequential order between cause and effect is not adequately captured in the case of cross-sectional data. Statistical relationships between the suspected cause and suspected effect can be conditioned only for observed heterogeneity so that the CIA remains rather strong. If panel data are used, the sequential order between cause and effect is secured and effects can be conditioned for time-constant unobserved heterogeneity. However, this comes at the expense of the effects of time-constant respondent characteristics not being estimated without additional assumptions (Brüderl and Ludwig 2015), a central disadvantage when testing models on decision-making.

2.1.3 Controlled Respondent Sampling and Replicability

In the sociology of education, theories typically refer to specific populations, such as (university) students or employers so that empirical tests must of course rely on subjects who represent this target population (Stroebe et al. 2018, 388). However, the question of external validity might be especially meaningful and important for the sociology of education, because it might be of interest whether experimental

findings of desirable intervention effects promise benefits also for other groups or on a larger scale (Mook 1983, 380). Heterogeneous sampling and replications can help to assess the external validity. Yet, researchers will not *make* generalisations but only test them.[3]

Through simple replicability, FSE facilitate theory-testing in different settings, with different treatment and outcome variables, with different samples or in diverse geographical locations (Mutz 2011; Auspurg and Hinz 2015). FSE can be carried out with convenient respondent samples or with random samples drawn from a population (Wallander 2009; Sauer et al. 2011). Heterogeneous sampling offers the potential for analyses of subgroups and interaction effects.

Compared to FSE, replications of laboratory and field experiments with different samples or even in other regions are far more complicated. Since subjects from very specific populations (e. g. entrepreneurs) or very different groups are difficult to recruit, laboratory experiments are usually carried out with very homogeneous (student) samples so that the knowledge gained is sometimes questioned (Levitt and List 2007). In field experiments, compared to FSE, it is also very difficult to carry out a controlled selection of study participants from a defined population. Observational studies are typically based on total populations or drawn probability samples, though panel data may be limited by panel mortality. In contrast to experiments, indefinite replications are not possible.

2.1.4 Social Desirability Bias

Finally, FSE is sometimes reported to be strengthening construct validity by systematically suppressing a possible social desirability bias (SDRB, Krumpal 2013). The indirect approach and the high degree of realism permit unobtrusive measurements of sensitive information (Alexander and Becker 1978; Auspurg et al. 2015; Walzenbach 2019). Compared to other methods, FSE induced less socially desirable response behaviour (Armacost et al. 1991), especially when third persons are described in vignettes (Finch 1987). One reason might be the lack of direct interaction with the experimenters (Mutz 2011). To a small extent, however, SDRB can also occur (Collet and Childs 2011; Markovsky and Eriksson 2012), because survey methods are reactive. However, as long as the subjects are aware that they are currently participating in an investigation, socially desired responses are possible in all study designs. Unobtrusive field experiments are an exception, in which true behaviour in real-world situations can be measured non-reactively.

3 The initial conceptualisation of external validity (Campbell 1957, 297; Campbell and Stanley 1963, 5) is often erroneously interpreted as implying that general hypotheses can be derived from empirical observations. From a deductive Popperian (1959) viewpoint, it is impossible to logically prove that a cause-effect relation will also hold in a different population or situation based on singular observations. Instead, the domain of applicability of a hypothesis is specified by the theory and the aim is to test whether a predicted effect actually occurs in an experiment, not to generalise. Accordingly, "diversification of subject populations does not make experimental findings more externally valid" (Stroebe et al. 2018, 387).

2.2 Pitfalls of Factorial Survey Experiments in Empirical Tests

2.2.1 Design Violations

Pitfalls in the application of FSE can typically result from their special construction. Any violations of the balanced and orthogonal experimental design or of random assignment are threats to internal validity, as with laboratory and field experiments. Such disturbances can result for instance from the exclusion of illogical cases in level combinations, from a biased vignette sample, by a selective item or unit non-response, or by analysing small subgroups.

Several issues of internal validity concern the stable unit treatment value assumption (SUTVA), which is fulfilled if there is no inference between units "[…] leading to different outcomes depending on the treatments other units received and there are no versions of treatments leading to 'technical errors' […]" (Rubin 1980, 591). Originally, the SUTVA referred to inferences between the units of the experimental and control group. Therefore, violations of the SUTVA can occur in all types of experiments, especially in field experiments. When applying FSE, however, respondents are repeatedly assigned to several vignettes and thus, technically, to different experimental groups. The presentation order of the vignettes may evoke carry-over effects, learning effects, and fatigue effects so that treatments of previous vignettes affect outcomes regarding subsequent vignettes. Yet, in a recent study, vignettes in a random order produced results similar to those presented in a fixed order (Sauer et al. 2020).

Further, "technical errors" concerning SUTVA may be especially reflected by the following problems in data collection. Certain combinations of dimension levels can sometimes be illogical or implausible so that the dimensions are not taken seriously by the respondents (Auspurg et al. 2009). Single dimensions can be overlooked, especially when presented in a text format, rather than in a tabular form (Shamon et al. 2019). Dimensions that vary on a particularly large number of levels also attract more attention-biasing responses, known as number of levels effects (Verlegh et al. 2002). The presentation order of dimensions can lead to positional effects, where primarily the first (primacy effect) or the last (recency effect) dimension is considered. Such biases occur, however, especially with the low complexity of the vignettes (Auspurg and Jäckle 2017) or if respondents tend to give quick answers (Düval and Hinz 2020).

2.2.2 Response Inconsistency

Answering vignettes is cognitively more demanding than answering conventional question types. As a result, response patterns can occur, especially from older and less-educated respondents (Auspurg and Hinz 2015), in particular if the vignettes are very complex and the respondents are not familiar with the subject of the survey (Sauer et al. 2011). One challenge for construct validity is a potential violation of the information equivalence between the vignettes of a FSE (Dafoe et al. 2018). Certain

levels may activate different background beliefs so that comparability across levels is limited. Regarding response scales measurement equivalence must be assumed, i. e. responses obtained reflect respondents' true values. However, if the respondents are inconclusive regarding where to put their answer value, ceiling effects and censoring effects can occur (Jasso 2006). When different respondents report the same true value on a rating scale in different ways, this is called differential item functioning (King et al. 2004). Compared to occasionally recommended types of open response scales, however, rating scales still have fewer problems, such as item non-response (Sauer et al. 2020). Replications can be particularly restricted by limited survey equivalence, especially when FSE are conducted in different cultural contexts (Harkness 1998), i. e., if vignette content or response scales are understood differently. Yet again, as with other survey designs, violations of information equivalence, measurement equivalence, and survey equivalence cannot be completely ruled out when applying FSE (Eifler and Petzold 2019).

2.2.3 Hypothetical Bias

The main objection to construct validity when using FSE is, however, that both the treatments and the outcomes are only hypothetical. The complexity of a decision problem can never be fully simulated in vignettes (Hughes and Huby 2004). As a result, the decision-making conditions may be perceived differently than in real situations (Collett and Childs 2011). Since the answers do not have any consequences for the respondents, there may not be sufficient incentive compatibility (Friedman and Cassar 2004). As a result, a "hypothetical bias" may occur if preferences and perceptions of restrictions are not adequately activated when administering decision situations in surveys (Ajzen et al. 2004). An important methodological challenge in the construction of FSE is therefore to reduce the hypothetical bias as much as possible, by adequately describing the relevant vignette dimensions and relevant outcomes.

2.2.4 Behavioural Validity

Further, when testing theoretical models on educational decision-making with FSE, the question of behavioural validity arises (Petzold and Wolbring 2018; Eifler and Petzold 2019). Evidence regarding the correspondence between results obtained by FSE and results of studies that serve as behavioural benchmarks is mixed and further research is needed (for an overview, see: Petzold and Wolbring 2019). Although there is no validation study from educational research so far, in some studies, the results of FSE correspond largely with real behaviour (Hainmueller et al. 2015) or at least treatment effects could be replicated (Petzold and Wolbring 2018), while other studies report clear deviations (Pager and Quillian 2005). At this point, field experiments, natural experiments, and observational studies have a clear advantage over FSE, as they permit investigating real education-related behaviour in real-world situations.

3 Example: A Factorial Survey Experiment on the Decision to Study Abroad

Potentials, pitfalls, and evaluations are demonstrated by an example application on decision-making in tertiary education. In an FSE, the conditions of students' intentions to spend a period abroad are examined both, originally at a German university and replicated at a Chinese university. The FSE was part of the project MOHSL-Mobility of High Skilled Labour funded by the Federal Ministry of Education and Research (BMBF). The state of the art, theoretical arguments, and substantial results have been outlined elsewhere (Petzold and Moog 2018; Petzold 2018). Therefore, the focus is on methodological aspects.

3.1 Motivation

In the context of globalisation processes, the question arises under which conditions students decide to study abroad. In literature, far-reaching theoretical assumptions are often made about institutional restrictions and perceived benefits (Salisbury et al. 2009). However, empirical tests are usually indirect and mostly based on observational studies such as student and graduate surveys (Lörz et al. 2016). While the assessment of external validity is mostly acceptable, there are concerns about internal and construct validity. Mobile and non-mobile students are often compared, while subjective goals and situational restrictions are not or roughly operationalised. Institutional conditions are typically confounded with each other and with the characteristics of the students.

By contrast, FSE permit a detailed and direct test of theoretical assumptions. Economic, organisational, and social restrictions can be varied in vignettes, while expected returns can either be experimentally alternated or be measured at the respondents' level. It is possible to determine the relative weights that are attributed to the situational restrictions, individual resources, and expected benefits in the subjective reasoning of students considering study abroad. The findings are more detailed and the causal effects more trustworthy than in studies based on observational data.

3.2 Experimental Design

Researchers are limited in the number of possible dimensions for two reasons. First, to avoid inconsistent answers from too little information (boredom) or too much information (fatigue), it is recommended to vary approximately seven dimensions (Sauer et al. 2011; Auspurg and Hinz 2015). Second, the number of possible combinations increases exponentially. The larger the full factorial design (also called vignette universe) is, the more difficult it becomes to maintain orthogonality and balance when a selection of vignettes is made.

Based on the theory of rational choice (Opp 1999) and the theory of planned behaviour (Ajzen 1991), the specific treatment selection built up results of previous studies. Combining both approaches allows integration of subjective goals, individual

resources, situational restrictions, and subjective and personal norms as determinants of the intention to study abroad. Table 2 shows the varied dimensions and levels. A lower intention to study abroad is hypothesised with lower levels, which reflect higher costs or lower benefits, while a stronger intention is expected with higher levels, which reflect fewer restrictions or higher benefits.

Table 2 Vignette Dimensions

	Dimensions	Levels			Total
		1	2	3	
1	Exchange program	No program	Program		2
2	Financial scholarship	No scholarship	Scholarship		2
3	Exchange in group	Alone	In group		2
4	Related language skills	No skills	Elementary skills	Good skills	3
5	Reputation host university	Poorer than home university	Equal to home university	Better than home university	3
6	Host country preference	Not desired	Less desired	Strongly desired	3
7	Family's / friends' expectations	Expect to stay	Show no expectations	Expect study abroad	3
	Vignette universe (Cartesian product of all levels of all dimensions)				648

The decision to study abroad depends on many other conditions, which, however, cannot all be varied for the reasons mentioned above. In order to keep the information provided equivalent for all respondents, it is recommended to fix relevant but not alternating situational conditions as constant in the vignette introduction. For example, in this study, all vignettes refer to a fixed period of one semester abroad. One must account also for the possibility of illogical or implausible combinations of levels. Excluding them from the full factorial only afterwards is a distortion of balance and orthogonality. In this study, no combinations were excluded so that the Cartesian product of all dimensions and levels results in $2 * 2 * 2 * 3 * 3 * 3 * 3 = 648$ decision situations. Since these cannot be presented entirely, a selection must be made.

3.3 Vignette Sample

The desired orthogonality and level balance of the full experimental design may be biased when vignettes are sampled. The vignette sample is drawn either randomly (random design) or systematically (optimal design). Random designs are sufficient in many cases but imply the risk of violations, which should be checked. In contrast, with optimal designs, the quality of the reduced experimental design is determined by the D-efficiency, a standardised measure for the sampling bias, ranging from 0 (bias) to 100 (no bias) (Dülmer 2007).

In the example experiment, the vignette sample was drawn using the modified Federov search algorithm (Kuhfeld et al. 1994) with a D-efficiency of D = 97.05[4] and divided into fifteen decks, with eight vignettes each. Accordingly fifteen questionnaire versions were designed and every respondent expressed the intention to study abroad towards eight hypothetical situations.

3.4 Vignette Presentation and Response Scale

Vignette descriptions are usually presented as running text or in tabular form (Shamon et al. 2019; Sauer et al. 2020). Figure 2 shows the vignette texts. In the original study, the vignette order was not randomised due to restrictions from the self-administered paper and pencil questionnaire so that the vignette position should be controlled for in statistical analysis. In the replication study, vignette order could easily be randomised in the web survey.

Figure 2 Vignette Text

There is *an* / *no* exchange program at your university and you *have the* / *have no* opportunity to solicit a financial scholarship. You will study abroad *alone* / *with a group of fellows* in a country, which language you speak *not at all* / *elementary* / *well*. The host country holds *the first* / *the last* / *no position* on your personal favourite list, while the host university has a *better* / *equal* / *poorer* reputation than your home university. Your family and your friends would find it *bad* / *neither bad nor good* / *good* if you study abroad.

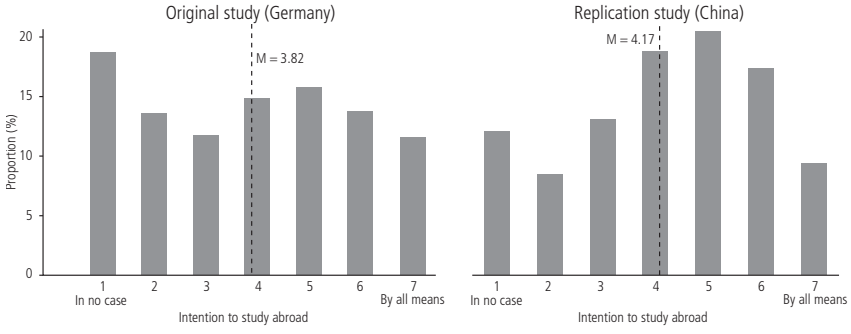
I intent to study abroad given these conditions

In no case ○ ○ ○ ○ ○ ○ ○ By all means

Rating scales are typically used with FSE (Wallander 2009). In the example, the queried intention to study abroad was expressed on a seven-point rating scale ranging from "in no case" to "by all means". Figure 3 shows the distributions of the measured outcomes for the original study and the replication study. There was a clear variance in both studies, indicating that the dimensions presented in the vignettes contained relevant information.

4 Resolution V design, where all the main effects and two-factor interactions are estimable free of each other (Kuhfeld et al. 1994, 546).

Figure 3 Distributions of the Intention to Study Abroad (Original Study and Replication Study)



3.5 Measurements of Expected Returns

To investigate whether respondents differ in their assessment of a period abroad regardless of the situational conditions, two scales for frequently mentioned expected returns are integrated. The beliefs in personality development and increasing job market chances were measured with three items each and sum scores were calculated based on consistency analyses. The effects of the scales on the outcome can thus be estimated separate from treatment effects but also interaction effects between situational restrictions and expected returns can be revealed.

3.6 Data Collection

The data of the original experiment were collected in 2012 from students of economics and engineering science of the University of Siegen. These disciplines show the highest and the lowest participation rates in international mobility in Germany. A standardised paper and pencil questionnaire was used on-site in classrooms. A total of 370 questionnaires were issued, of which 304 were reasonable to work after data cleaning. A replication at Northeastern University in Shenyang, PR China, was implemented as an online survey, which enabled recruitment via an invitation link sent by email and automatic randomisation. 147 students took part, of whom four were excluded due to item non-response. However, as the invitations were sent through the internal email lists of the departments, the total number of invitations is unknown.

3.7 Results

Each respondent expressed intentions regarding eight vignettes. The resulting hierarchical data structure (Hox et al. 1991; Jasso 2006) is captured by using multi-level models (Snijders and Bosker 2012). This allows for simultaneous estimations of treatment effects and effects of respondent characteristics. To evaluate the robustness of the findings at the vignette level, additional models with fixed effects for the respondents are also estimated (see section 4.1). Effect heterogeneity across certain experimental dimensions or respondent groups can be determined by modelling interactions. For this, multiplicative terms between vignette dimensions (level 1 interaction) or between a vignette dimension and a respondent characteristic (cross-level interaction) are included in a model (Auspurg and Hinz 2015, 96–97).

Table 3 shows the exemplary results. Models 1 to 5 are based on the original study and model 6 on the replication study. Model 1 contains all the vignette dimensions but no respondent characteristics. Due to the experimental design, the effects are additive and can be interpreted independently of one another. If the experiment was successful, the regression coefficients represent the causal effects of the treatments on the educational intention, i. e. to study abroad.

Model 2 is expanded by the respondents' characteristics. In contrast to level 1 based on experimental data, these level 2 coefficients cannot be interpreted causally, since the respondents' characteristics represent observational data. As with regressions based on cross-sectional data, the trustworthiness of such an effect must be assessed according to the selected control variables, which are used for conditioning.

Model 3 contains an additional multiplicative term that exemplarily models the interaction between two vignette treatments at level 1. The coefficients indicate a mutual substitution of the scholarship and the country preference in the individual calculation. For both dimensions, conditional regression coefficients are provided in this model.

Model 4 integrates an exemplary cross-level interaction between a level 1 (scholarship) and a level 2 variable (return on personal development). The conditional treatment effect can still be interpreted causally, provided that the internal validity is not disturbed (see below). The conditional effect of the respondent's characteristic still reflects only a correlation. The coefficient of the interaction term suggests a promotion effect: the greater the expected return on personal development, the more important the scholarship is when considering study abroad, and *vice versa.*

Finally, model 6 contains the results of the replication study in China. Some effects are substantially stable across both samples, but not others. If effect heterogeneity was predicted by theory, hypotheses would receive support. If effect heterogeneity was not predicted, the theory would need refinement.

Table 3 Regression Models on the Intention to Study Abroad (Original and Replication Study)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Vignette dimensions						
Exchange program (ref. no)	0.364***	0.364***	0.362***	0.363***	0.363***	0.081
Financial scholarship (ref. no)	0.925***	0.923***	1.069***	0.265	0.927***	0.508***
Exchange in group (ref. alone)	0.328***	0.329***	0.334***	0.332***	0.327***	0.114
Related language skills						
Elementary skills (ref. no skills)	0.548***	0.547***	0.546***	0.546***	0.551***	0.145
Good skills (ref. no skills)	1.295***	1.291***	1.278***	1.287***	1.298***	0.870***
Reputation host university						
Equal reputation (ref. poorer rep.)	0.567***	0.567***	0.565***	0.571***	0.569***	0.416***
Better reputation (ref. poorer rep.)	0.757***	0.759***	0.759***	0.762***	0.753***	0.759***
Host country preference						
Less desired (ref. not desired)	0.317***	0.320***	0.639**	0.326***	0.312***	0.609***
Strongly desired (ref. not desired)	1.309***	1.310***	1.671***	1.312***	1.305***	0.968***
Family's/friends' expectations						
Show no expectation (ref. expect stay)	0.316***	0.317***	0.311***	0.320***	0.314***	0.058
Expect study abroad (ref. expect stay)	0.412***	0.411***	0.411***	0.408***	0.411***	0.590***
Respondents' characteristics						
Belief in personality development		0.306***	0.305***	0.124		
Belief in job market benefits		0.070	0.070	0.070		
Study abroad experience (ref. no)		0.758***	0.757***	0.757***		
Age		–0.006	–0.006	–0.006		
Gender (ref. female)		–0.005	–0.004	–0.004		
In relationship (ref. single)		–0.304*	–0.302*	–0.305*		

Continuation of table 3 on the next page.

Continuation of table 3.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Economics (ref. engineering)		–0.037	–0.036	–0.036		
Master student (ref. bachelor student)		-0.014	–0.007	–0.013		
Interactions						
Country less desired * financial scholarship			–0.209			
Country strong desired * financial scholarship			–0.237+			
Bel. in personality dev. * financial scholarship				0.120+		
Constant	1.192***	-0.600	–1.740*	–0.534	1.196***	2.356***
σ_u	1.062	0.905	0.905	0.907	1.172	0.857
σ_e	1.384	1.384	1.384	1.383	1.384	1.369
ϱ	0.371	0.300	0.299	0.301	0.418	0.282
Wald χ^2 / F	834.84***	1185.29***	1219.75***	1221.87***	115.04***	181.635***
R^2	0.252	0.330	0.331	0.331	0.252	0.169
$R^2_{between}$	0.021	0.245	0.248	0.245	0.021	0.052
R^2_{within}	0.374	0.374	0.375	0.376	0.374	0.241
$N_{vignettes}$	2432	2432	2432	2432	2432	1144
$N_{respondents}$	304	304	304	304	304	143

Models 1–-4: Original study sample, random intercept fixed slopes regressions, b-coefficients, robust SE.
Model 5: Original study sample, fixed effects regression, b-coefficients.
Model 6: Replication study sample, random intercept fixed slopes regression, b-coefficients, robust SE.
+ $p<0.10$ *$p<0.05$, **$p<0.01$, ***$p<0.001$

4 Threats to Validity and Strategies of Evaluation

In this section, some typical threats to validity are highlighted and strategies of evaluation illustrated.

4.1 Internal Validity

4.1.1 Final Sample

To ensure internal validity, it must be evaluated whether the variation and randomisation in the final data set are still intact. If single vignettes are not judged by respondents, level balance and orthogonality can be violated, though vignettes from the original experimental design were perfectly assigned. All levels should be evenly distributed within the vignette dimensions and across levels of other vignette

Table 4 Vignette Dimensions: Distribution in Original Sample

Vignette dimensions (Sample)	N	%
Exchange program		
No program	1221	50.2
Program	1211	49.8
Financial scholarship		
No scholarship	1182	48.6
Scholarship	1250	51.4
Exchange in group		
Alone	1257	51.7
In group	1175	48.3
Related language skills		
No skills	825	33.9
Elementary skills	809	33.3
Good skills	798	32.8
Reputation host university		
Poorer than home university	839	34.5
Equal to home university	746	30.7
Better than home university	847	34.8
Host country preference		
Not desired	847	34.8
Less desired	785	32.3
Strongly desired	800	32.9
Family's / friends' expectations		
Expect to stay	795	32.7
Show no expectations	830	34.1
Expect study abroad	807	33.2
$N_{Vignettes}$	2432	100.0

Table 5 Vignette Dimensions: Correlations in Original Sample

Dimensions (sample)	1	2	3	4	5	6	7
1 Exchange program	1.000						
2 Financial scholarship	–0.067	1.000					
3 Exchange in group	0.025	–0.036	1.000				
4 Related language skills	0.037	0.033	0.068	1.000			
5 Reputation host university	0.006	–0.007	–0.028	–0.020	1.000		
6 Host country preference	0.053	–0.026	0.014	0.047	–0.007	1.000	
7 Family's / friends' expectations	–0.050	0.006	–0.021	0.009	–0.036	–0.003	1.000

Table 6 Vignette Dimensions: Distribution in Subgroup of Replication Sample

Vignette dimensions (Sample)	N	%
Exchange program		
No program	36	50.0
Program	36	50.0
Financial scholarship		
No scholarship	34	47.2
Scholarship	38	52.8
Exchange in group		
Alone	37	51.4
In group	35	48.6
Related language skills		
No skills	26	36.1
Elementary skills	22	30.6
Good skills	24	33.3
Reputation host university		
Poorer than home university	26	36.1
Equal to home university	23	31.9
Better than home university	23	31.9
Host country preference		
Not desired	25	34.7
Less desired	25	30.6
Strongly desired	25	34.7
Family's / friends' expectations		
Expect to stay	23	31.9
Show no expectations	24	33.3
Expect study abroad	25	34.7
$N_{Vignettes}$	72	100.0

Table 7 Vignette Dimensions: Correlations (r) in Subgroup of Replication Sample

	Dimensions (sample)	1	2	3	4	5	6	7
1	Exchange program	1.000						
2	Financial scholarship	–0.167	1.000					
3	Exchange in group	0.028	–0.138	1.000				
4	Related language skills	0.067	0.035	0.199	1.000			
5	Reputation host university	0.100	0.033	0.100	–0.080	1.000		
6	Host country preference	0.219	0.121	–0.086	0.120	–0.061	1.000	
7	Family's / friends' expectations	–0.170	–0.104	0.035	–0.121	–0.061	–0.102	1.000

dimensions, which is equal to a correlation of zero between all dimensions. In the final analysis sample, all levels within the dimensions are still approximately evenly distributed (Table 4) and all dimensions hardly correlate with one another (Table 5).

The quality of the randomisation can be easily evaluated using regression models. Since vignette and respondent characteristics should not be confused with one another, the treatment effects with and without control for respondent characteristics must be stable. In addition, the coefficients should not differ between multi-level models with random effects (RE) and fixed effects (FE) for the respondents' level (Hausman test), because unobserved heterogeneity controlled in the FE model should already be captured by randomisation (Wooldridge 2013, chapter 14). In the example (Table 3), there are only minimal deviations in the coefficients of the vignette dimensions between RE-models with and without covariates (model 1 and model 2). Also coefficients of RE-model 1 and FE-model 5 do not differ significantly (Hausman test: $\chi^2 = 12.45$; $p = 0.331$), which indicates successful randomisation.

4.1.2 Subgroup Analysis

Subgroup analyses can result in violations of the experimental design and randomisation, which is sometimes overlooked. This pitfall will be demonstrated by a constructed subgroup analysis: what treatment effects can be revealed for female engineering students from China? This subgroup comprises only 9 respondents, who assessed a total of 72 vignettes.

Though the levels within the single dimensions are equally distributed (Table 6), all vignette dimensions correlate more strongly than in the overall sample (Table 7). Accordingly, the experimental design on which this special subsample is based has no longer the required quality to justify a causal interpretation of the effects.

Randomisation is also affected by the small sub-sample size. As a comparison of the RE model with the FE model shows (Table 8), the vignette and respondent characteristics are confounded, as the coefficients differ significantly (Hausman test: $\chi^2 = 33.22$; $p < 0.000$). Regarding this subsample, internal validity is weak and the

Table 8 Regression Models on the Intention to Study Abroad (Subgroup of Replication Sample)

Vignette Dimensions	RE	FE
Exchange program (ref. no)	0.090	0.145
Financial scholarship (ref. no)	–0.228	–0.044
Exchange in group (ref. alone)	0.272	0.383
Related language skills		
Elementary skills (ref. no skills)	0.136	0.656
Good skills (ref. no skills)	1.504**	1.525***
Reputation host university		
Equal reputation (ref. poorer rep.)	0.081	0.046
Better reputation (ref. poorer rep.)	0.981+	0.780+
Host country preference		
Less desired (ref. not desired)	0.760	0.582
Strongly desired (ref. not desired)	0.821	0.871*
Family		
Show no expectations (ref. expect stay)	–0.504	–0.358
Expect study abroad (ref. expect stay)	–0.303	–0.157
Constant	2.803***	2.472***
σ_u	0.000	1.333
σ_e	1.291	1.291
ϱ	0.000	0.516
Wald χ^2 / F	22.32*	3.12**
R^2	0.271	0.250
$R^2_{between}$	0.599	0.203
R^2_{within}	0.369	0.397
$N_{vignettes}$	72	72
$N_{respondents}$	9	9

RE-Model: Random intercept fixed slopes regression, b-coefficients, robust SE.
FE-Model: Fixed effects regression, b-coefficients. + $p < 0.10$ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

coefficient should be interpreted only as a statistical correlation, or an interpretation should be dispensed with entirely.[5]

5 Since such subgroup analyses can always pose a threat to the factorial design and randomisation, the following conventional guidelines are proposed. Regarding checks of orthogonality through correlation analyses, the rules of thumb introduced by Cohen (1988) appear useful. "Small" correlations of $r \leq 0.1$ may reflect a sufficient degree of level balance, provided no other arguments suggest themselves. Regarding checks of randomisation using the Hausman test, a satisfied random effects assumption may reflect a sufficient degree of randomisation, i. e., if Hausman χ^2 is not significant.

4.2 Construct Validity

Information equivalence across the vignettes (Dafoe et al. 2018) can be violated. For example, a high reputation of the host university could be associated with a particular country different from a vignette including a university with a lower reputation. Similarly, the cut points the respondents use to make interpretations of levels can differ according to their background (e. g., two different students with equal language skills will consider their skills differently as being good or elementary). In addition, measurement equivalence regarding the rating scales must be sustained, i. e. responses obtained reflect only respondents' true value (King et al. 2004).

Although the questionnaire was translated and cross-checked several times by bilingual people, sufficient survey equivalence (Harkness 1998) can ultimately only be assumed in the cross-cultural replication. In the original study, a paper survey was used, in the replication study, a web survey was applied what can result in mode effects (De Leeuw 2005). In the original experiment, researchers and other students were also present in the room, which may have led to a stronger social desirability bias (Tourangeau and Yan 2007). Limited survey equivalence would thus be a potential explanation of why intentions and effects differ between the two experiments.

4.3 External Validity

In this study, assumptions have been (implicitly) made for contemporary undergraduate university students worldwide. Predictions were tested with a student sample of two disciplines from one German university and replicated with a sample of students of one Chinese University. The replication can help assess whether predictions hold and thus gain additional support for hypotheses (Popper 1959). While all hypotheses found support in the original study, not all effects could be replicated with the Chinese sample. This is unproblematic as long as this effect heterogeneity can be explained either by the original theory or a refined theory.

It must also be considered that the samples differ in their composition (Table 9) so that effect heterogeneity can rely on variables other than the cultural context. Such characteristics can be implemented as moderators into the original theory, and further empirical tests can be conducted.

Further, when a hypothesis for a specific population is tested, sampled subjects should represent this population (Stroebe et al. 2018). In the original study, the survey was carried out on-site in classrooms. Therefore, the course participants are adequately represented in the sample, but the population of students of the university is not. In the replication study, email lists were used to recruit students and self-selection based on interest is likely. The sample obtained does not represent the population of students at the university, and certainly not the population of all Chinese students.

One may argue that, given the non-probability sample of the respondents, the application of frequentist methods of statistical inference is not justified. However,

Table 9 Sample Characteristics

	Original sample (Germany) M (SD) / %	Replication sample (China) M (SD) / %
Age	23.2 (2.44)	25.1 (2.72)
Gender		
Male	65.8	47.6
Female	34.2	52.4
Relationship		
In relationship	43.1	31.5
Single	56.9	68.5
Discipline		
Economics	48.7	72.0
Engineering	51.3	28.0
Study level		
Bachelor student	74.7	61.6
Master student	25.3	38.4
Study abroad		
Experience	11.3	53.1
No experience	88.7	46.9
$N_{respondents}$	304	143
$N_{vignettes}$	2432	1144

since subjects were randomly assigned to treatments in a controlled probability procedure, the differences between treatments can be attributed to randomisation error, which permits a test of null hypotheses of treatment effects, though statistical inferences are restricted only to the actual respondent sample under study (Edgington 1966; Berk et al. 1995).

5 Conclusions

In this article, the potentials and pitfalls of FSE are discussed for empirical tests of theoretical explanations in the sociology of education. FSE can guarantee a high degree of internal validity due to manipulation and randomisation. The theory-driven data collection permits a direct test strategy. Theoretical elements of the decision situation are measured and causal effects are estimated. A potentially reduced social desirability bias may be advantageous regarding construct validity. Assessing external validity is facilitated by the easy implementation of heterogeneous samples, probability samples, and simple replicability. Accordingly, as with labora-

tory and field experiments, empirical tests with FSE offer a higher internal validity than tests with observational studies. Heterogeneous sampling or replication, just as with usual observational data, helps evaluate the external validity of a hypothesis. FSE can therefore be considered an ideal method for empirical tests of theoretical explanations of educational decisions in the sociology of education, provided the illustrated threats to validity can be avoided.

Yet, the method has two main disadvantages. First, the method holds uncertainty regarding the behavioural implications of any intentions obtained. Only those parts of the described general explanation model can be empirically tested that are related to the intention for educational investments. It must be assumed that the revealed effects on the tested educational intentions also correspond to real educational behaviour. This leads to a second issue. The analytical focus of the sociology of education is on social structures at the societal macro level, e. g., educational inequality. Since macrostructures are the aggregate of individual behaviour, statements based on the results obtained with FSE must necessarily remain speculative.

The FSE method is thus primarily a helpful complement to laboratory, field, and natural experiments on the one hand and cross-sectional and panel surveys on the other. An indirect test comparing educational decisions or achievements across social contexts based on large-scale survey data provides initial knowledge, investigates action-structuring components, and strengthens confidence regarding a theoretical model. A direct test using an FSE further provides empirical insights into the detailed causal mechanisms of decision-making. Thereby, FSE forces researchers to elaborate on underlying theoretical mechanisms through which causal effects may occur (Jackson and Cox 2013; Morgan and Winship 2015; Zangger and Becker 2019). In this way, the strengths and weaknesses of various approaches are balanced, and the methods contribute robust evidence to the cumulative advancement in the sociology of education.

6 References

Ajzen, Icek. 1991. The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes* 50(2): 179–211.

Ajzen, Icek, Thomas C. Brown, and Franklin Carvajal. 2004. Explaining the Discrepancy Between Intentions and Actions: The Case of Hypothetical Bias in Contingent Valuation. *Personality and Social Psychology Bulletin* 30(9): 1108–1121.

Alexander, Cheryl S., and Henry Jay Becker. 1978. The Use of Vignettes in Survey Research. *Public Opinion Quarterly* 42(1): 94–104.

Armacost, Robert L., Jamshid C. Hosseini, Sara A. Morris, and Kathleen A. Rehbein. 1991. An Empirical Comparison of Direct Questioning, Scenario, and Randomized Response Methods for Obtaining Sensitive Business Information. *Decision Sciences* 22(5): 1073–1099.

Auspurg, Karin, and Thomas Hinz. 2015. *Factorial Survey Experiments*. London / Thousand Oaks: Sage Publications.

Auspurg, Katrin, Thomas Hinz, and Stefan Liebig. 2009. Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey. *Methoden – Daten – Analysen* 3(1): 59–96.

Auspurg, Katrin, Thomas Hinz, Stefan Liebig, and Carsten Sauer. 2015. The Factorial Survey as a Method for Measuring Sensitive Issues. Pp. 137–150 in *Improving Survey Methods: Lessons from Recent Research*, edited by Uwe Engel, Ben Jann, Peter Lynn, Annette Scherpenzeel, and Patrick Sturgis. New York / Hove: Routledge.

Auspurg, Katrin, and Annette Jäckle. 2017. First Equals Most Important? Order Effects in Vignette-Based Measurement. *Sociological Methods and Research* 46(3): 490–539.

Becker, Rolf. 2000. Klassenlage und Bildungsentscheidungen. Eine empirische Anwendung der Wert-Erwartungstheorie. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 52(3): 450–474.

Becker, Rolf, and Anna Etta Hecken. 2009. Higher Education or Vocational Training? An Empirical Test of the Rational Action Model of Educational Choices Suggested by Breen and Goldthorpe and Esser. *Acta Sociologica* 52(1): 25–45.

Becker, Rolf, and Heike Solga (eds.). 2012. *Soziologische Bildungsforschung. Sonderheft 52 der Kölner Zeitschrift für Soziologie und Sozialpsychologie*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Berk, Richard A., Bruce Western, and Robert E. Weiss. 1995. Statistical Inference for Apparent Populations. *Sociological Methodology* 25: 421–458.

Boudon, Raymond. 1974. *Education, Opportunity, and Social Inequality. Changing Prospects in Western Societies*. New York: Wiley & Sons.

Breen, Richard, and John H. Goldthorpe. 1997. Explaining Educational Differentials. Towards a Formal Rational Action Theory. *Rationality and Society* 9(3): 275–305.

Brüderl, Josef. 2004. Die Überprüfung von Rational-Choice-Modellen mit Umfragedaten. Pp. 163–180 in *Rational-Choice Theorie in den Sozialwissenschaften. Anwendungen und Probleme*, edited by Andreas Diekmann, and Thomas Voss. München: Oldenbourg Verlag.

Brüderl, Josef, and Volker Ludwig. 2015. Fixed-Effects Panel Regression. Pp. 327–359 in *The SAGE Handbook of Regression Analysis and Causal Inference*, edited by Henning Best, and Christof Wolf. London et al.: Sage.

Buskens, Vincent, Werner Raub, and Marcel A.L.M. van Assen. 2014. *Micro-Macro Links and Microfoundations in Sociology*. London / New York: Routledge.

Campbell, Donald T. 1957. Factors Relevant to the Validity of Experiments in Social Settings. *Psychological Bulletin* 54(4): 297–312.

Campbell, Donald T. and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge: Harvard University Press.

Collett, Jessica L., and Ellen Childs. 2011. Minding the Gap: Meaning, Affect, and the Potential Shortcomings of Vignettes. *Social Science Research* 40(2): 513–522.

Cook, Thomas D. 2001. Sciencephobia. Why Education Rejects Randomized Experiments. *Education Next* 1(3): 63–68.

Dafoe, Allan, Baobao Zhan, and Devin Caughey. 2018. Information Equivalence in Survey Experiments. *Political Analysis* 26(4): 399–416.

Daniel, Annabell, Martin Neugebauer, and Rainer Watermann. 2019. Studienabbruch und Einstellungschancen auf dem Ausbildungsmarkt. Ein faktorieller Survey mit Arbeitgeber/innen. *Zeitschrift für Erziehungswissenschaft* 22: 1147–1174.

De Leeuw, Edith D. 2005. To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics* 21(2): 233–255.

De Wolf, Inge and Rolf Van Der Velden. 2001. Selection Processes for Three Types of Academic Jobs. An Experiment among Dutch Employers of Social Sciences Graduates. *European Sociological Review* 17(3): 317–330.

Di Stasio, Valentina. 2014. Education as a Signal of Trainability: Results from a Vignette Study with Italian Employers. *European Sociological Review* 30(6): 796–809.

Dülmer, Hermann. 2007. Experimental Plans in Factorial Surveys: Random or Quota Design? *Sociological Methods and Research* 35(3): 382–409.

Dülmer, Hermann. 2016. The Factorial Survey. Design Selection and its Impact on Reliability and Internal Validity. *Sociological Methods and Research* 45(2): 304–347.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences. A Design-Based Approach*. Cambridge: Cambridge University Press.

Düval, Sabine and Thomas Hinz. 2020. Different Order, Different Results? The Effects of Dimension Order in Factorial Survey Experiments. *Field Methods* 32(1): 23–37.

Edgington, Eugene. 1966. Statistical Inference and Nonrandom Samples. *Psychological Bulletin* 66(6): 485–487.

Eifler, Stefanie, and Knut Petzold. 2019. Validity Aspects of Vignette Experiments: Expected What-If Differences Between Reports of Behavioral Intentions and Actual Behavior. Pp. 393–416 in *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*, edited by Paul Lavrakas, Michael Traugott, Courtney Kennedy, Allyson Holbrook, Edith de Leeuw, and Brady West. Hoboken: John Wiley & Sons.

Elwert, Felix, and Christopher Winship. 2014. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology* 40: 31–53.

Erikson, Robert, and Jan O. Jonsson. 1996. Explaining Class Inequality in Education: The Swedish Test Case. Pp. 1–63 in *Can Education Be Equalized?* edited by Robert Erikson, and Jan O. Jonsson. Boulder: Westview Press.

Erlinghagen, Marcel and Karsten Hank. 2018. *Neue Sozialstrukturanalyse. Ein Kompass für Studienanfänger.* 2. Auflage. Stuttgart: UTB.

Esser, Hartmut. 1996. What is Wrong With "Variable Sociology"? *European Sociological Review* 12(2): 159–166.

Esser, Hartmut. 1999. *Soziologie. Spezielle Grundlagen. Band 1: Situationslogik und Handeln*. Frankfurt a.M. / New York: Campus Verlag.

Finch, Janet. 1987. The Vignette Technique in Survey Research. *Sociology* 21(1): 105–114.

Finger, Claudia. 2016. Institutional Constraints and the Translation of College Aspirations into Intentions– Evidence from a Factorial Survey. *Research in Social Stratification and Mobility* 46: 112–128.

Friedman, Daniel, and Alessandra Cassar. 2004. *Economics Lab. An Intensive Course in Experimental Economics*. London / New York: Routledge.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton & Company.

Goldthorpe, John H. 2001. Causation, Statistics, and Sociology. *European Sociological Review* 17(1): 1–20.

Grusky, David B. 1994. *Social Stratification: Class, Race, and Gender in Sociological Perspective*. Boulder: Westview Press.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. Validating Vignette and Conjoint Survey Experiments Against Real-World Behavior. *Proceedings of the National Academy of Sciences* 112(8): 2395–2400.

Harkness, Janet (ed.). 1998. *Cross-Cultural Survey Equivalence*. Zentrum für Umfragen, Methoden und Analysen. Mannheim: ZUMA

Hedström, Peter, and Richard Swedberg. 1996. Rational Choice, Empirical Research, and the Sociological Tradition. *European Sociological Review* 12(2): 127–146.

Hedström, Peter, and Petri Ylikoski. 2010. Rational Choice, Empirical Research, and the Sociological Tradition. *Annual Review of Sociology* 36: 49–67.

Holland, Paul W. 1986. Causal Mechanisms in the Social Sciences. *Journal of the American Statistical Association* 81(4): 945–960.

Hox, Joob J., Ita G. Kreft, and Piet L. J. Hermkens. 1991. The Analysis of Factorial Surveys. *Sociological Methods and Research* 19(4): 439–510.

Hughes, Rhidian, and Meg Huby. 2004. The Construction and Interpretation of Vignettes in Social Research. *Social Work & Social Sciences Review* 11(1): 36–51.

Jackson, Michelle, and D.R. Cox. 2013. The Principles of Experimental Design and Their Application in Sociology. *Annual Review of Sociology* 39: 27–49.

Jasso, Guillermina. 2006. Factorial Survey Methods for Studying Beliefs and Judgements. *Sociological Methods and Research* 34(3): 334–423.

Keller, Tamás. 2018. Dare to Dream: A Vignette Survey on Self-Selection in Secondary Education Track Choice. *Sociological Research Online* 23(2): 1–20.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review* 98: 191–207.

Krumpal, Ivar. 2013. Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality and Quantity* 47(3): 2025–2047.

Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt. 1994. Efficient Experimental Design With Marketing Research Applications. *Journal of Marketing Research* 31(4): 545–557.

Levitt, Steven D., and John A. List. 2007. What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *The Journal of Economic Perspectives* 21(2): 153–174.

Lindenberg, Siegwart. 1996. Die Relevanz theoriereicher Brückenannahmen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 48(1): 126–140.

Lörz, Markus, Nicolai Netz, and Heiko Quast. 2016. Why Do Students from Underprivileged Families Less Often Intend to Study Abroad? *Higher Education* 72(2): 153–174.

Markovsky, Barry, and Kimmo Eriksson. 2012. Comparing Direct and Indirect Measures of Just Rewards. *Sociological Methods and Research* 41(1): 199–216.

McDonald, Patrick. 2019. How Factorial Survey Analysis Improves Our Understanding of Employer Preferences. *Swiss Journal of Sociology* 45(2): 237–260.

Mook, Douglas G. 1983. In Defense of External Invalidity. *American Psychologist* 38(4): 379–387.

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* 2nd edition. Cambridge: Cambridge University Press.

Möser, Sara, David Glauser, and Rolf Becker. 2019. Valuation of Labour Market Entrance Positions among (Future) Apprentices - Results from Two Discrete Choice Experiments. *Journal of Choice Modelling* 33: 100180.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton: Princeton University Press.

Opp, Karl-Dieter. 1999. Contending Conceptions of the Theory of Rational Choice. *Journal of Theoretical Politics* 11(2): 171–202.

Pager, Devah, and Lincoln Quillian. 2005. Walking the Talk? What Employers Say Versus What They Do. *American Sociological Review* 70(3): 355–380.

Pearl, Judea. 2010. The Foundations of Causal Inference. *Sociological Methodology* 40(1): 75–149.

Petzold, Knut. 2018. Fachspezifische Entscheidungen zum Auslandsstudium. Ein experimenteller Test der Wert-Erwartungstheorie. *Zeitschrift für Erziehungswissenschaft* 21(4): 817–838.

Petzold, Knut, and Petra Moog. 2018. What Shapes the Intention to Study Abroad? An Experimental Approach. *Higher Education* 75(1): 35–54.

Petzold, Knut, and Tobias Wolbring. 2018. What Can We Learn From Factorial Surveys About Human Behavior? A Validation Study Comparing Field and Survey Experiments on Discrimination. *Methodology*. 15(1): 19-30.

Petzold, Knut, and Tobias Wolbring. 2019. Zur Verhaltensvalidität von Vignettenexperimenten: Theoretische Grundlagen, Forschungsstrategien und Befunde. Pp. 307–338 in *Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente.*, edited by Natalja Menold, and Tobias Wolbring. Wiesbaden: Springer VS.

Popper, Karl Raimund. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.

Rosenbaum, Paul R. 2010. *The Design of Observational Studies*. New York et al.: Springer.

Rössel, Jörg. 2009. *Sozialstrukturanalyse. Eine kompakte Einführung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Rossi, Peter H., and Andy B. Anderson. 1982. The Factorial Survey Approach: An Introduction. Pp. 15–67 in *Measuring Social Judgments. The Factorial Approach*, edited by Peter H. Rossi, and Steven L. Nock. Beverly Hills et al.: Sage Publications.

Rubin, Donald B. 1980. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association* 75(371): 591–593.

Rubin, Donald B. 2008. For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics* 2(3): 808–840.

Salisbury, Mark H., Paul U. Umbach, Michael B. Paulsen, and Ernest T. Pascarella. 2009. Going Global: Understanding the Choice Process of the Intent to Study Abroad. *Research in Higher Education* 50(2): 119–143.

Sauer, Carsten, Katrin Auspurg, Thomas Hinz, and Stefan Liebig. 2011. The Application of Factorial Survey in General Population Samples: The Effect of Respondent Age and Education on Response Times and Response Consistency. *Survey Research Methods* 5(3): 89–102.

Sauer, Carsten, Katrin Auspurg, and Thomas Hinz. 2020. Designing Multi-Factorial Survey Experiments: Effects of Presentation Style (Text or Table), Answering Scales, and Vignette Order. *methods, data, analyses* 14(2): 195–214.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston / New York: Houghton Mifflin Company.

Shamon, Hawal, Hermann Dülmer, and Adam Giza. 2019. The Factorial Survey: The Impact of the Presentation Format of Vignettes on Answer Behavior and Processing Time. *Sociological Methods and Research* DOI: 10.1177/0049124119852382.

Shi, Lulu P., Christian Imdorf, Robin Samuel, and Stefan Sacchi. 2018. How Unemployment Scarring Affects Skilled Young Workers: Evidence from a Factorial Survey of Swiss Recruiters. *Journal for Labour Market Research* 52: Article n°7.

Snijders, Tom A. B., and Roel J. Bosker. 2012. *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. Los Angeles, CA: Sage.

Stocké, Volker. 2010. Der Beitrag der Theorie rationaler Entscheidung zur Erklärung von Bildungsungleichheit. Pp. 73–94 in *Bildungsverlierer. Neue Ungleichheiten*, edited by Gudrun Quenzel. Wiesbaden: Springer VS.

Stroebe, Wolfgang, Volker Gardenne, and Bernard A. Nijstad. 2018. Do Our Psychological Laws Apply Only to College Students? External Validity Revisited. *Basic and Applied Social Psychology* 40(6): 384–395.

Thelin, Mikael and Thomas Niedomysl. 2015. The (Ir)Relevance of Geography for School Choice: Evidence from a Swedish Choice Experiment. *Geoforum* 67(1): 110–120.

Tourangeau, Roger, and Ting Yan. 2007. Sensitive Questions in Surveys. *Psychological Bulletin* 133(5): 859–883.

Trafimow, David. 2012. The Role of Auxiliary Assumptions for the Validity of Manipulations and Measures. *Theory & Psychology* 22(4): 486–498.

Verlegh, Peeter W. J., Hendrik N. J. Schifferstein, and Dick R. Wittink. 2002. Range and Number-of-Levels Effects in Derived and Stated Measures of Attribute Importance. *Marketing Letters* 13(1): 41–52.

Wallander, Lisa. 2009. 25 Years of Factorial Surveys in Sociology: A Review. *Social Science Research* 38(3): 505–520.

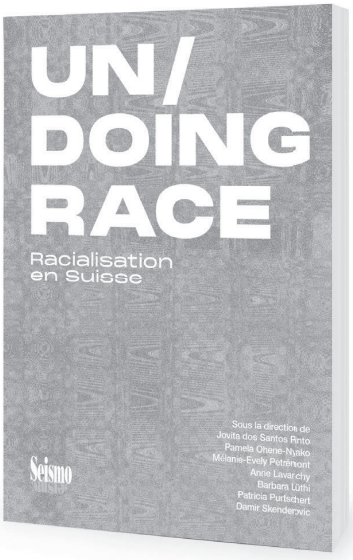
Walzenbach, Sandra. 2019. Hiding Sensitive Topics by Design? An Experiment on the Reduction of Social Desirability Bias in Factorial Surveys. *Survey Research Methods* 13(1): 103–121.

Webster, Murray Jr., and Jane Sell. 2007. *Laboratory Experiments in the Social Sciences*. Amsterdam: Academic Press / Elsevier.

Wooldridge, Jeffrey M. 2013. *Introductory Econometrics: A Modern Approach.* 5th Edition. Mason, OH: South-Western Cengage Learning.

Zangger, Christoph, and Rolf Becker. 2019. Experiments in the Sociology of Education: Causal Inference and Estimating Causal Effects in Sociological Research on Education. Pp. 153–171 in *Research Handbook on the Sociology of Education*, edited by Rolf Becker. Cheltenham, UK: Edward Elgar Publishing.

Éditions Seismo
Sciences sociales et questions de société



Quelles sont les significations de la race, de la racialisation et du racisme en Suisse ? Comment les phénomènes de racisme et de racialisation sont-ils liés à son héritage colonial ? Comment le traitement du racisme a-t-il évolué au cours de l'histoire ? Quel est le rôle du militantisme antiraciste, en particulier celui des Noirs et des personnes racisées ? En abordant ces questions, l'ouvrage montre comment le racisme est enraciné dans les structures des sociétés modernes. Comme le font ressortir les contributions, le racisme structurel et quotidien est également présent en Suisse, dans les domaines sociaux les plus divers. L'ouvrage propose, d'une part, des concepts et des approches permettant de saisir les processus et les mécanismes de racialisation. Il vise, d'autre part, à favoriser l'échange et à la circulation des connaissances pour examiner les régimes de racialisation. L'ouvrage jette ainsi les bases d'une réflexion scientifique critique en Suisse sur le racisme et l'utilisation des concepts analytiques liés à la race.

Jovita dos Santos Pinto, Pamela Ohene-Nyako, Mélanie Pétrémont, Anne Lavanchy, Barbara Lüthi, Patricia Purtschert, Damir Skenderovic (dir.)

Un/Doing Race
Racialisation en Suisse

ISBN 978-3-03777-252-2
320 pages, 15.5 × 22.5 cm
Fr. 38.– / Euro 33.–

Aussi disponible en allemand

Anne Lavanchy est anthropologue et professeure à la HES-SO Genève en Travail social. **Barbara Lüthi** est professeure assistante en histoire à l'Université de Cologne. **Pamela Ohene-Nyako** est assistante-doctorante en histoire contemporaine à l'Université de Genève. **Mélanie-Evely Pétrémont** est doctorante en géographie à l'Université de Genève. **Patricia Purtschert** est philosophe, professeur d'études genres et co-directrice du Centre Interdisciplinaire en Etudes Genre de l'Université de Berne. **Jovita dos Santos Pinto** est doctorante à l'IZFG de l'Université de Berne. **Damir Skenderovic** est professeur d'histoire contemporaine à l'Université de Fribourg.

Éditions Seismo, Zurich et Genève **www.editions-seismo.ch** **livre@editions-seismo.ch**