

Factorial Survey Experiments to Predict Real-World Behavior: A Cautionary Tale from Hiring Studies

Andrea G. Forster,^a Martin Neugebauer^b

a) Utrecht University; b) Karlsruhe University of Education

Abstract: Factorial surveys (FSs) are increasingly used to predict real-world decisions. However, there is a paucity of research assessing whether these predictions are valid and, if so, under what conditions. In this preregistered study, we sent out $N = 3,002$ applications to job vacancies in Germany and measured real-world responses. Eight weeks later, we presented nearly identical applicant profiles to the same employers as a part of an FS. To explore the conditions under which FSs provide valid behavioral predictions, we varied the topic sensitivity and tested whether behavioral predictions were more successful after filtering out respondents who gave socially desirable answers or did not exert sufficient effort when answering FS vignettes. Across conditions, the FS results did not correspond well with the real-world benchmark. We conclude that researchers must exercise caution when using FSs to study (hiring) behavior.

Keywords: factorial survey experiment; field experiment; audit study; behavioral validity; validation study; hiring

Reproducibility Package: The code and data needed to reproduce the analyses are available at the Open Science Framework: <https://osf.io/x2tcp/>

Citation: Forster, G. Andrea and Martin Neugebauer. 2024. "Factorial Survey Experiments to Predict Real-World Behavior: A Cautionary Tale from Hiring Studies" *Sociological Science* 11: 886-906.


Received: April 26, 2024

Accepted: August 23, 2024

Published: September 24, 2024

Editor(s): Arnout van de Rijt, Stephen Vaisey

DOI: 10.15195/v11.a32

Copyright: © 2024 The Author(s). This open-access article has been published under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited. 

FACTORIAL survey (FS) experiments, also known as vignette studies, have recently gained tremendous popularity in the social sciences. They combine the ability to estimate how humans form multidimensional judgments through experimental variation with the opportunity to survey large representative samples (Auspurg and Hinz, 2015; Mutz, 2011; Rossi et al., 1974).

Although initially developed to measure attitudes or normative judgments, FSs are now commonly used to predict real-world decisions. In a recent review of 441 FS experiments published between 1982 and 2018, Treischl and Wolbring (2022) found that 45 percent of all FS studies aimed to approximate actual behavior, with the proportion increasing significantly since 2010. For instance, researchers have used FSs to investigate whether voters discriminate against women running for office (Schwarz and Coppock, 2022), whether incentives spur COVID-19 vaccination uptake (Klüver et al., 2021), and why employers are less likely to invite individuals who have been unemployed long term for job interviews (Van Belle et al., 2018). The central question is: Do behavioral intentions correspond to real-world behavior, allowing us to estimate actual decisions with relatively little effort within the scope of a survey? Or are the critics (e.g., Bertrand and Mullainathan, 2001; Collett and Childs, 2011; Eifler, 2007) correct in asserting that response biases and the hypothetical nature of FSs render them unsuitable for predicting "true" behavior?

Research on this question has remained limited and inconclusive, despite its fundamental importance. Some studies have indicated that FSs have low-behavioral

validity, highlighting significant disparities between decision intentions and a behavioral real-world benchmark (Eifler, 2007; Findley et al., 2017; Pager and Quillian, 2005; Wulff and Villadsen, 2020). Other studies have found a higher level of agreement regarding the relative effects being studied, even if the level of behavior is not always correctly predicted in FSs (Diehl et al., 2013; Eifler, 2010; Groß and Börensen, 2009; Hainmueller et al., 2015; Nisic and Auspurg, 2009; Petzold and Wolbring, 2019).

Critically, the few existing studies have not adequately ensured that the experimental design of the FS aligns well with the behavioral benchmark. Most notably, participants differed between FS and benchmark in almost all validation studies conducted to our knowledge (except for Pager and Quillian (2005) and Petzold and Wolbring (2019)). However, to assess whether inferences can be drawn from behavioral intentions to real-world behavior, the experimental designs of an FS and a study attempting to replicate the FS in a real-world setting should be as similar as possible. Only then can we rule out other factors causing discrepancies in the results, such as differences in samples, formulation of treatment conditions, operationalization of dependent variables, or the extent of information available in the decision-making situation.

In addition, to our knowledge, no study has systematically examined the conditions under which FSs can accurately predict real-world behavior. Survey research has shown that social desirability bias (SDB) leads to distorted self-reports, especially concerning sensitive topics, such as delinquency or ethnic discrimination (Crowne and Marlowe, 1960; Krumpal, 2013). Does this imply that the behavioral intentions stated in FSs correspond more closely to actual behavior for less sensitive issues? In addition, we know that respondents differ in their tendency to provide socially desirable responses or to put in sufficient effort when answering a survey (Crowne and Marlowe, 1960; Huang et al., 2012). Does this imply that FSs more accurately depict “true” behavior when respondents who give socially desirable answers or do not make an effort into survey responses are filtered out? Answers to these questions are essential for researchers to understand for which topics and respondents they can employ FSs to predict actual behavior effectively.

Against this backdrop, we aimed to make two contributions through our study. First, we conducted a validation test of an FS by ensuring that the experimental designs of the FS and the study replicating the FS in a real-world setting were maximally similar. Our application case is employer hiring decisions, a topic often studied using FSs that is essential for understanding the role of employers as gatekeepers in the labor market and their impact on perpetuating inequalities in school-to-work transitions and individuals’ careers (Bills, Di Stasio, and Gerxhani, 2017). Our behavioral benchmark is a field experiment (FE), in which we sent out $N = 3,002$ applications to actual vacancies on the German apprenticeship labor market. Eight weeks later, we asked the same employers, without their knowledge of participating in the FE, to evaluate substantively identical applicant profiles in an FS.¹

Second, we examined the conditions under which FSs can adequately predict actual behavior. We varied the topic’s sensitivity by including a sensitive job applicant characteristic (ethnicity) and a non-sensitive “meritocratic” characteristic

(education). We also assessed how respondent characteristics, such as their tendency to provide socially desirable answers and their effort in responding, influenced the correspondence between actual and reported behavior.

Overall, we found that, regardless of the sensitivity of the topic and the respondent's tendency to provide socially desirable answers or put effort into responding, the results from the FS did not correspond well with the results of the FE. We discuss the implications of these findings and recommend exercising great caution when studying real-world behavior with FS experiments.

Factorial Survey Intentions and the Real World

A common criticism of FSs is that they can measure only behavioral intentions not actual behavior (e.g., Collett and Childs, 2011; Eifler, 2007; Pager and Quillian, 2005). As the theory of planned behavior (Ajzen, 1991) outlines, researchers consider behavioral intentions as the best predictor of behavior. However, behavioral intentions often do not match actual behavior due to a lack of behavioral control or the difficulty of forecasting one's behavior in a given situation (Armitage and Conner, 2001). Therefore, many scholars fear that FS may be useless for predicting real-world behavior.

To date, it is largely unclear whether these concerns are justified. The few existing validation studies have paid too little attention to ensure that an FS closely mimics an actual decision situation. Variations in item wording, the measurement of the dependent variable, and the selection of the analysis sample are known to influence survey responses (e.g., Bertrand and Mullainathan, 2001). Therefore, diverging results in previous validations may simply reflect a failure to match the FS with the behavioral benchmark. In line with Petzold and Wolbring (2019), we argue that, ideally, FS and observation of real behavior should differ only in measuring intention or actual behavior; all other parameters should closely match. To our knowledge, only two studies (Pager and Quillian, 2005; Petzold and Wolbring, 2019) approach this ideal; however, important deviations remain, particularly regarding the measurement of the outcome. Although decisions in real life are typically dichotomous, FS guides recommend measuring the outcome on a continuous probability scale (Auspurg and Hinz, 2015), and the aforementioned studies have followed this differential scaling (Pager and Quillian, 2005; Petzold and Wolbring, 2019).²

A second, closely related criticism of FSs pertains to the fact that they create artificial decision situations, a phenomenon sometimes referred to as hypothetical bias (Gutfleisch, Samuel, and Sacchi, 2021; Petzold and Wolbring, 2019). For instance, many FS studies present short text vignettes, which typically involve substantially simplifying a real-world decision situation. In addition, real-world decisions often involve actual costs, but FS decisions do not. Furthermore, a lack of psychological realism can arise when unsuitable respondents are recruited. Although it is reasonable to expect that experience with the situation under study is essential to obtain valid estimates (Mize and Manago, 2022), many FS studies confront respondents with a decision situation they do not encounter in the real world (e.g., asking individuals without hiring experience to select job candidates).

Against this backdrop, we present a validation study in which the FS and behavioral benchmark are maximally aligned while hypothetical bias is minimized. Through this approach, we aim to conduct a “fair” validation test, ensuring that the FS feels as natural as possible so that the psychological processes involved in the survey mirror those in real life (Aronson, Wilson, and Brewer, 1998).

Under What Conditions Should Factorial Surveys Provide Valid Predictions of Behavior?

We suspect that actual behavior can be predicted with FSs only under certain conditions, such as when the topic is not sensitive or when respondents are willing to put sufficient effort into answering the survey. To date, no studies have systematically investigated these conditions. We address this research gap, focusing on two well-known response biases: SDB and insufficient effort responding (IER).

Social Desirability Bias

SDB—respondents’ inclination to answer in a socially acceptable manner rather than providing truthful information—is a well-known challenge in survey research (Crowne and Marlowe, 1960; Krumpal, 2013). FS experiments are often considered less susceptible to SDB than single-item questions because respondents must consider multiple dimensions simultaneously (Auspurg et al., 2014). However, for topics associated with strong social norms or ethical concerns, SDB may still be present in FSs. This occurs because respondents tend to pay more attention to the sensitive dimension, even when they must evaluate multiple dimensions at the same time (Krumpal, 2013; Walzenbach, 2019). Consequently, higher topic sensitivity is likely to result in lower FS validity.

Furthermore, SDB can be described as a response style that individuals with a strong inclination toward social desirability tend to exhibit more than others (Crowne and Marlowe, 1960; Kemper, Beierlein, and Bensch, 2012; Paulhus, 2002). These respondents often have a significant need for approval and tend to overstate their desirable behaviors or deny common but undesirable behaviors. Therefore, assessing disposition toward social desirability in the FS may allow researchers to filter respondents who are more susceptible to SDB than others.

Insufficient Effort Responding

A similar argument can be made regarding IER, the tendency of respondents to put forth inadequate effort when answering survey questions, resulting in unreliable or biased data (Huang et al., 2012). Although FSs tend to offer more variety than single-item questions, which should counteract IER, some respondents put in little effort when answering questions about vignettes, partly because they question the general value of surveys (Rogelberg et al., 2001).

Summary of Theoretical Considerations

In summary, we propose that an FS can effectively measure real behavior when it exhibits a high degree of psychological realism and pertains to a less sensitive topic and when respondents are filtered to ensure honest and effortful responses.

Use Case: Employer Hiring Decisions

We tested our theoretical expectations empirically using employer hiring decisions as a prominent application case. Although unobtrusive FEs have traditionally been used to study hiring behavior by sending fictitious applications to real job vacancies (for an overview, see Baert (2018)), FSs offer several advantages: they are more cost effective, allow more dimensions to be manipulated, and facilitate the collection of respondent characteristics. This makes FSs, at least in principle, particularly suited to study the mechanisms behind employer decision making. Importantly, FSs are also ethically safe, as they do not require deceiving employers. Consequently, researchers increasingly use FSs to investigate employer behavior (e.g., Auer et al., 2019; Damelang et al., 2019; Di Stasio and van de Werfhorst, 2016; Neugebauer and Daniel, 2022; Protsch and Solga, 2015).

Specifically, we studied the likelihood of receiving an invitation to a job interview for entry-level apprenticeship positions in Germany. Apprenticeships represent an important route into the German labor market. They are particularly well suited to our experiments, as the hiring process follows a relatively standardized procedure, making it easier to develop convincing experimental materials that fit a large number of positions. We have selected apprenticeship positions in the upper segment of vocational occupations in four occupational fields (electronics, laboratory, administration, and media) to cover a wide range of companies, industries, required skill sets, and gender compositions.

To investigate the influence of topic sensitivity, we varied sensitive and less sensitive applicant characteristics. The sensitive characteristic is ethnicity, which has been extensively explored with FEs (for an overview, see Lippens, Vermeiren, and Baert, 2013) and, to a lesser extent, with FSs (Auer et al., 2019; Baert and De Pauw, 2014; Van Beek, Koopmans, and Van Praag, 1997; Damelang and Abraham, 2016). Ethnic hiring discrimination has been observed in various contexts, such as applications for apprenticeships (Schneider, Yemane, and Weinmann, 2014), despite being recognized as socially inappropriate and illegal (Barr, Lane, and Nosenzo, 2018). Consequently, respondents in an FS might be reluctant to acknowledge that they consider an applicant's ethnic background when making decisions.

The less sensitive characteristic is higher education non-completion. Distinguishing between applicants based on their educational background aligns with meritocratic ideals and should not evoke strong social norms. Specifically, we compared the hiring chances of applicants who completed upper-secondary education with those who also pursued a brief period of college education but then discontinued their studies in favor of vocational training. Both upper-secondary graduates and college non-completers are typical applicants for apprenticeships in the upper segment of vocational occupations, where possessing an upper-secondary degree

is not uncommon (Neugebauer and Daniel, 2022).³ Relatively few studies have considered the topic of higher education non-completion (for an FS example, see Neugebauer and Daniel (2022)), and, to our knowledge, not a single FE has done so. However, a growing number of FE and FS studies have examined how other aspects of education and skills contribute to labor market integration (Di Stasio and van de Werfhorst, 2016; Mai, 2020; Protsch and Solga, 2015). In the context of our use case, we developed the following hypothesis regarding topic sensitivity:

H1: The true FE effect of higher education non-completion on the likelihood of receiving an invitation for a job interview should be more readily replicated with an FS than the true effect of an applicant's ethnic origin.

Concerning the influence of social desirability as a respondent characteristic, we hypothesized the following:

H2: Both non-completion and ethnic origin effects should be more readily replicated in the FS for respondents with a low disposition to provide socially desirable answers.

Finally, we hypothesized the following regarding IER:

H3: Both non-completion and ethnic origin effects should be more readily replicated in the FS for respondents who show a high level of effort when completing the survey.

Data and Methods

Sample and Data Collection Procedure

Following a sequential design, we submitted fictitious applications for real apprenticeship positions (FE). Eight weeks later, we presented nearly identical FS vignettes to the same employers and collected information on respondents' SDB and IER (for a visualization, see Fig. 1).

To obtain a representative sample of real employers, we web-scraped all job advertisements for 24 different occupations (details are shown in the online supplement 1.1) from four occupational fields (electronics, laboratory, administration, and media) advertised between February 1, 2022 and November 30, 2022 on the official job portal of the Federal Employment Agency, on which 70 percent of all apprenticeships in Germany are advertised (BIBB, 2015). Within a few days of an advertisement appearing, we applied if we could do so via email. Many companies advertised multiple positions. We applied only once per company to minimize the effort for the companies and the risk of detection for us. In total, we sent out $N = 3,002$ applications.

To generate realistic application materials, we reviewed real applications and conducted 16 qualitative interviews with job counselors and recruiters. The materials, which we optimized through several pretests, included a cover letter, CV, and school certificate (for details, see the online supplement 1.2). We optimized the cover letters for our four occupational fields, highlighting skills and interests

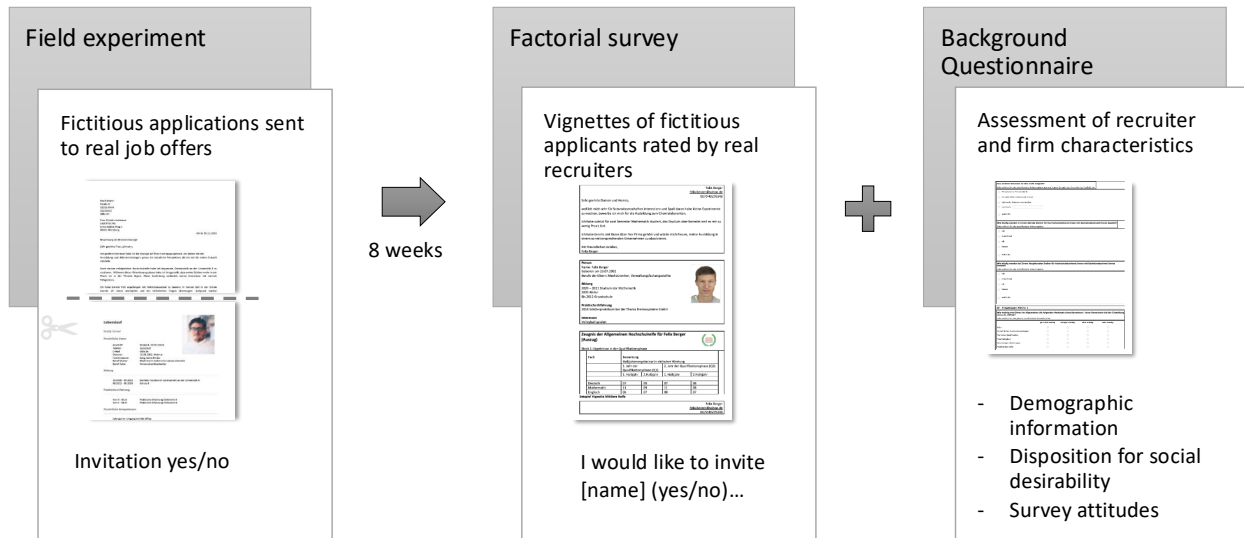


Figure 1: Data collection sequence.

specific to each occupation. After submitting one application per employer, we recorded whether an applicant received a positive (invitation for an interview or assessment test) or negative (rejection or no response) reply (for more details, see the online supplement 1.3).

Eight weeks later, within the framework of an FS, we presented nearly identical applications to the same employers. For this purpose, we invited them to complete an online survey using the same email address we had used to send the application. In the FS, we informed the employers of the fictitious nature of the applicants; however, we used a cover story to disguise the true aim of the study. In total, 480 employers completed the entire survey, including measures for SDB and IER (response rate: 16 percent), resulting in $8 \times 480 = 3,840$ vignette ratings. As in all surveys, the question of selective participation arose. We tested selection based on observables and Figure 2 illustrates the geographic distribution and firm size distribution across both samples. For these and other observables, the FS sample closely aligns with the FE sample. We will return to this point later.

To minimize hypothetical bias, we carefully transformed the FE application materials into an FS format. To accomplish this, we created rather elaborate vignettes that closely resembled the FE materials, in contrast to the brief text vignettes commonly found in the literature. Each vignette is comprised of three components: a concise cover letter, a tabular resume, and an abbreviated high school diploma, presented to fit on one computer screen page. The online supplement 1.2 provides an illustrative comparison between the FE and the “adapted” FS application materials. We also closely replicated the actual decision-making process. Based on our qualitative pre-studies, recruiters typically review all applications and make an initial assessment before deciding whether to extend an invitation for the next stage (interview or assessment center). After each of the eight vignettes, we prompted the recruiters to assess, on an 11-point scale from 0 percent to 100 percent, the likelihood of inviting the candidate to the next stage. Once they had evaluated all

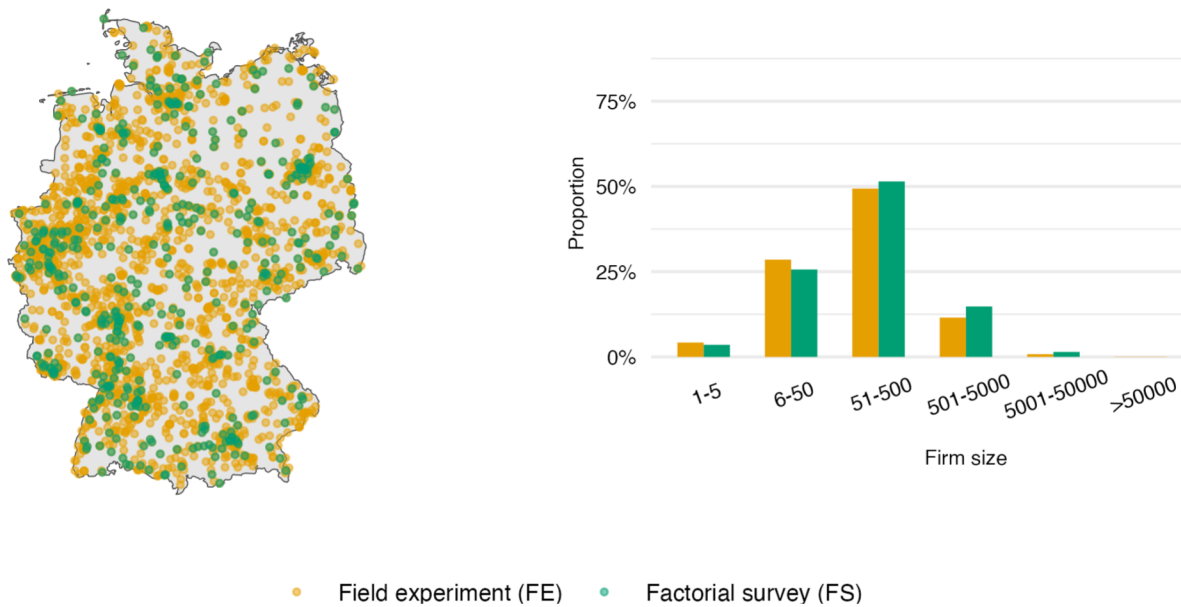


Figure 2: Geographic distribution and firm size distribution in FE and FS samples.

eight vignettes, we presented them with the entire pool of applicants on one page, along with key information and their own prior ratings. On this page, they could make a final binary decision on which candidates they wished to invite. We used this binary evaluation as our dependent variable to make the FS comparable to the FE. Sensitivity analyses with the 11-point scale outcome yielded substantively equivalent findings (see the online supplement 4.6).

Experimental Variation

To test topic sensitivity (H1) in both studies, we randomly varied the ethnic background (indicated by the applicant's name) and whether an applicant had dropped out of higher education or not (displayed on their CV and cover letter). We chose typical German and Turkish names because Turks are the largest ethnic minority group in Germany, and research has shown them to be discriminated against in the labor market (Schneider et al., 2014). We drew the randomly assigned names of the applicants from a list of the most common German and Turkish names in the respective cohorts. In addition, we varied applicants' gender and, in the case of higher education dropouts, their field of study (these specific details are not the focus of this article and are thus not discussed further). Table 1 provides an overview.

Based on our pre-studies and the existing literature (BIBB, 2015), we know that individuals with an intermediate secondary school leaving certificate (Mittlerer Schulabschluss) also apply for the selected apprenticeship positions. Our FE applicants competed with these individuals, potentially affecting their likelihood of receiving an invitation. To create a comparable pool of competitors in the FS, we constructed additional applicants with an intermediate secondary school leaving

Table 1: Experimental dimensions and levels

Dimension	Field Experiment (FE)	Factorial Survey (FS)
Ethnic origin	0 = German (e.g., Julia Fischer) 1 = Turkish (e.g., Zeynep Yilmaz)	0 = German (e.g., Anna Wagner) 1 = Turkish (e.g., Ayse Sahin)
Education	0 = Intermed. high school degree 1 = Upper sec. high school degree (Abitur) 2 = Upper sec. high school degree (Abitur) + some college without degree	0 = Intermed. high school degree 1 = Upper sec. high school degree (Abitur) 2 = Upper sec. high school degree (Abitur) + some college without degree
Gender	0 = Male (e.g., Mehmet) 1 = Female (e.g., Zeynep)	0 = Male (e.g., Ahmet) 1 = Female (e.g., Ayse)

Additional dimensions were school grades and social origin (fixed in FE and varied in FS). In the FE, there were no applicants with intermediate high school degree. More information on these considerations can be found in the online supplement 1.

certificate. In the FE, we fixed these characteristics to avoid underpowering the experiment. To enhance the realism of the vignettes and reduce repetitiveness among the respondents, we varied school grades and other attributes in the FS. We also randomly varied some characteristics (e.g., hobbies) while keeping the level constant. The online supplement 1.3 provides an overview of these additional dimensions and their coding.

This experimental variation resulted in an orthogonal full factorial design of 144 vignettes, which were allocated in 18 sets of eight vignettes following principles of d-efficiency, as recommended by Auspurg and Hinz (2015). We randomly assigned one of these sets to each recruiter.

SDB and IER

Following the vignette ratings, we measured a respondent's disposition to provide socially desirable answers (H2) via the Social Desirability-Gamma Short Scale (KSE-G) developed by Kemper et al. (2012). The scale consists of six 5-point Likert scale items that capture whether a respondent tends to deny socially undesirable impulses and attribute positive traits to themselves (e.g., "In an argument, I always remain factual and objective"; Cronbach's $\alpha = 0.64$).⁴

We measured IER (H3) in two ways. First, we used the average time spent evaluating the vignettes, assuming that shortened response time indicated an absence of cognitive processing (Huang et al., 2012). Second, we used a shortened version of the attitudes to surveys scale of Stocké (2003) with six 5-point Likert scale items that gauged the effort respondents put into answering (e.g., "I usually try very hard to answer correctly in surveys"; Cronbach's $\alpha = 0.66$). We standardized the SDB and IER scales to zero mean and unit variance.

Table 2 provides descriptive statistics. It indicates that the four occupational fields are not equally sized due to the varying number of advertised positions among fields. However, it also shows little selectivity regarding the occupational field in the FS compared with the FE. In the survey, we measured respondents'

Table 2: Descriptive statistics

	Field Experiment (FE)			Factorial Survey (FS)		
	Mean	SD	N	Mean	SD	N
Average invitation probability	0.54	0.50	3,002	0.59	0.49	3,840
Occupational field			3,002			3,840
Electronics	0.22		655	0.13		512
Laboratory	0.13		398	0.15		560
Administration	0.42		1,261	0.50		1,944
Media	0.23		688	0.22		824
Responsible for candidate selection						
Respondent alone	—		—	0.16		632
Respondent and colleagues	—		—	0.78		2,984
Colleagues	—		—	0.06		224
Social desirability scale (standardized)	—	—	—	0	1	3,840
Attitudes to survey scale (standardized)	—	—	—	0	1	3,840
Average processing time per vignette (in sec)	—	—	—	30.78	11.63	3,840

role in the hiring process. Fortunately, the results show that only six percent of all respondents have colleagues making the hiring decisions, whereas 94 percent of respondents are solely or jointly responsible for making hiring decisions. Thus, we reached our target respondents with the FS.

Analytical Strategy

To assess whether behavioral intentions stated in FSs correspond more closely to actual behavior for less sensitive topics (H1), we compared the effects of applicants' education and ethnicity on invitation probability between the FE and FS using linear probability models with a dichotomous measure of invitation probability as the dependent variable. Next, we assessed the significance of the difference between the two experimental conditions by pooling the data and calculating interaction effects between the variables and a dummy variable for the experimental condition. All other applicant characteristics (e.g., gender) were independent of the two treatments by design (however, we included them in all regression models to improve statistical precision). We also included controls for the occupational field and a dummy variable for wave, as we submitted applications in both the spring and fall of 2022. To account for the nested data structure in the FS, we clustered standard errors at the respondent level. Sensitivity analyses using logistic and multilevel random-effect regressions yielded substantively consistent results (see the online supplement 4.7).

To examine whether an FS estimate better reflects reality when respondents have a low disposition to provide socially desirable answers (H2), we categorized respondents into three groups—low, intermediate, and high disposition—by dividing our FS sample into three groups at the 33rd and 66th percentiles (cut-off points: $-0.29, 0.53$). Although the cut-off points are arbitrary, different cut-off points led to

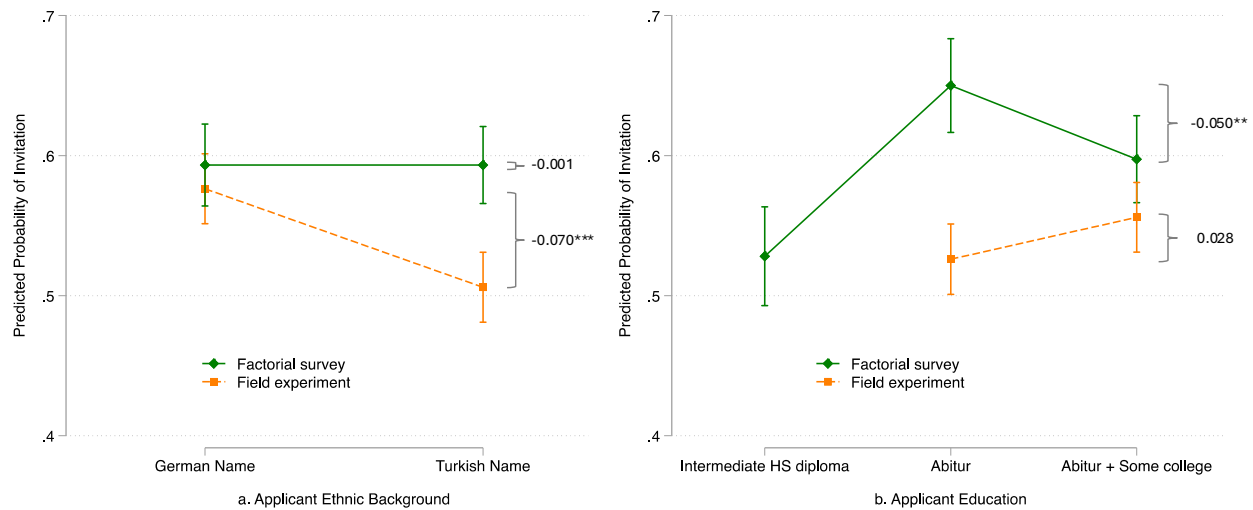


Figure 3: Predicted probabilities of invitation by applicant ethnic background and education. *Note:* Based on models in the online supplement 2, Table S7. Covariates: Other applicant characteristics, occupational field, and wave. All covariates are fixed at the mean. $N_{FE} = 3,002$; $N_{FS} = 3,840$.

the same finding. Then, we estimated the FS model separately for these subgroups and checked if those respondents with low SDB came closer to the “true” effect.

For H3, we employed a similar approach by examining whether respondents who invested more time in the survey (i.e., put in more effort) corresponded more closely to the behavioral benchmark. We divide the FS sample into respondents with low, intermediate, and high response times by splitting the sample at the 33rd and 66th percentiles (cut-off points: 24.50, 33.25).⁵ Finally, we also formed three groups based on the attitudes to surveys scale, which is our alternative measure of survey effort, by splitting our FS sample at the 33rd and 66th percentiles (cut-off points: -0.26, 0.42).

Results

Topics with Varying Sensitivity (H1)

Figure 3 displays the effects of applicants’ ethnic background and education on invitation probability (for the corresponding regression tables, see the online supplement 2). The figure shows the predicted probabilities of invitation, along with 95 percent confidence intervals, from both the FE (dashed orange line) and the FS (solid green line).

Figure 3a illustrates the differing results for ethnic background between the FE and FS. In the FE, we found a negative effect of having a Turkish name, amounting to a difference of seven percentage points, indicating that applicants of Turkish origin were significantly less frequently invited for an interview or assessment than native German applicants. This effect is consistent in both size and direction with previous FE research on this topic in Germany (Schneider et al., 2014). However, in

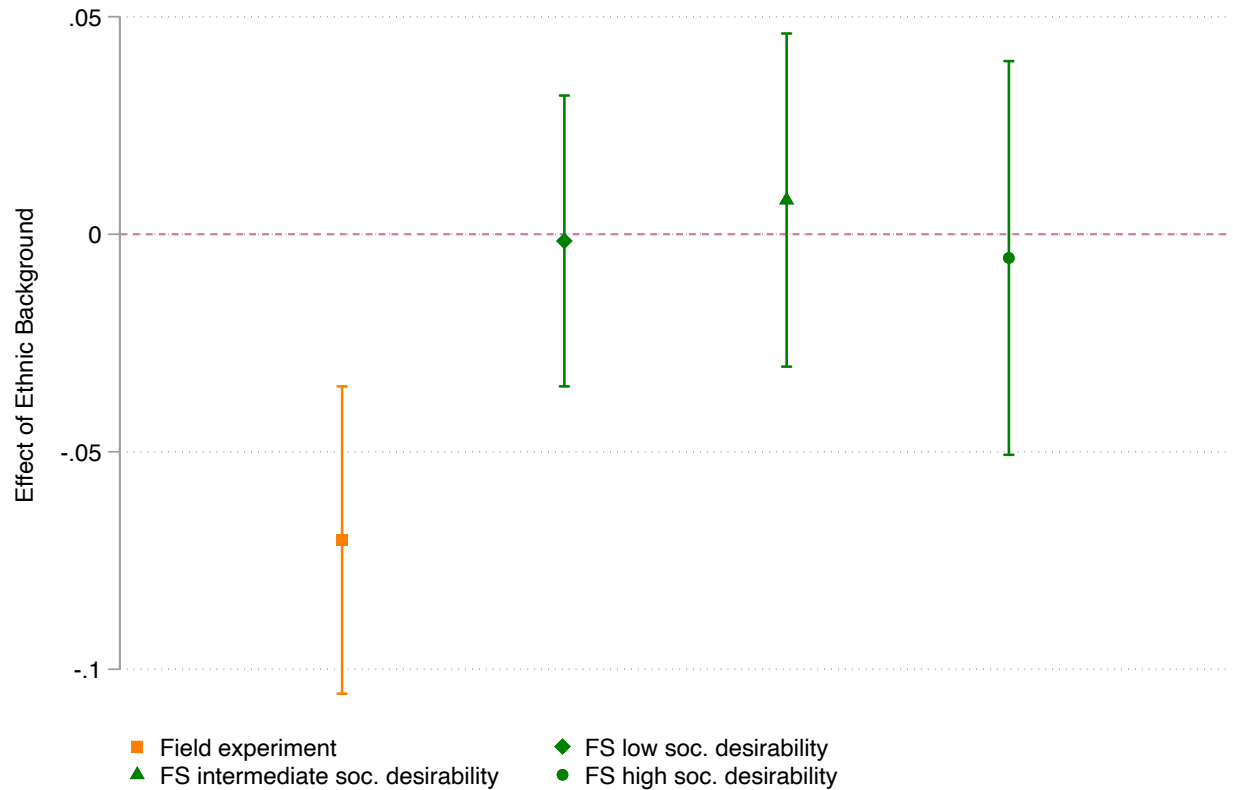


Figure 4: Coefficient plot of the effect of migration background for the FE and three levels of disposition for social desirability in the FS. *Note:* Based on models in the online supplement Table S8. $N_{FE} = 3,002$; $N_{FSlow} = 1,600$; $N_{FSintermed} = 1,272$; $N_{FShigh} = 968$.

the FS, we found no significant difference in invitation probability between German and Turkish names. The disparity between the coefficients in the two experiments was statistically significant at the 0.01 level ($b = 0.070$, $SE = 0.023$).

In contrast to ethnicity, evaluating applicants based on their educational trajectory is legitimate and socially appropriate. Figure 3b displays a positive, albeit statistically insignificant, advantage of 2.8 percentage points in the FE for individuals with an upper-secondary education (Abitur) and some college compared to those with Abitur only. However, we do see a negative and significant effect of -5 percentage points for the FS. Here, applicants with profiles with some college are less often invited than those with Abitur only. The difference between the two experimental conditions is significant at the 0.01 level ($b = -0.080$, $SE = 0.026$). This result is somewhat surprising and does not align with previous FS research on the topic, which found a null effect for college non-completion versus Abitur only, albeit for different occupations (Neugebauer and Daniel, 2022). In addition, in the FS, those with only an intermediate high school diploma had a theoretically plausible significant disadvantage (12.2 percent points) compared to those with Abitur.

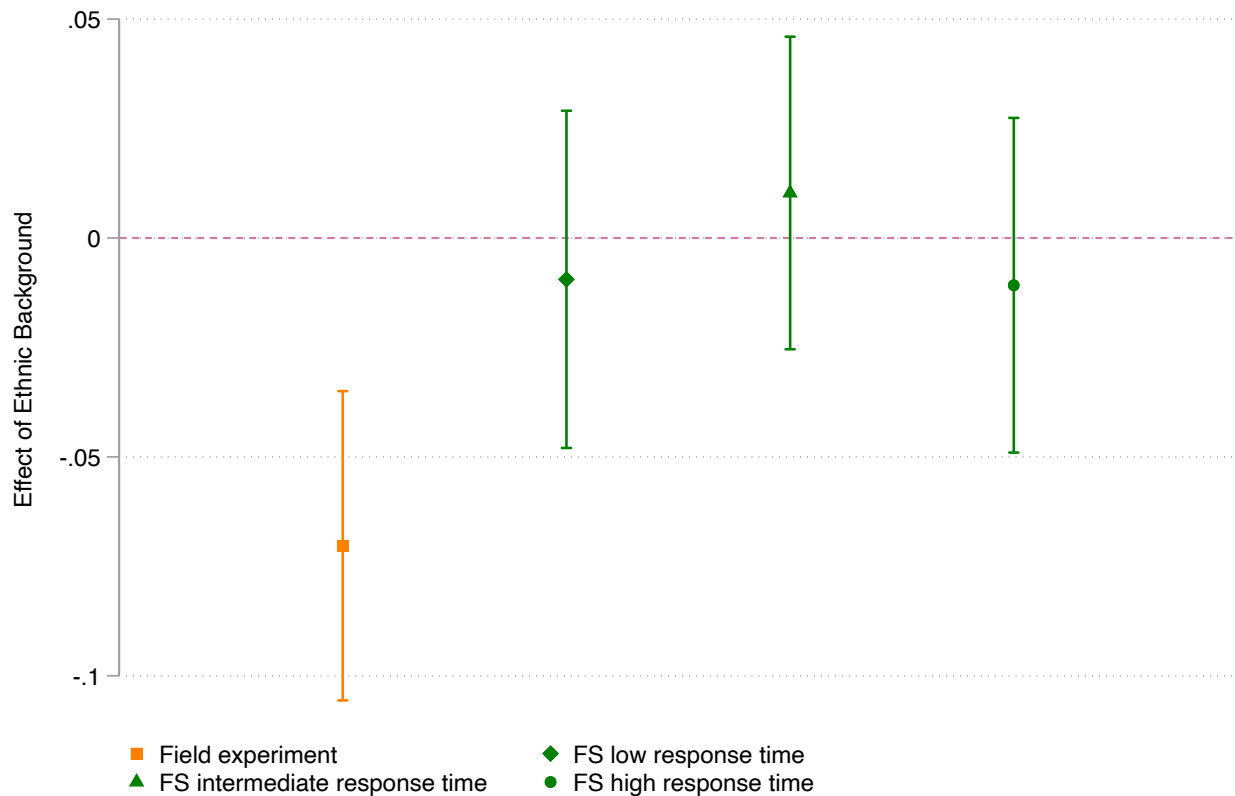


Figure 5: Coefficient plot of the effect of migration background for different levels of response time in the FS. Note: Based on models in the online supplement Table S9. $N_{FE} = 3,002$; $N_{FSlow} = 1,280$; $N_{FSintermed} = 1,264$; $N_{FShigh} = 1,296$.

Overall, we must reject H1. Although we expected biased results for ethnic discrimination, we also anticipated a higher alignment of the FS with the “true” effect of the FE in relation to college non-completion. However, our results suggest that replicability through the FS is poor regardless of the topic’s sensitivity.

Disposition for Social Desirability (H2)

Our study also asked whether FSs are close to actual behavior when surveying respondents with a low inclination to respond in socially desirable ways. Figure 4 provides an answer to this question by showing the effect of ethnic background in the FE and for three groups of respondents in the FS: those with low, intermediate, and high SDB (we concentrate on ethnicity here; analogous results for college non-completion, yielding the same conclusion, can be found in the online supplement 3).

For all three groups of respondents in the FS, the effect of ethnic discrimination is insignificant, in contrast to its negative effect of -0.07 in the FE. We did not find a higher level of correspondence between FS and FE results for respondents with a low inclination to provide socially desirable responses; therefore, we must reject H2.

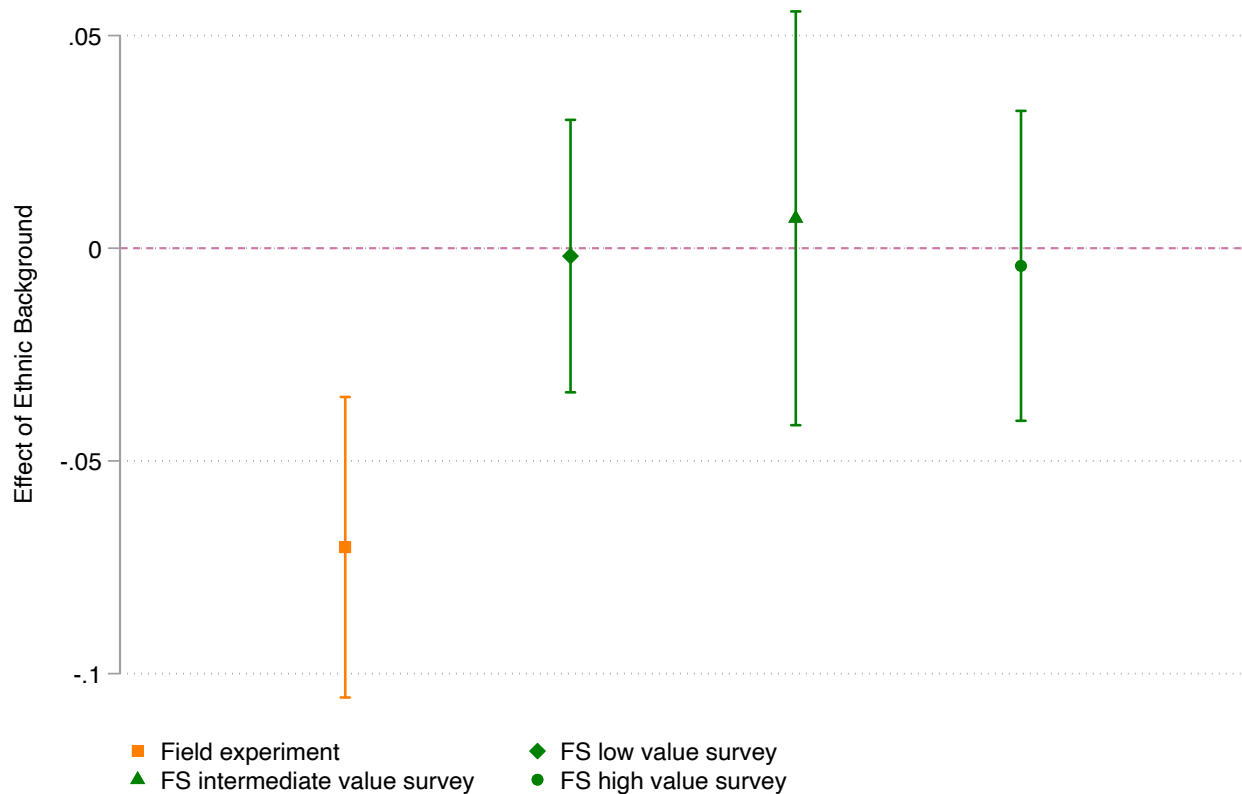


Figure 6: Coefficient plot of the effect of migration background for different levels of survey appreciation in the FS. *Note:* Based on models in the online supplement Table S10. $N_{FE} = 3,002$; $N_{FSlow} = 1,720$; $N_{FSintermed} = 936$; $N_{FShigh} = 1,184$.

Insufficient Effort Responding (H3)

Finally, we tested whether FSs correspond more closely to actual behavior when respondents put more effort into answering the survey. Again, we show only the analyses for the effect of ethnic discrimination here; the analyses regarding college non-completion are reported in the online supplement 3. We begin with the vignette response time as a measure of effort. Figure 5 shows the results for the FE and the groups of respondents in the FS with low, intermediate, and high levels of response time, that is, effort. The ethnic discrimination effect of seven percentage points from the FE could not be replicated in any of the three groups.

The same pattern emerged when relying on our alternative effort measurement based on the attitudes to surveys scale. Figure 6 shows that the effect of ethnic background remains insignificant and close to zero for respondents who put in low, intermediate, and high efforts. Based on these findings, we must reject H3: We did not find a higher correspondence between the FE and the FS results for respondents who showed a high level of effort.

Robustness Checks

The results presented so far cast substantial doubt on the accuracy of FSs in predicting real-world hiring behavior. Several possible objections to our design remain; however, we addressed these through various sensitivity analyses reported in more detail in the online supplement. None of these sensitivity analyses altered the conclusions from our primary analyses:

1. In our FS design, each employer received a set of eight vignettes, whereas in the FE, we sent only one application to each employer. One might suspect that, due to this repetition, respondents would be more likely to discover the topic behind the experimental manipulation (i.e., ethnic background), which may have increased SDB. To test whether this occurred, we ran an analysis using only the first vignette for each FS respondent (online supplement 4.1).
2. One plausible source of measurement error arises from cases in which the FS respondent was different from the person reviewing the real-world application. If hiring decisions vary between managers within firms, this may lead to disagreement between the FS and FE. To address this concern, we restricted the sample to respondents who indicated they were solely responsible for hiring apprentices (online supplement 4.2).
3. It is possible that not all employers perceived our applicant profiles as realistic, which may have led to biased responses. To test this, we restricted the sample to recruiters who indicated in the survey that our applicant profiles closely resembled the typical applicants they encounter in real world (61 percent) (online supplement 4.3).
4. In the FS, we varied two additional dimensions: socioeconomic status and achievement. This additional variation could have altered employers' evaluation of the profiles. To address this concern, we restricted the FS sample to vignettes showing applicants with intermediate SES and achievements, as in the FE (online supplement 4.4).
5. We explored various specifications of the dependent variable in the FE (different categorization of reactions of employers) and the FS (continuous outcome) (online supplement 4.5).
6. Sample selection bias could also cause the FE and FS results to differ, independent of the behavioral validity (Berk, 1983). To determine whether the weak alignment between the two experiments is due to a lack of behavioral validity or selection bias, we restricted the FE sample to those respondents who also participated in the FS to ensure that we considered exactly the same respondents in both experiments (online supplement 4.6).

None of the listed sensitivity analyses resulted in a stronger alignment of the FS with the behavioral benchmark of the FE.

Summary and Discussion

Scholars have long aimed to predict human behavior through surveys or experiments. Recently, FSs, in which researchers present respondents with multidimensional decision situations, have been used more and more frequently for this purpose (Treischl and Wolbring, 2022). However, few studies have examined how well answers to hypothetical vignette scenarios allow us to predict decisions in real-world settings.

In this study, we tested the predictive validity of FSs through a prominent use case: employer hiring decisions. A first key contribution is that we contrasted the hypothetical behavior in the FS with a maximally similar real-life situation to rule out alternative explanations for deviating results. A second key contribution is that, to our knowledge, our study is the first to investigate the conditions under which FSs allow valid behavioral predictions. We aimed to identify situations in which FSs allow good predictions and can thus replace more costly or ethically questionable procedures. We hypothesized that researchers should design the vignette scenarios realistically, limit themselves to non-sensitive topics, and identify respondents who are motivated and likely to provide honest responses.

Our findings are sobering. In our study, the behavioral intentions stated in the FS did not translate to actual behavior, rendering the self-reports in the FS unsuitable for predicting actual behavior. This proved true not only for the sensitive topic of ethnic hiring discrimination but also for a non-sensitive characteristic (education). Moreover, the behavioral prediction did not improve when we restricted the FS sample to respondents who showed a low propensity for socially desirable responses or who tried to answer the FS accurately.

This aligns with two previous validation studies in the field of hiring behavior. Pager and Quillian (2005) found that employers who indicate openness to hiring ex-offenders in a telephone survey do not “walk the talk” when confronted with applicants with a criminal record in an FE. Wulff and Villadsen (2020) reached similar conclusions compared to our study concerning ethnic discrimination, which they found to be present in an FE but which did not replicate in an FS conducted with an online access panel. To exclude the possibility that this failure to replicate results in an FS stemmed from a poor equivalence of FS and FE, our FS design improved upon these previous studies. We put great care and deliberation into developing the materials, ensuring a high level of psychological realism and a strong alignment between FS and benchmark FE. We also ensured that both experiments involved the same employers. However, even with this maximally aligned design, we did not achieve validity of the FS results. Therefore, given what we can deduce from our study, we must conclude that FSs are not suited for studying hiring behavior. Moreover, we must carefully re-evaluate the conclusions from hiring studies that use FSs as their method of analysis.

Our article contributes to the debate on whether we can predict behavior through surveys or experiments (Barabas and Jerit, 2010; Bertrand and Mullainathan, 2001; Charness and Fehr, 2015; Hainmueller et al., 2015). Previous contributions have mainly focused on the generalizability of laboratory experiments (Levitt and List, 2007) and, more recently, discrete choice experiments (Quaife et al., 2018). We

extend this literature with a systematic validation of FS, a method that is currently widely used in sociology and related sciences.

Naturally, this study has scope limitations, as it was conducted at a particular time and in a particular space. It leaves open the question of whether the reported findings can be generalized to other hiring or decision situations. We can only speculate that FSs may work better for low-cost decisions such as forwarding an email (see e.g., Petzold and Wolbring, 2019), as opposed to inviting an applicant for a job interview, which involves relatively high costs of time and monetary resources for a firm. Another aspect of the experimental design that deserves more attention in future research is the psychological realism of the FS materials. Our study aimed to maximize realism; however, the potential effect of less realistic materials is not entirely clear. A few studies have dealt with this topic, but conclusive results have not yet been obtained (see Gutfleisch et al., 2021). In sum, cumulative research and additional validation attempts are required to obtain more robust answers.

FSs were developed to assess how people form beliefs, normative judgments, and attitudes (Auspurg and Hinz, 2015; Jasso, 2006); they were not developed to approximate actual behavior. When used for their original purpose, FSs deliver essential insights into sociologically relevant judgment principles (Auspurg, Hinz, and Sauer, 2017). We suggest continuing to use FSs for these types of questions. Although we acknowledge the limitations of any single study, our findings serve as a cautionary note on the behavioral validity of FS experiments, challenging their growing popularity.

Notes

- 1 The data collection process and the hypotheses for both of our experiments are preregistered at the Open Science Framework: <https://osf.io/2jrgu/>.
- 2 In addition, the measured behavior diverges in Pager and Quillian's (2005) seminal study, in which they compared employers' willingness to hire ex-offenders, as stated in a telephone vignette and an actual hiring situation. While the vignette asked employers to rate their likelihood of hiring the applicant, the actual hiring study measured whether the applicant was invited back for an interview—two behaviors with potentially different decision processes.
- 3 Individuals with an intermediate secondary school certificate also apply for these positions. As we explain further below, we took this into account in our design.
- 4 Kemper et al. (2012) suggest splitting the items into two separate factors. In our study, analyses using this approach yielded the same results as combining all six items into one scale.
- 5 To exclude cases in which the respondent might have been distracted from the web survey, we omitted vignettes for which the respondent stayed on a page for more than 120 sec. Naturally, it is challenging to determine whether a respondent was distracted (e.g., by talking to a colleague) or if they were carefully considering their decision on the vignette. Therefore, we tested different cut-off points ranging from 20 to 1,000 sec. The choice of cut-off point did not affect the results. These additional analyses are available upon request.

References

- Ajzen, Icek. 1991. "The Theory of Planned Behavior". *Organizational Behavior and Human Decision Processes* 50(2):179–211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Armitage, Christopher J. and Mark Conner. 2001. "Efficacy of the Theory of Planned Behaviour: A Meta-Analytic Review." *British Journal of Social Psychology* 40(4):471–99. <https://doi.org/10.1348/014466601164939>
- Aronson, Elliot, Timothy D. Wilson, and Marilyn B. Brewer. 1998. "Experimentation in Social Psychology." Pp. 99–142 in *The Handbook of Social Psychology*, Vols. 1-2, 4th ed. New York, NY: McGraw-Hill.
- Auer, Daniel, Giuliano Bonoli, Flavia Fossati, and Fabienne Liechti. 2019. "The Matching Hierarchies Model: Evidence from a Survey Experiment on Employers' Hiring Intent Regarding Immigrant Applicants." *International Migration Review* 53(1):90–121. <https://doi.org/10.1177/0197918318764872>
- Auspurg, Katrin and Thomas Hinz. 2015. *Factorial Survey Experiments*. Vol. 175. Thousand Oaks, CA: Sage Publications. <https://doi.org/10.4135/9781483398075>
- Auspurg, Katrin, Thomas Hinz, Stefan Liebig, and Carsten Sauer. 2014. "The Factorial Survey as a Method for Measuring Sensitive Issues." Pp. 159–72 in *Improving Survey Methods*. London: Routledge. <https://doi.org/10.4324/9781315756288-21>
- Auspurg, Katrin, Thomas Hinz, and Carsten Sauer. 2017. "Why Should Women Get Less? Evidence on the Gender Pay Gap from Multifactorial Survey Experiments." *American Sociological Review* 82(1):179–210. <https://doi.org/10.1177/0003122416683393>
- Baert, Stijn. 2018. "Hiring Discrimination: An Overview of Almost All Correspondence Experiments Since 2005." Pp. 63–77 in *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. M. Gaddis. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-71153-9_3
- Baert, Stijn and Ann-Sophie De Pauw. 2014. "Is Ethnic Discrimination Due to Distaste or Statistics?" *Economics Letters* 125(2):270–3. <https://doi.org/10.1016/j.econlet.2014.09.020>
- Barabas, Jason and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104(2):226–42. <https://doi.org/10.1017/S0003055410000092>
- Barr, Abigail, Tom Lane, and Daniele Nosenzo. 2018. "On the Social Inappropriateness of Discrimination." *Journal of Public Economics* 164:153–64. <https://doi.org/10.1016/j.jpubeco.2018.06.004>
- Berk, Richard A. 1983. "An Introduction to Sample Selection Bias in Sociological Data." *American Sociological Review*, 48(3):386–98. <https://doi.org/10.2307/2095230>
- Bertrand, Marianne and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *American Economic Review* 91(2):67–72. <https://doi.org/10.1257/aer.91.2.67>
- BIBB. 2015. *Datenreport zum Berufsbildungsbericht 2016: Informationen und Analysen zur Entwicklung der beruflichen Bildung*. Bundesinstitut für Berufsbildung (BIBB).
- Bills, David B., Valentina Di Stasio, and Klarita Gërkhani. 2017. "The Demand Side of Hiring: Employers in the Labor Market." *Annual Review of Sociology* 43(1):291–310. <https://doi.org/10.1146/annurev-soc-081715-074255>
- Charness, Gary and Ernst Fehr. 2015. "From the Lab to the Real World." *Science* 350(6260):512–3. <https://doi.org/10.1126/science.aad4343>

- Collett, Jessica L. and Ellen Childs. 2011. "Minding the Gap: Meaning, Affect, and the Potential Shortcomings of Vignettes." *Social Science Research* 40(2):513–22. <https://doi.org/10.1016/j.ssresearch.2010.08.008>
- Crowne, Douglas P. and David Marlowe. 1960. "A New Scale of Social Desirability Independent of Psychopathology." *Journal of Consulting Psychology* 24(4), 349–54. <https://doi.org/10.1037/h0047358>
- Damelang, Andreas and Martin Abraham. 2016. "You Can Take Some of It with You!: A Vignette Study on the Acceptance of Foreign Vocational Certificates and Ethnic Inequality in the German Labor Market." *Zeitschrift für Soziologie* 45(2):91–106. <https://doi.org/10.1515/zfsoz-2015-1005>
- Damelang, Andreas, Martin Abraham, Sabine Ebensperger, and Felix Stumpf. 2019. "The Hiring Prospects of Foreign-Educated Immigrants: A Factorial Survey among German Employers." *Work, Employment and Society* 33(5):739–58. <https://doi.org/10.1177/0950017018809897>
- Di Stasio, Valentina and Herman G. van de Werfhorst. 2016. "Why Does Education Matter to Employers in Different Institutional Contexts? A Vignette Study in England and the Netherlands." *Social Forces* 95(1):77–106. <https://doi.org/10.1093/sf/sow027>
- Diehl, Claudia, Veronika A. Andorfer, Yassine Khoudja, and Karolin Krause. 2013. "Not in My Kitchen? Ethnic Discrimination and Discrimination Intentions in Shared Housing Among University Students in Germany." *Journal of Ethnic and Migration Studies* 39(10):1679–97. <https://doi.org/10.1080/1369183X.2013.833705>
- Eifler, Stefanie. 2007. "Evaluating the Validity of Self-Reported Deviant Behavior Using Vignette Analyses." *Quality & Quantity* 41(2):303–18. <https://doi.org/10.1007/s11135-007-9093-3>
- Eifler, Stefanie. 2010. "Validity of a Factorial Survey Approach to the Analysis of Criminal Behavior." *Methodology* 6(3):139–46. <https://doi.org/10.1027/1614-2241/a000015>
- Findley, Michael G., Brock Laney, Daniel L. Nielson, and J. C. Sharman. 2017. "External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation." *The Journal of Politics* 79(3):856–72. <https://doi.org/10.1086/690615>
- Groß, Jochen and Christina Börensen. 2009. "Wie valide sind Verhaltensmessungen mittels Vignetten? Ein methodischer Vergleich von faktoriellem Survey und Verhaltensbeobachtung." *Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen*, 149–78. https://doi.org/10.1007/978-3-531-91380-3_7
- Gutfleisch, Tamara, Robin Samuel, and Stefan Sacchi. 2021. "The Application of Factorial Surveys to Study Recruiters' Hiring Intentions: Comparing Designs Based on Hypothetical and Real Vacancies." *Quality & Quantity* 55(3):775–804. <https://doi.org/10.1007/s11135-020-01012-7>
- 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):2395–400. <https://doi.org/10.1073/pnas.1416587112>
- Huang, Jason L., Paul G. Curran, Jessica Keeney, Elizabeth M. Poposki, and Richard P. DeShon. 2012. "Detecting and Detering Insufficient Effort Responding to Surveys." *Journal of Business and Psychology* 27(1):99–114. <https://doi.org/10.1007/s10869-011-9231-8>
- Jasso, Guillermina 2006. "Factorial Survey Methods for Studying Beliefs and Judgments." *Sociological Methods & Research* 34(3):334–423. <https://doi.org/10.1177/0049124105283121>

- Kemper, Christoph J., Constanze Beierlein, and Doreen Bensch. 2012. "Eine Kurzsкала zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens." *Technical report*. GESIS.
- Klüver, Heike, Felix Hartmann, Macartan Humphreys, Ferdinand Geissler, and Johannes Giesecke. 2021. "Incentives Can Spur COVID-19 Vaccination Uptake." *Proceedings of the National Academy of Sciences* 118(36):e2109543118. <https://doi.org/10.1073/pnas.2109543118>
- Krumpal, Ivar. 2013. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review." *Quality & quantity* 47(4):2025–47. <https://doi.org/10.1007/s11135-011-9640-9>
- Levitt, Steven D. and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21(2):153–74. <https://doi.org/10.1257/jep.21.2.153>
- Lippens, Louis, Siel Vermeiren, and Stijn Baert. 2023. "The State of Hiring Discrimination: A Meta-Analysis of (Almost) All Recent Correspondence Experiments." *European Economic Review* 151:104315. <https://doi.org/10.1016/j.euroecorev.2022.104315>
- Mai, Quan D. 2020. "Unclear Signals, Uncertain Prospects: The Labor Market Consequences of Freelancing in the New Economy." *Social Forces*, 99(3):895–920. <https://doi.org/10.1093/sf/soaa043>
- Mize, Trenton D. and Bianca Manago. 2022. "The Past, Present, and Future of Experimental Methods in the Social Sciences." *Social Science Research* 108:102799. <https://doi.org/10.1016/j.ssresearch.2022.102799>
- Mutz, Diana C. 2011. *Population-Based Survey Experiments*. New Jersey: Princeton University Press. <https://doi.org/10.23943/princeton/9780691144511.001.0001>
- Neugebauer, Martin and Annabell Daniel. 2022. "Higher Education Non-Completion, Employers, and Labor Market Integration: Experimental Evidence." *Social Science Research* 105:102696. <https://doi.org/10.1016/j.ssresearch.2022.102696>
- Nisic, Natascha and Katrin Auspurg. 2009. "Faktorieller Survey und klassische Bevölkerungsumfrage im Vergleich - Validität, Grenzen und Möglichkeiten beider Ansätze." Pp. 211–45 in *Klein aber fein!: Quantitative empirische Sozialforschung mit kleinen Fallzahlen*, edited by P. Kriwy and C. Gross. Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91380-3_9
- Pager, Devah and Lincoln Quillian. 2005. "Walking the Talk? What Employers Say Versus What They Do." *American Sociological Review* 70(3):355–80. <https://doi.org/10.1177/000312240507000301>
- Paulhus, Delroy L. 2002. "Socially Desirable Responding: The Evolution of a Construct." Pp. 49–69 in *The Role of Constructs in Psychological and Educational Measurement*, 0 ed. Mahwah, NJ: Erlbaum.
- Petzold, Knut and Tobias Wolbring. 2019. "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. <https://doi.org/10.1027/1614-2241/a000161>
- Protsch, Paula and Heike Solga. 2015. "How Employers Use Signals of Cognitive and Noncognitive Skills at Labour Market Entry: Insights from Field Experiments." *European Sociological Review* 31(5):521–32. <https://doi.org/10.1093/esr/jcv056>
- Quaife, Matthew, Fern Terris-Prestholt, Gian L. Di Tanna, and Peter Vickerman. 2018. "How Well Do Discrete Choice Experiments Predict Health Choices? A Systematic Review and

- Meta-Analysis of External Validity." *The European Journal of Health Economics* 19(8):1053–66. <https://doi.org/10.1007/s10198-018-0954-6>
- Rogelberg, Steven G., Gwenith G. Fisher, Douglas C. Maynard, Milton D. Hakel, and Michael Horvath. 2001. "Attitudes toward Surveys: Development of a Measure and Its Relationship to Respondent Behavior." *Organizational Research Methods* 4(1):3–25. <https://doi.org/10.1177/109442810141001>
- Rossi, Peter H., William A. Sampson, Christine E. Bose, Guillermina Jasso, and Jeff Passel. 1974. "Measuring Household Social Standing." *Social Science Research* 3(3):169–90. [https://doi.org/10.1016/0049-089X\(74\)90011-8](https://doi.org/10.1016/0049-089X(74)90011-8)
- Schneider, Jan, Ruta Yemane, and Martin Weinmann. 2014. "Diskriminierung am Ausbildungsmarkt: Ausmaß, Ursachen und Handlungsperspektiven." *SVR*.
- Schwarz, Susanne and Alexander Coppock. 2022. "What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments." *The Journal of Politics* 84(2):655–68. <https://doi.org/10.1086/716290>
- Stocké, V. 2003. "Einstellungen zu Umfragen." *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. <https://doi.org/10.6102/zis218>
- Treischl, Edgar and Tobias Wolbring. 2022. "The Past, Present and Future of Factorial Survey Experiments: A Review for the Social Sciences." *Methods, Data, Analyses* 16(2):30. <https://doi.org/10.12758/mda.2021.07>
- Van Beek, Krijn W. H., Carl C. Koopmans, and Bernard M. S. Van Praag. 1997. "Shopping at the Labour Market: A Real Tale of Fiction." *European Economic Review* 41(2):295–317. [https://doi.org/10.1016/S0014-2921\(96\)00037-2](https://doi.org/10.1016/S0014-2921(96)00037-2)
- Van Belle, Eva, Valentina Di Stasio, Ralf Caers, Valentina De Couck, and S. Baert. 2018. "Why Are Employers Put Off by Long Spells of Unemployment?" *European Sociological Review* 34(6):694–710. <https://doi.org/10.1093/esr/jcy039>
- Walzenbach, Sandra. 2019. "Hiding Sensitive Topics by Design?: An Experiment on the Reduction of Social Desirability Bias in Factorial Surveys." *Survey Research Methods* 13:103–21. <https://doi.org/10.18148/srm/2019.v1i1.7243>
- Wulff, Jasper N. and Anders R. Villadsen. 2020. "Are Survey Experiments as Valid as Field Experiments in Management Research? An Empirical Comparison Using the Case of Ethnic Employment Discrimination." *European Management Review* 17(1):347–56. <https://doi.org/10.1111/emre.12342>

Acknowledgments: Both authors contributed equally to this study. We would like to thank Lukas Zielinski, Stefan Gunzelmann, Tim Skroblien, Pablo Neitzsch, and Franz Geiger for their help with the design of the experiments and the collection of the data. Furthermore, we would like to thank Katrin Auspurg, Annabell Daniel, Tamara Gutfleisch, Knut Petzold, and Katharina Stückradt as well as 16 professional experts (recruiters and job counselors) for their feedback on our experimental design and materials. Finally, we would like to thank the participants of ECSR 2022, ACES 2022, DGS 2022, the ISOL paper seminar, the Research Colloquium Sociology (University of Bern), and the Research Colloquium Analytical Sociology (LMU Munich) for their feedback on earlier versions of this article. This research was funded by the German Federal Ministry of Education grant number 16PX21011.

Andrea G. Forster: Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands. E-mail: a.g.forster@uu.nl

Martin Neugebauer: Karlsruhe University of Education, Bismarckstr. 10, 76133 Karlsruhe, Germany. E-mail: martin.neugebauer@ph-karlsruhe.de