sociological science

Validating Factorial Survey Experiments: Response to Comment

Andrea G. Forster,[a] Martin Neugebauer[b]

a) Utrecht University; b) Karlsruhe University of Education

In Forster and Neugebauer (2024), we examine to what extent a factorial survey (FS) on invitations of fictitious applicants can replicate the findings of a nearly identical field experiment conducted with the same employers. In addition to exploring the conditions under which FSs provide valid behavioral predictions, we varied the topic sensitivity and tested whether behavioral predictions were more accurate after filtering out respondents who provided socially desirable answers or did not exert sufficient effort in responding to FS vignettes. Across these conditions, the FS results did not align well with the real-world benchmark. We conclude that researchers must exercise caution when using FSs to study (hiring) behavior. In this rejoinder, we respond to the critique of our study by Pickett (2025).

General Remarks

Pickett argues that we dismiss an entire type of experiment while overlooking the fact that the field experiment (FE) and factorial survey (FS) experiment are not sufficiently comparable to allow for a valid assessment of the findings from both. Although Pickett raises some substantive points of criticism, which we will address further below, we argue that his main criticism is flawed in two key ways.

First, we do not dismiss an entire type of experiment. In the discussion section of our article, we explicitly highlight the scope limitations of our study and emphasize that the method might work well in different decision contexts, such as low-cost decisions. We also note that FSs were developed to assess how people form beliefs, normative judgments, and attitudes, and when used for that purpose, they deliver essential insights into sociologically relevant judgment principles (Forster and Neugebauer 2024, p. 902).

Second, our validation is as comparable to real-life situations as the FS method permits, or at least much more so than other available validation studies (Pager and Quillian 2005; Wulff and Villadsen 2020). Pickett's critique focuses on specific design differences between the FE and FS, such as the fact that the applicant profiles in the FS provided less detailed information than in the FE, or that the FS respondents were aware they were part of a study, whereas FE participants were not, or that applicant names and photos were not identical across experiments. Before we go into the details, we feel it is necessary to make a more fundamental comment. FE and FS are different methods and, therefore, they are never exactly the same. Certain characteristics are inherent to the method and changing them would render the entire method pointless. Stronger even, FE and FS necessarily have to be different in certain respects to work as a standalone method and to make them contenders

Citation: Forster, Andrea G., Martin Neugebauer. 2024. "Validating Factorial Survey Experiments: Response to Comment" Sociological Science 12: 106-114.

Received: December 6, 2024

Accepted: December 6, 2024

Published: January 27, 2025

Editor(s): Arnout van de Rijt

DOI: 10.15195/v12.a6

A male applicant with a Turkish name and an upper secondary school degree (Abitur) is applying for an apprenticeship with your company. His academic performance is average.

Figure 1: Example vignette from a typical FS.

for a validation study in the first place. For example, an FE where respondents are informed about fictitious applications beforehand is as nonsensical as the idea that survey participants are unaware that they are taking part in a survey. Similarly, the adaption of reality in an FS is unavoidable as the instrument has to be functional in a survey setting. For instance, a vignette should fit on a computer screen and should not overwhelm respondents with irritatingly detailed information such as the exact postal address and telephone number. On the other hand, in an FE, an application must be realistic enough to "pass" as a real application, that is, it must necessarily include an address and telephone number. In our FS study, we strived to replicate all the key elements of a real application process as accurately as possible, in contrast to typical vignette studies, which are usually based on brief textual descriptions such as shown in Figure 1.

We discuss the reduction of hypothetical bias that is achieved by our study in detail in the article. However, there is a limit to perfect alignment between the two methods. In this sense, Pickett's critique is akin to criticizing a painter for not making their painting look like a photograph. Our FS "painting" is a significant improvement over previous attempts, however, we do not aim for an exact representation of reality, as then we would need to take a photograph, as shown in the analogy in Figure 2. Equivalently, if we do not run an FS but two FEs, we cannot validate an FS.

Detailed Reply Regarding Specific Design Choices

In the following sections, we take a closer look at the individual points of criticism that were raised by Pickett (2025).

Included Factor Levels in FE and FS

Pickett criticizes that our two experiments include different factor levels. Some factors such as socioeconomic status (SES) and achievement are fixed in the FE, whereas they vary in the FS. Education, one of our dimensions of interest, has two levels in the FE (Abitur, HE dropout), whereas in the FS a third level (intermediate high school degree) is added. Differences in factor levels can be problematic, as the effect of a factor also depends on the number of its levels. Moreover, the effect in an FS depends on the distribution of the other factors.

Although this argument has some validity, it is important to acknowledge that it is impossible to gain control over the factor levels within the FE. One inherent difference between FEs and FSs is that in an FE, fictitious applicants compete with other real-world applicants about whom we can make assumptions but whose entering in the selection process we cannot control. For this reason, an exact match

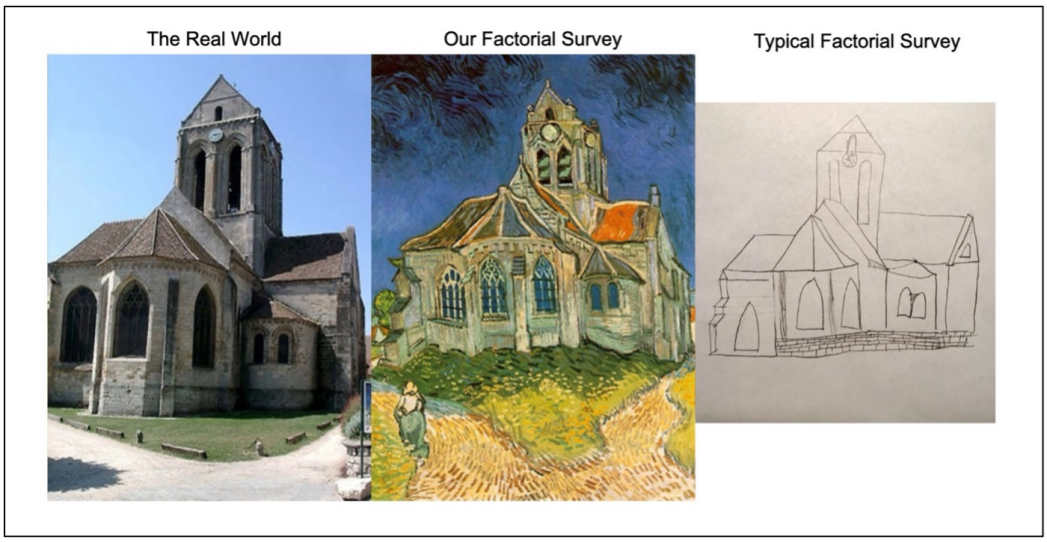

Figure 2: Illustration based on Vincent van Gogh's The Church at Auvers. Sources: https://www.restinpieces.co.uk/blogs/news/real-life-locations-world-famous-paintings (left and middle panel) and 9-year-old Mathilda Neugebauer (right panel)

of the factor levels in both experiments would not solve the problem: The applicant pools remain different and consequently the two empirical estimands are also different. In our view, this is another problem with FSs when used to estimate real-world behavior: The levels of the real-life decision factors can hardly be determined and reproduced exactly.

The best thing to do is to map the real distribution as accurately as possible with rich contextual knowledge. This is exactly what we have done. As we explain in the article: "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 certificate." (p. 893)

A design alternative–which, as we described above, would not have solved the problem any better–would have been to implement all 108 variations[1] of the FS in the FE as well. However, this would have meant that we would have needed a much larger FE sample to achieve the same power. As it was already necessary to send out 3,000 applications (which took 10 months) for the smaller number of ($2 \times 2 \times 2 = 8$) variations[2] to achieve sufficient power, we decided against this alternative strategy.

Applicant Characteristics

Another point of criticism is that we do not use exactly the same names and photos in our experiments and that we do not provide evidence that the used names and photos are information equivalent.

Summarizing our approach regarding applicant characteristics such as photos or names, we aimed to send applications with nearly identical characteristics. The reason why we did not send completely identical applicant profiles across FE and FS is that we did conduct both experiments with the same employers. Sending completely identical applicant profiles would have raised suspicion (the employers would have noticed that they had been part of an FE earlier when seeing the FS).[3]

Regarding names, Pickett mainly criticizes that those simultaneously convey information about other characteristics (e.g., age, ethnicity, gender, SES, and religion) that may impact the outcome. Although this is an important concern, both our FS and FE included explicit markers of other key applicant characteristics such as age, school achievement, hobbies, and also SES (signaled through their parents' occupations). This should greatly diminish the problem of information equivalency by experimentally controlling possible signals that a name can send over and above ethnicity and gender. Including information about SES may seem peculiar to the international reader, but it is still a common habit in Germany to include the names and occupations of parents on the CV when applying for apprenticeship positions. In addition, we chose names that send unclear social class signals. Following other research in the German context (see e.g., Wenz and Hoenig 2020), we opted for "modern" names (e.g., Tim), which are less clear in terms of social background signals compared to Anglo-American sounding names (e.g., Kevin, Jacqueline), which are more likely to be associated with the lower class and traditional first names (e.g., Maximilian, Sophie) that tend to signal a higher educational level of the family. Finally, we aimed to choose very common first and last name combinations for the FE and–to ensure alignment of the experiments–also for the FS as we wanted to make it difficult for employers to make investigations online about the fictitious candidates. The final list consisted of six common, modern German first names (Tim, Niklas, Leon, Anna, Lea, and Julia) and six Turkish first names (Can, Ahmet, Mehmet, Ayse, Hatice, and Zeynep) (selection based on Destatis 2005; Rodriguez 2010, which were combined with typical last names: Fischer, Wagner, Schmidt, Schneider, and Weber [German names] and Yilmaz, Demir, Sahin, Özdemir, and Yildirim [Turkish names]).

Regarding the photos, we provide detailed information on the selection process in the article and the online supplement, but we are happy to spell this out in more detail. Because it is common practice in Germany to include a photo with application materials, a total of 10 portrait photos (eight for the FS and two for the FE) were needed to equip the application documents and vignettes with different photos. We first considered using pretested photos from a face database (e.g., Ma et al. 2015). However, we wanted to use photos that were perceived as equally German and Turkish, so that names and photos would not be confounded, and such photos were not available. Hence, we created our own photos. To obtain suitable photos depicting five male and five female applicants, we initially selected 30 stock photos that were considered similar by three evaluators (research assistants). These photos all showed individuals with similar facial expressions who appeared to be around 19 years old. The photos were edited to show a consistent frame. Subsequently, we conducted a pretest with these 30 photos using a sample of 100 respondents that we accessed via the online panel Prolific. Table 1 shows the

Table 1: Intro and questions from the photo pretest (translated from German original).

Introduction: Below you will see photos of different people. For each person, we ask for your subjective assessment of a number of characteristics. Please respond as spontaneously as possible. There are no "right" or "wrong" answers; what matters are your personal associations!	
Question	**Answer Options**
How old do you think the person is?	Open question
The ethnic or cultural background of a person is often not easily recognizable. How would you most likely categorize this person?	5-point Likert scale ranging from 1 (clearly Turkish) to 5 (clearly German)
How attractive is this person to you?	10-point Likert scale ranging from 1 (not at all attractive) to 10 (very attractive)
How intelligent does this person appear?	10-point Likert scale ranging from 1 (not at all intelligent) to 10 (very intelligent)
How friendly does this person appear?	10-point Likert scale ranging from 1 (not at all friendly) to 10 (very friendly)

introduction of the pretest survey as well as the questions and answer options (translated from German).

We showed 15 randomly chosen photos to each respondent and had them rated on the dimensions age, ethnicity, attractiveness, intelligence, and friendliness. The 10 photos were selected for the experiments that came closest to the desired age (19 years old) and were as close as possible to the mean on the other four dimensions.[4] Two of these photos were randomly chosen for the FE. Both were edited in Photoshop to add the same simple black shirt, ensuring that they were as similar as possible in terms of clothing as well. The applicants in the FS wore slightly different but similarly unobtrusive clothing (e.g., simple blue shirt). The allocation of the photos in the FS was randomized across respondents.

Finally, to check how photos and names affected our conclusions, we can add controls for the specific first names and photos in the FS. Table 2 shows that doing so does not change any of our effects of interest. In the case of names, an incremental F-test between the models with and without a control for names shows that the effects of the specific first names are also jointly zero ($F(5, 479) = 1.88, p = 0.10$). Regarding photos, the incremental F-test is significant ($F(6, 479) = 3.80, p = 0.001$) meaning that there is a non-zero effect of photos on invitation probability. Thus, our selection of photos is not ideal. However, this effect does nothing to change our effects of interest. We do not have any reason to believe that these patterns would be different in the FE or between the two experiments.

Additional Fixed Factors in the FE

Pickett criticizes that we included additional fixed factors in the FE that were not included in the FS, such as a story about relocation intentions, full names of schools and internship companies, and full school certificates. Coming back to the illustration in Figure 2: Equal treatments in two experiments does not necessarily mean that comparability is maximized. There are inherent differences between

Table 2: Adding photo or name as an additional control in our models.

	FS Original Model	FS with Control for Name	FS with Control for Photo
Applicant education (Ref = Abitur)			
intermediate HS	-0.122^{*}	-0.122^{*}	-0.122^{*}
	(0.018)	(0.018)	(0.018)
Abitur + some college	-0.050^{*}	-0.050^{*}	-0.050^{*}
	(0.016)	(0.016)	(0.016)
Applicant Ethnic Background (Ref = German)			
Turkish	−0.001	−0.001	−0.002
	(0.011)	(0.024)	(0.011)
Control for first name	no	yes	no
photo	no	no	yes
Intercept	0.687^{*}	0.700^{*}	0.700^{*}
	(0.049)	(0.051)	(0.051)
Number of observations	3,840	3,840	3,840
R squared	0.16	0.16	0.16

Standard errors in parentheses. Significance levels: $^{*}p < .01, +p < .05$. Models also control for gender, occupational field, achievement, SES, and wave.

FSs and FEs: Application materials in FEs must work in real life–they need to satisfy the demands of a complete and competitive job application. FSs must be sufficiently similar and also fulfil some requirements inherent to a survey, such as not overburdening the respondents.

For example, we added a relocation narrative in the FE that was necessary to do justice to the nature of the local apprenticeship labor market in Germany and to make our experiments practically feasible. In Germany, employers still occasionally reply by mail. Therefore, we had to provide a physical address for our fictitious applicants to which employers could direct reactions. As it was not feasible to acquire access to postal addresses all over Germany, we chose one address in an unobtrusive apartment building in a mixed residential neighborhood in an average German city–Siegen–as the place of living of the applicant. We set up mail forwarding from this address to our university address to obtain all mail that was directed to the applicants at this address. From this address, our applicants applied to positions all over Germany. However, the apprenticeship labor market is very local, meaning that applicants often apply within or around their hometown. To make the application more credible, we added a sentence to the application that motivated the decision to move from Siegen to the location of the apprenticeship position. Saying: "I would like to move back to [location of apprenticeship], to be able to live and work in my hometown again. This is why the position is very attractive to me." In the CV of the applications, it was visible that they were born in the location of the apprenticeship but had lived in Siegen for the past few years. We explored this sentence in the qualitative interviews that we did before sending out the FE and the story was judged as being very credible by all of our 16 respondents. We also tested whether the distance between the address in the town of Siegen and

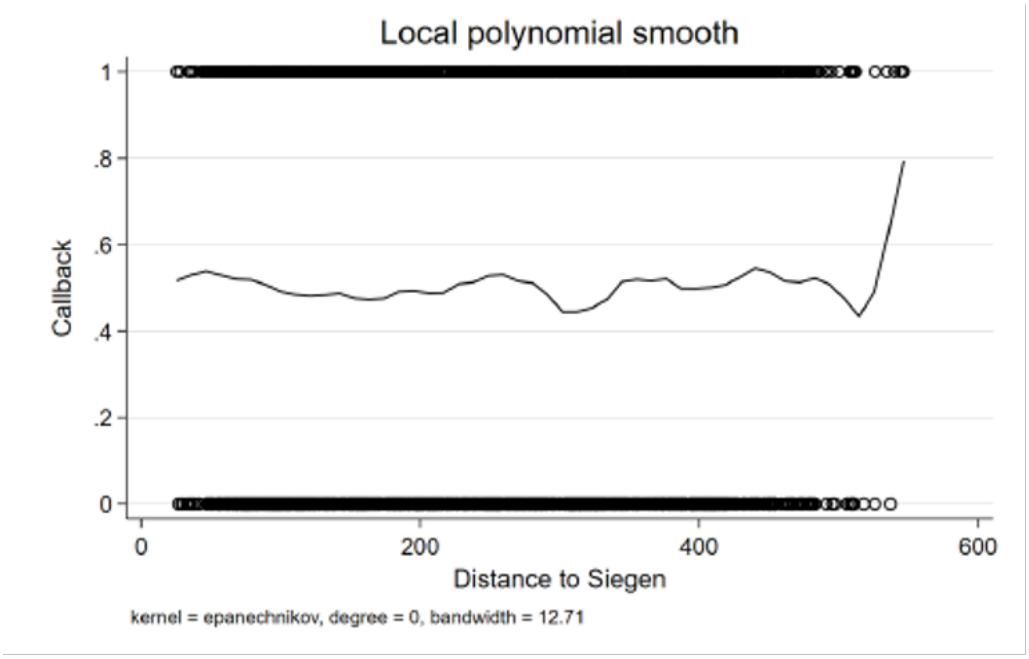

Figure 3: Callback probability by distance between current place of living and apprenticeship location.

the employment location had an effect on the likelihood of being invited. This was not the case (see Figure 3).

On the other hand, adding such a detailed story to eight consecutive vignettes would have drawn too much attention in the FS setting. In oder not to make the FS seem unrealistic, we omitted this story in the vignettes. Providing the same detailed information on the vignettes that we included in the FE would have clearly overburdened the respondents in the online-survey setting.

Experimental Setting

Finally, Pickett criticizes three aspects of the experimental setting. First, he notes that the FS was not anonymous. As we outlined in the introduction to this rejoinder, certain differences between FE and FS are inherent to the method, and an FS can never be anonymous–respondents always know they are in a survey.

Second, Pickett criticizes the use of a sequential design as in the FS employers are "pretested" (they already participated in a previous experiment, namely the FE), whereas they are "unpretested" in the FE. He does not give any indication if a pretesting effect should be expected theoretically nor in which direction the sequential design could influence the outcome. We cannot see why the sequential design should be inferior to other design options–rather we can see many advantages. An alternative would have been to split the sample of recruiters into FE and FS conditions and to send half of the sample an invitation to the FS and the other half a fictitious FE application. Although both designs are valid options in our opinion, we think that it is a stronger test of validity if both experiments are

run with the same employers. Furthermore, the larger sample size in the sequential design increases the power of our experiments.

Third, Pickett mentions a loose set of other design choices that, in his opinion, lead to a possible interaction between specific instruments and the treatment: (1) he notices that the dependent variable (DV) is not aligned in the two experiments. However, we argue that the DV in the FS is modeled exactly after how we can expect employers to look at applications in the real hiring process (FE). In the qualitative pre-study, recruiters told us that they first screen each application before making dichotomous decisions about actual interview invitations. This is precisely how we modeled the FS decision. Furthermore, as Pickett mentions himself, the different measures (scale, dichotomous, only first evaluation) do not show any differences in our robustness checks. (2) Pickett criticizes that the pool of applicants in the FS is restricted and all of them wear the same clothes. This is not correct, as only the two FE applicants (who were never both presented to the same employer) wear the same shirt. In the FS, the clothes vary slightly. (3) Pickett criticizes that FS respondents had to scroll down to find info. This is not true. As can be seen from the vignette example in Figure 1 in Forster and Neugebauer (2024), all info fitted neatly on one screen. Of course, compared to a typical FS, there was more information for the respondents to process. However, as we detailed throughout this rejoinder, we tried to find a good compromise between showing only short text vignettes and presenting full application materials.

Conclusion

As with any research project, we had to weigh up different design alternatives. With the information provided in the previous sections, readers can evaluate our choices for themselves. In our view, Pickett's critique does not alter the conclusions we draw in Forster and Neugebauer (2024).

An FS is inherently different from the real world. Respondents will always know that they are part of a study, and that reality is often more complex. This is precisely the issue of hypothetical bias and social desirability bias (SDB) that we discuss in our article. We made an effort to identify conditions under which FSs can successfully mimic real-world decisions. Our goal was to create a "recipe" to cook up an FS that delivers valid results. Accordingly, our pre-registered hypotheses proposed that FSs would be effective if we excluded participants with high SDB tendencies or focused on non-sensitive topics. However, the sobering findings of our study contradict these hypotheses.

As we have already outlined in the article, our study has scope limitations, as it was conducted at a specific time and within a specific context. It remains an open question whether the reported findings can be generalized to other hiring or decision-making scenarios. However, the study raises fundamental questions about the validity of FSs for measuring behavior, challenging their growing popularity.

We believe that as a scientific community, we can only progress by critically acknowledging the limitations of FSs. As stated in our article, it is essential to continue exploring the boundary conditions under which FSs are appropriate for predicting behavior. We would therefore welcome further validation efforts by

other researchers to examine whether our findings can be generalized to different decision-making contexts, countries, or other settings.

Notes

1 2 (ethnicity) $\times$2 (gender) $\times$3 (education) $\times$3 (achievement) $\times$3 (SES) = 108 profiles. For details, see the online supplement of Forster and Neugebauer (2024).

2 2 (ethnicity) $\times$2 (gender) $\times$2 (education).

3 An alternative design would have been to split the sample of employers into two experimental conditions, which we will detail in the experimental setting section.

4 As the four dimensions not always overlap in distance from the mean, we made qualitative evaluations of combinations of the photos at hand to get to a satisfying set of photos for the experiments.

References

BIBB. 2015. *Datenreport zum Berufsbildungsbericht 2016: Informationen und Analysen zur Entwicklung der beruflichen Bildung*. BIBB.

Destatis. 2005. *Statistisches Jahrbuch 2005.* Wiesbaden: Statistisches Bundesamt.

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

Ma, Debbie S., Joshua Correll, and Bernd Wittenbrink. 2015. "The Chicago face database: A free stimulus set of faces and norming data." *Behavior Research Methods* 47(4):1122–35. https://doi.org/10.3758/s13428-014-0532-5

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

Pickett, Justin T. 2025. "Invalidating Factorial Survey Experiments Using Invalid Comparisons is Bad Practice: Learning From Forster and Neugebauer (2024)." *Sociological Science* 12: 97-105. https://doi.org/10.15195/v12.a5

Rodriguez, Gabriel. 2010. "Turksprachige Namen in Deutschland. Statistik und Tendenzen in der Turksprachigen Vornamengebung." *Namenkundliche Informationen* 97:95–107. https://doi.org/10.58938/ni448

Wenz, Sebastian E. and Kerstin Hoenig. 2020. "Ethnic and Social Class Discrimination in Education: Experimental Evidence from Germany." *Research in Social Stratification and Mobility* 65:100461. https://doi.org/10.1016/j.rssm.2019.100461

Wulff, Jesper 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

Andrea G. Forster: Utrecht University. E-mail: a.g.forster@uu.nl (Corresponding author)
Martin Neugebauer: Karlsruhe University of Education.
E-mail: martin.neugebauer@ph-karlsruhe.de.