sociological science

Invalidating Factorial Survey Experiments Using Invalid Comparisons Is Bad Practice: Learning from Forster and Neugebauer (2024)

Justin T. Pickett

University at Albany, SUNY

Abstract: Forster and Neugebauer's (2024) invalidation study is invalid. Their conclusion that factorial survey (FS) experiments "are not suited for studying hiring behavior" (P. 901) is unjustified, because their claim that they conducted a field experiment (FE) and FS with "nearly identical" designs is false (P. 891). The two experiments included: (1) different factor levels (for three factors), (2) different unvalidated applicant names (to manipulate ethnicity), (3) different applicant photos, (4) different fixed factors (e.g., applicant stories about moving), and (5) different experimental settings (e.g., testing, instrumentation, and conditions of anonymity). In the current article, I discuss each of these major design differences and explain why it invalidates Forster and Neugebauer's (2024) comparison of their FE and FS findings. I conclude by emphasizing that social scientists are better served by asking why FE and FS findings sometimes differ than by assuming that any difference in findings across the experimental designs invalidates FS.

Keywords: factorial survey experiment; field experiment; hiring; survey methodology; race/ethnicity; bias

Citation: Pickett, T. Justin. 2025. "Invalidating Factorial Survey Experiments Using Invalid Comparisons Is Bad Practice: Learning from Forster and Neugebauer (2024)" Sociological Science 12: 97-105.

Received: September 27, 2024

Accepted: October 1, 2024

Published: January 27, 2025

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

DOI: 10.15195/v12.a5

AFTER obtaining different findings in a field experiment (FE) and a factorial survey (FS) experiment, Forster and Neugebauer (2024) dismissed the latter as invalid, ignoring the fact that nothing about their experiments was similar enough to allow a valid comparison of the estimates from them. Their study is but a recent example of this misguided practice. The most influential is Pager and Quillian's (2005) study of employers; those researchers found that an in-person FE and a telephone FS yielded different findings and concluded that survey-based studies of hiring discrimination are invalid, ignoring key design differences between their FE and FS.[1] For instance, telephone surveys are the worst mode for satisficing, sensitive questions (Holbrook, Green, and Krosnick 2003; Kreuter, Presser, and Tourangeau 2008), and vignettes—it is hard for respondents to remember a paragraph (or more) of detailed information after it is read aloud to them (Auspurg and Hinz 2015).

In Forster and Neugebauer (2024), the FS was administered online, yet the results still differed from an FE, leading them to conclude that "FSs are not suited for studying hiring behavior" (P. 901). This conclusion is erroneous because it ignores a fundamental fact about quantitative research: studies with dissimilar designs have different "empirical estimands," even if they have the same "theoretical estimand" (Lundberg, Johnson, and Stewart 2021, P. 541). Although Forster and Neugebauer (2024, P. 891) claimed that their FE and FS had "nearly identical" designs, the two experiments included: (1) different factor levels (for three factors), (2) different unvalidated applicant names (to manipulate ethnicity), (3) different

applicant photos, (4) different fixed factors (e.g., applicant stories about moving), and (5) different experimental settings (e.g., testing, instrumentation, and conditions of anonymity). Let us consider each in turn.

The FE and FS Included Different Factor Levels

The FS in Forster and Neugebauer (2024) included extra (more extreme) levels of three factors—job applicants' school grades (three levels), socioeconomic status (SES) (three levels), and education (four levels).[2] In contrast, the FE omitted extreme levels (see Tables 1 and S4). In the FE, the levels of two of the factors were fixed (i.e., reduced from three levels to one), such that all FE applicants had intermediate (i.e., satisfactory) grades and intermediate SES (i.e., were skilled workers). The lowest level of the third factor, education, was dropped, reducing it from four to three levels: "in the FE, there were no applicants with [an] intermediate high school degree" (P. 894). It turns out that the levels of school grades and education that were missing from the FE were also the strongest predictors of hiring choices in the FS (Table S7).

There are three reasons that the inclusion of different factor levels in Forster and Neugebauer's (2024) FE and FS invalidates any comparison of the findings across the experiments. First, adding factor levels in factorial experiments changes findings by affecting how respondents weight them in their decisions (De Wilde, Cooke, and Janiszewski 2008; Verlegh, Schifferstein, and Wittink 2002). As Auspurg and Hinz (2015, P. 20) emphasized, "dimensions [i.e., factors] defined with more levels attract more attention from respondents and, consequently, show a higher impact on judgments than dimensions with fewer levels." Therefore, it is unsurprising that the extra levels of school grades and education that Forster and Neugebauer (2024) added in their FS exerted a relatively high impact on hiring decisions in that experiment (Table S7).

Second, in addition to any "number-of-levels effects" (Auspurg and Hinz 2015, P. 20), the range covered by a factor's levels also matters (Mellers and Cooke 1994; Verlegh et al. 2002). Specifically, narrowing the range of levels for a factor, such as applicant education, affects its importance in respondents' decisions as well as the importance of other factors (e.g., applicant ethnicity), especially if the removed levels are the extreme ones (Verlegh et al. 2002). By including extreme factor levels in the FS (for school grades, SES, and education) but removing them from the FE, Forster and Neugebauer (2024) changed both the range and the number of levels for these three factors. These design changes alone may explain all the observed differences in their FS and FE findings.

Third, when analyzing data from factorial experiments (FE or FS), the main effect of any specific factor (e.g., applicant ethnicity) is conditional on the distribution of the other factors (e.g., applicant grades, SES, and education), because it averages over them (de la Cuesta, Egami, and Imai 2022). This issue goes beyond the problem of confounding (or aliasing) in fractional factorial designs (Auspurg and Hinz 2015). Even in full factorial designs, the empirical estimand for any factor's main effect is tied to the specific randomization distribution (i.e., uniform vs. nonuniform) used to assign the levels of all other factors (de la Cuesta et al. 2022). In Forster and

Neugebauer (2024), the FE was a fractional factorial design (omitting levels of three factors), whereas the FS was a full factorial design with a nonuniform randomization distribution (e.g., 25 percent of applicants had intermediate high school degrees and 50 percent were college dropouts). Put plainly, the FE estimand in Forster and Neugebauer (2024) for any specific factor (e.g., applicant ethnicity) is its effect when the distribution of grades, SES, and education is fixed so that all applicants have satisfactory grades, intermediate SES, and more than an intermediate high school degree. In contrast, the FS estimand is the effect when averaging over a different (full but nonuniform) distribution of these three factors. The two estimands are not the same and thus are not comparable.

Of course, Forster and Neugebauer (2024) would counter that they conducted a "robustness check" to deal with this issue, estimating the FS models only for vignettes describing applicants with satisfactory grades and intermediate SES (Table S14). However, this analysis is inadequate, because the FS respondents received (and simultaneously considered) a set of eight applicants who varied on these factors (from low to high grades and from low to high SES) as well as on education. Restricting the analysis to vignettes with specific levels of grades and SES does not change the fact that those vignettes were evaluated in the context of (and often immediately after) other vignettes with other levels of these two factors, nor does it change the fact that those vignettes included a level for education that was not used in the FE. Forster and Neugebauer (2024) might counter that they got the same findings when they only used the first vignette that respondents saw (Table S11). That, too, is an inadequate analysis, because among the vignettes that were evaluated first in the FS set ($n = 480$), only 38 matched the FE distribution of levels for the three factors—that is, had satisfactory grades, intermediate SES, and more than an intermediate high school degree (my analysis of the public data). In fact, among all the vignettes evaluated in the FS ($n = 3,840$), only 8 percent ($n = 310$) matched the FE distribution for applicant grades, SES, and education. Of these 310 "matching" vignettes, 272 (or 88 percent) were evaluated after other non-matching vignettes were considered first. Clearly, the two experiments differed widely in design. Their findings cannot be compared.

The FE and FS Included Different Applicant Names

Forster and Neugebauer (2024) manipulated applicant ethnicity (German vs. Turkish) using names—a perilous approach. As studies published in this journal (Crabtree et al., 2022; Johfre, 2020) and elsewhere have shown, names simultaneously convey information about many characteristics (e.g., age, ethnicity, gender, SES, and religion) (Crabtree et al. 2023; Darolia et al. 2016; Martiniello and Verhaeghe 2023). The problem is that names are essentially instrumental variables designed to influence respondents' perceptions of a specific applicant characteristic (e.g., ethnicity), and they are only valid if the exclusion restriction (or information equivalence assumption) is met—that is, if the name only affects the outcome via the intended perceptual mechanism and no other (Dafoe, Zhang, and Caughey 2018). Consequently, names used in a single experiment to signal different ethnicities (e.g., German vs. Turkish) may result in effects that are not actually due to perceived

ethnicity. As importantly, when different names are used to signal the same ethnicity (e.g., Turkish) in two experiments (e.g., FE vs. FS), the effects may differ across them even if applicant ethnicity is perceived correctly in both, because there may be perceptual differences in other factors signaled by the different names used in each.

There are two serious problems with how Forster and Neugebauer (2024) used the "names" method in their experiments. First, they used unvalidated names in both the FS and FE, choosing them simply by selecting "from a list of the most common German and Turkish names" (P. 893). No evidence is provided that these names are information equivalent on factors other than ethnicity. Notably, other experimentalists are increasingly turning to lists of validated names to deal with this issue (Crabtree et al. 2023). The second problem is that Forster and Neugebauer (2024) used different names in the FS and FE. In fact, they only used a single name in the FE for each ethnicity/gender group (i.e., Leon/Julia Fischer for the German man/woman; Mehmet/Zeynep Yilmaz for the Turkish man/woman). In the FS, they used a larger set of names, including different first names and different last names. However, there is zero evidence that the different names used in the FE and FS are information equivalent either within or between the experiments, so Forster and Neugebauer's (2024) entire approach in manipulating ethnicity is questionable, and the comparison of "ethnicity" effects across the FS and FE is invalid.

The FE and FS Included Different Applicant Photos

Although Forster and Neugebauer (2024) manipulated applicant ethnicity using names, they also included photos of the applicants that were not intended to affect perceived ethnicity. Specifically, in the FE, the same photos were used for applicants of different ethnicities—that is, the same photographed applicant was given both a German and a Turkish name. This also happened in the FS. Importantly, however, different photos were used in the FE and FS. Across both experiments, "a total of 10 portrait photos" were used (P. S6). Of these 10 photos, two "were reserved for the field experiment" (P. S7) and the other eight were used in the FS.

Forster and Neugebauer (2024) never explained how they chose the two photos for the FE nor did they ever provide convincing evidence that the utilized photos were information equivalent either within or between the FE and FS. They note that the 10 utilized photos were pretested with online panelists "for attractiveness, age and other characteristics" and were judged to be the "most similar" of the photos evaluated (P. S6). However, it is not clear what specific characteristics were evaluated in the pretest nor is it clear how similar the evaluations of the 10 photos were. What we do know is that photos "suffer from a lower standardization" than words and often cannot be compared validly in experiments (Auspurg and Hinz 2015, P. 74). Specifically, photos of people of the same ethnicity and gender are often judged to be very different on a whole set of factors, such as dominance, trustworthiness, dangerousness, warmth, happiness, temper, disgust, and sadness (Ma, Correll, and Wittenbrink 2015). Unless Forster and Neugebauer (2024) tested a full set of characteristics to ensure that all 10 photos were information equivalent on all relevant factors and were also perceived as equally German and Turkish (because they were used with names signaling both ethnicities), the findings both

within and between their experiments are questionable (Auspurg and Hinz 2015; Mutz 2011). This is especially true given that only two photos were used in the FE, one for each gender. Consequently, if either of those FE photos lacked information equivalence with the FS photos—on its own or when paired with a specific ethnic name—it would invalidate any comparison of the FE and FS findings.

The FE and FS Included Different Fixed Factors

Forster and Neugebauer (2024) included many additional fixed factors in the FE that were not included in the FS. For example, the FE included a personal narrative about the applicant's residential location and reason for moving; "In the FE, we included a short note on place of living in the cover letter . . . we constructed a short story that explained the reasons for relocating" (P. S4). No such narrative (or story) was included in the FS. The FE included real school names (e.g. "University of Siegen"), but the FS did not. The FE included "a full school leaving certificate" (P. S6), but the FS did not. The FE included names of internship companies (e.g., "BGH Edelstahl Siegen GmbH in Siegen"), but the FS did not.

Part of the reason that more information was included in the long, detailed FE application materials, was that Forster and Neugebauer (2024) wanted all the information in the FS "to fit on one computer screen page" (P. 892). So, they used cover-letter "excerpts" in the FS that were "considerably shorter" (pp. S5–S6). The FS cover-letter excerpts were also written differently (at a lower writing level) than the longer FE cover letters (see pp. S4–S5). The problem is that all the factors that were present in the FE but not in the FS may condition the effect of other applicant characteristics (e.g., education), undermining the comparability of estimates across the two experiments. In addition, putting all the applicant information on one computer screen in the FS was a bad idea. As Mutz (2011, P. 87) explained, "expecting subjects to scroll down in order to see additional material [is] not a safe bet." Instead, she recommends placing the information on multiple pages and interspersing it with questions.

The FE and FS Used Different Experimental Settings

The FS was administered eight weeks after the FE "to the same employers . . . using the same email address" (P. 892). There are several serious problems in this approach. First, the FS was not anonymous. Forster and Neugebauer (2024, P. S23) had the identifying information necessary to "look at exactly the same respondents in both experiments." Obviously, the FE effects observed under the assumed anonymity may differ from those observed in an FS where respondents are aware that their identities are known (Tourangeau and Yan 2007). Technically, the estimand in such an FS would be the causal effect of a given factor (averaging over the distribution of other factors) when you know a researcher is watching you.

The second problem is testing or more specifically the interaction of pretesting and treatment. Forster and Neugebauer (2024) used fake applicants in the FE, then declined (or ignored) the positive responses (e.g., interview invitations) that 54

percent of those fake applicants received from employers (Table 2), and then eight weeks later sent the same employers eight pictures of applicants who were dressed exactly like those earlier fake applicants: "All applicants were photoshopped to wear the same plain black button-down shirt" (pp. S6–S7). Methodologists have long warned that such pretesting can change treatment effects because of reactivity (Shadish, Cook, and Campbell 2002). As Campbell and Stanley (1963, pp. 5–6) emphasized, "a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of . . . [those] for the unpretested universe from which the experimental respondents were selected." Put plainly, Forster and Neugebauer's (2024) FE results are for unpretested employers, whereas the FS results are for pretested ones; the two are not comparable.

The third problem is instrumentation or more specifically the interaction of the specific instruments used and treatment (Campbell and Stanley 1963; Shadish et al. 2002). Unlike in the FE, the FS involved respondents first rating each of the eight applicants "on an 11-point scale from 0 percent to 100 percent" (P. 892). Unlike in the FE, the FS involved respondents choosing which applicants to interview "with the entire pool of applicants on one page, along with key information and their own prior ratings" (P. 893). Unlike in the FE, the FS involved respondents only considering a set of eight applicants all pictured wearing the same clothes. Unlike in the FE, the FS involved respondents having to scroll down to see all applicant information. Fortunately, this is one place where Forster and Neugebauer's (2024) supplementary analyses provide some comfort. Specifically, their finding that the results were similar when using the 0–100 scale and when analyzing only the first vignette that respondents evaluated ($N = 480$) are somewhat reassuring (Tables S11 and S16). However, these supplementary analyses only get at some of the instrumentation differences that may moderate treatment effects in the FE and FS, and they fail to deal with any of the other issues outlined above (e.g., the use of different factor levels, names, and photos in the FE and FS).

Conclusion

Forster and Neugebauer's (2024) invalidation study is invalid. Their conclusion that "FSs are not suited for studying hiring behavior" (P. 901) is unjustified, because their claim that the FE and FS had "maximally similar" designs is false (P. 887). The only thing their study tells us is that if you conduct a FE and FS with the same employers a few weeks apart, and those experiments have very different designs, the results may differ. Social scientists would be much better served by asking why FE and FS findings sometimes differ than by assuming that any difference in findings across the designs invalidates FS.

We know that well-conducted FS often yield causal effects that are reasonable and that mirror real-world behavior (Auerbach and Thachil 2018; Hainmueller, Hangartner, and Yamamoto 2015). For example, employers in FS prefer job applicants who have more education, more work experience, better references, and clean records (i.e., no criminal convictions) (Bushway and Pickett 2024). We also know that FE is plagued by design issues (e.g., the use of invalid names) that social

scientists too frequently overlook (Crabtree et al. 2022; Darolia et al. 2016; Heckman 1998). According to Neumark (2012, P. 1129), some of the most damaging criticisms of FE have "been ignored in the literature" casting "serious doubt on the validity of the evidence from these studies." Consequently, there are two important takeaways about experimental validation studies. The first is that FE and FS can both have flaws, so it is a mistake to assume that the FE in an FE–FS comparison is always right. The second is that even well-conducted (and valid) FE and FS may have different findings if they have different empirical estimands.

Notes

1 Pager and Quillian (2005, P. 374) went so far as to argue that the difference in their FE and FS findings suggested a need to question whether the broader survey evidence of "a liberalizing of racial attitudes among white Americans . . . [has] any necessary correspondence to the incidence of discrimination."

2 Although Table 1 provides three education levels in the FS and two in the FE, there were actually four and three levels, respectively, because the field of study (language or mathematics) was appended to the highest education category (i.e., some college but dropped out).

References

Auerbach, Adam M. and Tariq Thachil. 2018. "How Clients Select Brokers: Competition and Choice in India's Slums." *American Political Science Review* 112(4): 775–91. https://doi.org/10.1017/S000305541800028X

Auspurg, Katrin and Thomas Hinz. 2015. *Factorial Survey Experiments*. Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781483398075

Bushway, Shawn D. and Justin T. Pickett. 2024. "Direct Incentives May Increase Employment of People with Criminal Records." *Criminology and Public Policy*. https://doi.org/10.1111/1745-9133.12681

Campbell, Donald T. and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Boston, MA: Houghton Mifflin Company

Crabtree, Charles S., S. Michael Gaddis, John B. Holbein, and Edvard N. Larsen. 2022. "Racially Distinctive Names Signal Both Race/Ethnicity and Social Class." *Sociological Science* 9:454–72. https://doi.org/10.15195/v9.a18

Crabtree, Charles, Jae Y. Kim, S. Michael Gaddis, John B. Holbein, Cameron Gauge, and William W. Marx. 2023. "Validated Names for Experimental Studies on Race and Ethnicity." *Scientific Data* 10:1–10. https://doi.org/10.1038/s41597-023-01947-0

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26(4):399–416. https://doi.org/10.1017/pan.2018.9

Darolia, Rajeev, Cory Koedel, Paco Martorell, Katie Wilson, and Francisco Perez-Arce. 2016. "Race and Gender Effects on Employer Interest in Job Applicants: New Evidence from a Resume Field Experiment." *Applied Economic Letters* 23(12):853–56. https://doi.org/10.1080/13504851.2015.1114571

de la Cuesta, Brandon, Naoki Egami, and Kosuke Imai. 2022. "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution." *Political Analysis* 30(1):19–45. https://doi.org/10.1017/pan.2020.40

De Wilde, Els, Alan D. J. Cooke, and Chris Janiszewski. 2008. "Attentional Contrast During Sequential Judgments: A Source of the Number-of-Levels Effect." *Journal of Marketing Research* 45(4):437–49. https://doi.org/10.1509/jmkr.45.4.437

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. https://sociologicalscience.com/articles-v11-32-886/

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments Against Real-World Behavior." *PNAS* 112(8):2395–400. https://doi.org/10.1073/pnas.1416587112

Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12(2):101–16. https://doi.org/10.1257/jep.12.2.101

Holbrook, Allyson L., Melanie C. Green, and Jon A. Krosnick. 2003. "Telephone versus Face to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias." *Public Opinion Quarterly* 67(1):79–125. https://doi.org/10.1086/346010

Johfre, Sasha S. 2020. "What Age is a Name?" *Sociological Science* 7:367–90. https://doi.org/10.15195/v7.a15

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys." *Public Opinion Quarterly* 72(5):847–65. https://doi.org/10.1093/poq/nfn063

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. "What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86(3):532–65. https://doi.org/10.1177/00031224211004187

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

Martiniello, Billie and Pieter-Paul Verhaeghe. 2023. "Different Names, Different Discrimination? How Perceptions of Names Can Explain Rental Discrimination." *Frontiers in Sociology* 8:1125384. https://doi.org/10.3389/fsoc.2023.1125384

Mellers, Barbara A. and Alan D. J. Cooke. 1994. "Trade-Offs Depend on Attribute Range." *Journal of Experimental Psychology: Human Perception and Performance* 20(4):1055–67. https://doi.org/10.1037/0096-1523.20.5.1055

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press. https://doi.org/10.23943/princeton/9780691144511.001.0001

Neumark, David. 2012. "Detecting Discrimination in Audit and Correspondence Studies." *Journal of Human Resources* 47(4):1128–57. https://doi.org/10.1353/jhr.2012.0032

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

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

Tourangeau, Roger and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133(5):859–83. https://doi.org/10.1037/0033-2909.133.5.859

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 Letter* 13(1):41–52. https://doi.org/10.1023/A:1015063125062

Justin T. Pickett: School of Criminal Justice, University at Albany, SUNY. E-mail: jpickett@albany.edu.