This is the accepted manuscript (post-print version) of the article. The content in the accepted manuscript version is identical to the final published version, although typography and layout may differ.

How to cite this publication
Please cite the final published version:


Publication metadata


Author(s): Wulff, Jesper; Villadsen, Anders Ryom.

Journal: European Management Review.

DOI/Link: 10.1111/emre.12342

Document version: Accepted manuscript (post-print)

Document license:

General Rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

If the document is published under a Creative Commons license, this applies instead of the general rights.
Are Survey Experiments as Valid as Field Experiments in Management Research? An Empirical Comparison using the Case of Ethnic Employment Discrimination

Abstract

Field experiments have long been the gold standard in studies of organizational topics such as ethnic discrimination in recruitment. The recent use of survey experiments, also known as experimental vignettes, suggests that some researchers believe that survey experiments could be used as an alternative to field experiments. In this study we put this notion to the test. We perform a field experiment followed by two survey experiments on ethnic discrimination in recruitment. While the results of our field experiment are consistent with previous evidence on discrimination, one survey experiment concludes no difference between native and immigrant employees while another concludes positive discrimination. These results should invoke caution in researchers wanting to investigate organizational topics using survey experiments.

Keywords: Research methods, field experiment, survey experiment, discrimination
1. Introduction

Experiments are the gold standard for investigating causal relationships. In organizational research, they remain the most powerful technique we have for determining the existence, direction and magnitude of a causal relation between two variables. The random assignment of research subjects to either a treatment or control group eliminates possible alternative effects of unobserved factors, whose existence is almost always a concern when dealing with observational data (Hamilton and Nickerson, 2003).

Experiments come in many shapes and sizes. Most experiments in management have been done in the laboratory (Scandura and Williams, 2000). While lab experiments provide the desirable control over inessential influences, they may be criticised for being too artificial to actually generalise to conditions in organizations outside the laboratory (Fromkin and Streufert, 1976). Field experiments offer a potential solution to the problem of artificiality in lab experiments by trading control for naturalism (Whitley, 1996). The organizational researcher manipulates the variable of interest and observes its potential effect in natural settings. This increases the realism of the results. Field experiments have been widely used in studies of employment discrimination where researchers have applied for real job openings with artificial applications that vary on a dimension of interest such as name, gender, or picture (for one of the first applications see Bertrand, M., & Mullainathan, 2004).

One particular boundary might prevent a wider adoption of the field experimental approach: the question of feasibility. Some research questions simply require impractical field experimental research designs. For instance, in the study of discrimination in the hiring process, the researcher may want to experimentally manipulate not only the characteristics of the applicant, but also the characteristics of the job (Greenberg and Tomlinson, 2004). Other field experiments may be
excessively expensive or unethical to employ. It would, for instance, be difficult to imagine researchers randomly assigning some firms to unethical behaviour (say use of child labour) to observe the consequences for recruitment when compared to a control group.

Researchers across the social sciences believe that survey experiments, also known as experimental vignettes, may pose an attractive alternative to field experiments (Aguinis and Bradley, 2014; Barabas and Jerit, 2010; Mullinix, Leeper, Druckman, and Freese, 2015; Shah, Bartlett, Carpenter, Nicholas, and Hemingway, 2014). Survey experiments combine the control of the experiment with the generalisability of the survey when used on representative samples. For instance, in the case of discrimination in recruitment researchers may manipulate the characteristics not only of the applicant but also of the job applied for in a survey vignette as suggested relevant in previous research (Cortina, 2008). Respondents, randomly assigned to different versions of the vignette, then rate applicants on different dimensions.

Recent empirical studies suggest that researchers believe that survey experiments constitute a capable design for studying organizational topics as discrimination and gender differences in management (Baekgaard and George, 2018; Baert and De Pauw, 2014; Brescoll, 2011). However, using survey experiment designs for topics such as discrimination implies that respondents’ behavioural intentions are the same as their real behaviour. That is, a respondent to a survey vignette should rate an applicant with an immigration background low if the participant discriminates against immigrants in his/her actual work-life. While some researchers argue that self-report measures on perceptions correlate highly with actual behaviour (De Dreu, Evers, Beersma, Kluwer, and Nauta, 2001), the theory of planned behaviour (Ajzen, 1991) suggests that this may be a substantial concern, especially in studies of sensitive or tabooed topics such as discrimination. It is argued that the artificial and simplified surroundings of a survey simply
cannot create the relevant real-world pressures individuals face (King, Hebl, Botsford Morgan, and Ahmad, 2013). Further, subjects may be less likely to go against established societal norms and values when filling in a survey where no future consequences are faced (King and Bruner, 2000). In line with these criticisms, Aguinis and Bradley (2014) highlight that establishing an adequate level of ‘immersion’ in the survey experiment is paramount. Yet, this may be next to impossible when exploring processes involving implicit biases that are less likely to be triggered when the consequences of a decision are less tangible than a future work relation (Cortina, 2008).

While research has warned that survey experiments may in some cases be invalid, little research has empirically explored this proposition. Therefore, we approach the question from an empirical angle. We perform a field experiment followed by two different survey experiments in the context of ethnic discrimination in recruitment. This is a sensitive topic where research subjects’ intended and real behaviour is likely to diverge and thus a tough test of survey experiments. Our goal was not to conduct ‘perfect’ experiments (if such a notion is in any way meaningful) but to study ethnic employment discrimination using the techniques that are currently prevalent and compare the results. Our findings suggest a substantial difference in conclusions when comparing our three studies. While the field experiment suggests a substantial negative discrimination of Middle-Eastern name applicants, the survey experiments suggest no and even positive discrimination, respectively. Together, our results should invoke caution for organizational researchers in the use of survey experiments when studying topics like discrimination in recruitment. Survey experiments can provide important insights for management researchers but are no panacea. Best practices must be followed notably including a realistic level of immersion. When this is not accomplished, as in our studies here, we fear that
survey experiments may lead to erroneous conclusions underestimating ethnic employment discrimination.

In the following, we briefly describe our empirical context of job discrimination and the design of our experiments. Next, we present our results before ending the paper with a brief discussion.

2. The Empirical Context: Discrimination in Recruitment

To compare field and survey experiments we turn to the case of ethnic employment discrimination in recruitment. This is an example of a socially sensitive research area where “there are potential consequences or implications, either directly for the participants in the research or for the class of individuals represented by the research” (Sieber and Stanley, 1988: 49).

We do not wish to imply that either field experiments or survey experiments are exclusively relevant for a sensitive topic such as discrimination. Indeed, we see both methods having wide usability in organization and management studies. However, we find this context relevant for different reasons. First, it is a setting where both field experiments (Bertrand and Mullainathan, 2004; Pager, Western, and Bonikowski, 2009) and survey experiments (Baert and De Pauw, 2014) have been used. Second, while field experiments arguably are the gold standard, they may be difficult to invoke for topics that are hard to observe and manipulate. In business and management research there are many such topics at both micro and macro levels. At the individual level, sensitive topics include workplace discrimination, bullying and harassment. It is difficult to get reliable and valid information about such practices from both the victim and, especially, the perpetrator when asking directly. Often such practices also go unreported and do not figure in organizational records. At the organizational level, corporate wrong-doing such as
fraudulent accounting, unpermitted pollution, or a dangerous work environment are sensitive topics that may be difficult to study with conventional methods such as questionnaires or archival searches and all but impossible to study using field experimental methods.

Discrimination in recruitment has been intensively studied (King et al., 2013). For close to 50 years, field experiments in the form of correspondence tests have been used frequently to investigate discrimination in recruitment (Zschirnt and Ruedin, 2016). Researchers send two or more applications based on fictitious candidates with substantially identical qualifications to a real employer. In the study of discrimination against applicants with an immigration background, the applications only differ with respect to the name of the applicant signaling ethnicity or race. A difference in callback rates of the two equally qualified groups may arguably be attributed to discrimination (Jackson and Cox, 2013). Using this approach, numerous studies have documented strong evidence of discriminatory behaviour against applicants with an immigration background across a variety of different countries (e.g. Bertrand and Mullainathan, 2004; Carlsson and Rooth, 2007; Hoque and Noon, 1999; Midtbøen, 2015).

The field experimental design is highly relevant for studying discrimination and it provides researchers with a strong basis for defending their conclusions (King et al., 2013). Importantly, researchers can test the discrimination level directly while receiving results that are easy to interpret (Jackson and Cox, 2013). Because the applications are identical, except for the name of the applicant, differences in callback rates can only be due to the name. In addition, this type of experiment allows a higher degree of construct validity as it generates an outcome in real world situations (Shadish, Cook, and Campbell, 2002), and participants are not artificially affected by social desirability pressures in addition to what they normally face in their work. Thus, the participants’ responses to the résumés are more representative of naturally occurring events
(King et al., 2013). Finally, when field results from field experiments are repeated across different contexts, claims about external validity are possible (Chatterji, Findley, Jensen, Meier, and Nielson, 2016).

Still, field experiments studying discrimination are not without disadvantages. Realism comes at the expense of loss of control. Limitations arise when manipulating possibly important aspects like job description, number of applications received and formalization of recruitment procedures (Midtbøen, 2015). Further, one may question the morality of deceiving employers and interfering in a real recruitment process. While steps can be taken to mitigate the ethical costs (e.g. quickly responding to callbacks), the implicit deception of the employer is unavoidable. These challenges are not exclusive for studies of discrimination and similar considerations must be made for field experiments in other settings. The researcher must decide whether the benefits of performing the experiment outweigh the limitations and costs.

Survey experiments combine survey sampling methods with experimental manipulation (Mutz, 2011). Continuing the example of recruitment discrimination, researchers have relied on self-reported data for many years (Deitch et al., 2003) yet with the lack of causal interpretation made possible with experiments. To mitigate this shortcoming researchers can design a factorial survey where short, hypothetical descriptions (vignettes) of candidates and/or jobs are used to affect the respondents’ attitudes towards a job candidate. This opens for manipulating not only the ethnicity of a job candidate, but also the characteristics of the job or organization albeit they are now fictitious.

Regarding discrimination, as suggested by Cortina (2008) one could suspect that different organizational characteristics affect the likelihood of discrimination. Researchers cannot normally manipulate organizational characteristics as it is possible in a survey experiment. In the
two survey experiments presented below, we illustrate this by exploring the importance of job complexity and the degree of performance monitoring. We expected that discrimination may be less pronounced for jobs that represent less uncertainty for the recruiter, i.e. simple jobs that are easy to monitor.¹ Thus, survey experiments investigating discrimination offer the researcher a stricter control of which factors to manipulate and how.

A further benefit of the survey experiment is that it offers an appealing level of external validity when performed on a representative sample. In principle, results should therefore both be generalizable to the population and provide the possibility of testing causal hypotheses (Jackson and Cox, 2013). Finally, previous research suggests that self-administered questionnaires produce higher levels of reporting of sensitive behaviors (abortion, sexual behavior, illicit drug use, etc.) than face interviews (Tourangeau, Rasinski, Jobe, Smith, and Pratt, 1997).

Survey experiments are, however, not necessarily the free lunch they sound to be. Importantly, the survey experiment is conducted in an artificial setting using hypothetical scenarios. This invokes lower construct validity best understood through the problem of intent versus realism²: Asking recruiters what they would have done in a hypothetical situation does not have to correspond with that they would actually do in practice (Midtbøen, 2016). The concept of self-deception is especially relevant when respondents unintentionally present a favorable, yet inaccurate version of themselves (Dalal and Hakel, 2016). Research suggests that

---

¹ In Study 2 we also manipulated that some candidates represent a better fit with organizational culture. This was done to imitate some of the information that a recruiter would normally have but we could not provide as we only showed a very simplified version of a resume.

² We are indebted to an anonymous reviewer for pointing out the connection between the concept of construct validity and the intent versus realism problem.
this distortion in a socially acceptable direction takes place even if respondents fill out the survey anonymously (Zerbe and Paulhus, 1987).

In addition to social desirability effects, participant behavior may be affected by the artificiality of the questions, how the survey is administered and the lack of attention of the respondent to the questions asked (Eifler, 2010; Jackson and Cox, 2013). In fact, research suggests that the artificial environment of the survey experiment could be even more of a problem than for laboratory experiments where subjects respond with stronger responses in the theoretically predicted direction (Collett and Childs, 2011). Together, this suggests that survey experiments of a topic like discrimination may give a biased picture of the actual discrimination that takes place in the real world. This is due to factors such as unintended self-deception that makes subjects conform to established norms and values.

Survey experiments have several desirable features and are a relatively cost-efficient way of collecting data. They are also widely used in other social science fields such as political science and public administration. While still used to a very limited extent, survey experiments hold promise also for management and business research with possible applications to a range of topics including discrimination. Hence, it is important to know of the strengths and weaknesses of this method. In the following, we compare a field experiment and two survey experiments in the study of ethnic employment discrimination.

3. Method and Experiment design

In the following we describe the design of our experiments. Sample descriptions from the experiments can be found in the appendix. The design of our field experiment is made to closely match recent high-quality field experiments of discrimination (Bertrand and Mullainathan, 2004; Carlsson and Rooth, 2007). The design and descriptions used for our survey experiments are
inspired by examples from the survey experiment literature (Collett and Childs, 2011; Jasso, 2006) and actual survey experiments made on sensitive organizational topics such as ethnic discrimination and attitudes towards the behaviour of female managers (Baert and De Pauw, 2014; Brescoll, 2011).

We collect our field- and survey experimental data from different sources. We do this to replicate as closely as possible the data collection methodology currently used in each branch of studies. While drawing our samples from the same source would have increased the internal validity of our own study, it would have deviated substantially from the current praxis. As we aim to have our studies replicate how organizational researchers within each methodological branch would have investigated recruitment discrimination, we opted for collecting our data using the current standard. For the field experiment, this meant sending out real applications for real job openings. For the survey experiments, we drew our samples from the YouGov panel as this panel is used extensively in survey experimental research (e.g., Ahlquist, Mayer, and Jackman, 2014; Nyhan and Reifler, 2015; Olsen, 2017).

The research design includes three studies. The first is our baseline study where we conduct a field experiment. The following two studies are different versions of a survey experiment designed to reveal discriminatory behaviour among managers and employees. All studies are set in Denmark and conducted within a narrow timeframe not to allow external events to affect the attitudes towards immigrants in different ways in the three studies.

3.1 Study 1: Field Experiment

In Study 1 we set up a field experimental research design to examine ethnic discrimination in Denmark. We sent out 888 artificial applications to 222 real job openings from four occupational
categories between February and July 2015. We applied for openings within the following job categories: School teacher, physiotherapist, office assistants and financial controller. We created a bank of applications with six applications for each job category. Each application consisted of a general biography and a detailed résumé.

Using data from Statistics Denmark, we identified the six most often used names for citizens with Danish and a non-western origin, respectively. For each job advertisement, we drew four applications at random from the relevant résumé bank. Next, two Danish and two Middle-Eastern names (one female and one male each) were drawn at random from each name list and attached to the applications.

Our outcome variable of interest *Callback* is assigned the value 1 if an applicant is called for an interview and 0 otherwise. Our other variable *Middle-Eastern name* is assigned the value 1 if the application came attached with a Middle-Eastern sounding name and 0 if it came with a Danish name. Discrimination is then measured by cross-tabulating the two variables and observing to which extent the frequency of callbacks changes depending on the assigned name (Carlsson and Rooth, 2007; Pager, 2007).

---

3 We acknowledge the ethical challenges surrounding this research methodology. While no IRB approval was required for the study, we conducted it based on a call for such research from the government-financed Danish National Centre for Social Research (SFI) (Jensen, Weibel, Tørslev, Knudsen, and Jacobsen, 2012). We promptly declined any interview invitation we received and were conscious not to oversample the population applying for considerably fewer jobs than most existing studies.

4 The Danish names were Peter Jensen, Jens Nielsen, Lars Hansen, Anne Pedersen, Kirsten Andersen, Hanne Christensen while the Middle-Eastern names were Ali Aslam, Mohammad Ashraf, Ahmad Afzal, Mariam Yildrim, Fatima Akhtar, Yasmin Yilmaz.
3.2 Study 2: Survey Experiment with Forced Choice

In Study 2 we changed the research design to a survey experiment. We recruited a representative sample of 2491 Danish managers from the YouGov panel. In June 2016, 706 managers completed the survey resulting in a response rate of 28.3%.

The experiment was divided into two parts. In the first part, the respondent received at random a description of either a complex or simple job. The respondent was asked to imagine that his/her firm was hiring a candidate for the job in the description. Next, the respondent was shown two candidate descriptions. One candidate was described as being very qualified for the position, but not as good a fit with the company culture. The other candidate was characterized as less qualified than candidate one, but a better fit with the company culture. The two descriptions were attached with a Danish or Middle-Eastern name at random. The respondent was asked to select the preferred candidate for an interview. In the second part, the experiment was repeated with job descriptions presenting a job where performance was difficult to monitor and one where it was easy. Following the same procedure, the respondent was presented with two job descriptions similar to above with a Danish or Middle-Eastern name attached at random.

Our outcome variable of interest Callback is assigned the value 1 if a candidate description is chosen by the respondent to be called for an interview and 0 otherwise. Our other variable Middle-Eastern name is assigned the value 1 if the application came attached with a Middle-Eastern sounding name and 0 if it came with a Danish name. As in Study 1, discrimination is then measured by cross-tabulating the two variables and observing to which extent the frequency of callbacks changes depending on the assigned name.
3.3 Study 3: Survey Experiment with Continuous Outcome

One may argue that the set-up in Study 2 makes the experimental manipulation too obvious to the respondent. In Study 3, the manipulation is hidden from the participant who is only presented with one of two possible candidate descriptions attached with one name. This should make it less likely that the respondent becomes too aware of the manipulation. In addition, we employ a continuous measure of discrimination using a series of Likert-scale questions as this is a popular way of measuring opinions about sensitive organizational topics.

We recruited 276 public employees from the YouGov Panel in February 2017 with a response rate of 27%. The experimental set-up was very similar to that of the second part of Study 2. The respondent was presented with one randomly selected job description describing a job where performance was either difficult or easy to monitor. Either a Danish or Middle-Eastern name was selected at random and attached to the description presented to the respondent. Finally, the respondent was asked to rate the candidate on a series of six questions described in detail below.

The main variable of interest in this study is Candidate preference. Inspired by Beart and de Pauw (2014) and Brescoll (2011), we measured this variable as a composite index of six Lickert-scale items assessed by the respondent on a scale from one (completely disagree) to seven (completely agree). The items included “The candidate would fit well in our workplace”, “I would be happy to work together with the candidate”, “My colleagues would be happy to work with the candidate”, “The candidate would perform satisfactory in this position”, “If it were up to me, it would be highly likely that the candidate would be offered a job in my organisation” and “I would invite the candidate to a job interview”. Together, the items formed a reliable scale5 (α

5 All data analysis was performed using the R statistical software environment (R Core Team, 2016). Reliability measures we computed using the psych package (Revelle, 2017) and Bayes factors were computed using the BayesFactor package (Morey and Rouder, 2015). Bayarri and Mayoral’s (2002) equality-of-effect-size Bayes factor test was computed using the R code in Verhagen and Wagenmakers (2014).
We assessed name-based discrimination by comparing the group means of the composite candidate preference measure.

4. Results

In this section, we present the results from our experiments. We only present results relevant for comparing discrimination between the experiments.

4.1 Study 1: Field Experiment

124 out of 444 Danish-named applications received a callback for an interview, while only 81 Middle-Eastern-named candidates were invited. This results in a difference of about 10%-points in callback probability favouring the Danish-named applicants. A $\chi^2$-test suggested a rejection of the null hypothesis of independence suggesting named-based discrimination, $\chi^2(1) = 11.19, p < .001$. A Bayes factor (BF) of 25.15 suggested that the alternative that callback and Middle-Eastern name are dependent is 25 times more likely than the null hypothesis. This constitutes strong evidence in favour of name-based discrimination (Jeffreys, 1961). 30,000 draws from the posterior distribution reveal that there is 95% probability that the name-caused callback difference is between 4 and 15%-points.

4.2 Studies 2 and 3: Survey Experiments

In Study 2, the results from the two experiments were very similar and showed virtually no difference in preference for a candidate based on their name. The Danish candidate was chosen 50% of the time (second round: 52%) which was not enough to reject the null hypothesis of independence at conventional alpha levels, $\chi^2(1) = .003, p = .96$ (second round: $\chi^2(1) = 1.50, p =$ ...
The corresponding BF of .1 favoured the null hypothesis, which is equivalent to the null hypothesis of independence being 10 times more likely than the alternative (second study: BF = .22). This is strong (second round: substantial) evidence that candidate preference does not depend on name ethnicity. Pooling the samples from the two experiments in Study 2 did not change these conclusions, $\chi^2(1) = .89, p = .35; \text{BF} = .11$.

In Study 3, the candidates with a Middle-Eastern name were rated with a mean score of 5.04 while the mean score was 4.51 for Danish-named candidates. Thus, on average, Middle-Eastern-named candidates were preferred over Danish-named candidates, mean difference = .53, $t = 3.67$, $p < .001$. It is 78 times more likely that the means are different than the alternative (BF = 78.36), which is strong evidence in favour of respondents having a higher preference for Middle-Eastern-named candidates. 30,000 draws from the posterior distribution revealed that there is 95% probability that the higher preference for candidates with a Middle-Eastern name is between .23 and .78.

5. Discussion

To our knowledge, our study is the first to empirically assess whether field and survey experiments yield comparable results in studies of workplace discrimination or other topics in organizational research. Our results suggest that in the study of recruitment discrimination researchers may arrive at different conclusions when settling for survey experiments instead of field experiments. On this basis, we argue that field experiments should still be the preferred design when studying a sensitive organizational topic like discrimination.

Our studies are representative with respect to both design and results when compared to studies that are conducted to investigate hiring discrimination. First, using a field experimental design, we were able to closely replicate the results of past experiments on ethnic discrimination.
In a recent meta-analysis across 36 studies, Zschirnt and Ruedin (2016) estimated a median relative callback rate of 1.44, which is very close to our estimated relative callback rate of 1.41. The estimated callback rate in Study 1 also closely resembles those estimated in relatively recent field experiments performed in Norway (Midtbøen, 2016) and Sweden (Carlsson and Rooth, 2007), which are 1.34 and 1.50, respectively. Second, using two different survey experimental designs in Studies 2 and 3, we found no and positive discrimination, respectively. While these conclusions are both theoretically and empirically improbable, they resemble those found in a recent survey experiment on hiring discrimination by Baekgaard and George (2018). Baekgaard and George (2018) use a design that is in many ways similar to our Study 3: Both experiments were performed in February 2017 using an online survey sent to public sector staff, consider the hiring for administrative public sector positions and use similar names to represent ethnic majorities (e.g. Sofie) and minorities (e.g. Mohammed). Baekgaard and George (2018) ask Flemish local politicians how competent the applicant is on a scale from 0 to 10, and contrary to their expectations, they find that local politicians, on average, rate applicants with Arab sounding names 0.268 ($p = 0.005$) points higher. By computing Bayarri and Mayoral’s (2002) equality-of-effect-size Bayes factor we can compare the effect found in Baekgaard and George (2018) directly to the one found in Study 3. This procedure estimates that it is 10 times more likely than not that the effects from the two studies are the same (BF = 10.13). While Study 3 is by no means an exact replication of Baekgaard and George (2018), this comparison certainly suggests that our survey experiments are representative of studies being conducted in this field of research – not only in terms of design, but also in terms of the obtained results.

Field experiments have been widely used in the study of discrimination (Midtbøen and Rogstad, 2012; Zschirnt and Ruedin, 2016). While this methodology has some distinct
advantages, it also has limitations that must encourage researchers to look for alternative methods for increasing our understanding of this important topic. Survey experiments may constitute such an alternative due to very desirable features including generalizability and wide researcher discretion. Especially the latter is important as it enables rigorous studies that take the context much more into account (Cortina, 2008). Despite the apparent advantages, our paper warns researchers that survey experiments come with their own challenges some of them possibly hard to mitigate (Aguinis and Bradley, 2014).

Our results suggest that the survey experiment may create conditions that are not transferable to sensitive real-life situations when adequate immersion cannot be achieved. There are several reasons why researchers might observe differences between field and survey experiments. One explanation is that there might be a significant difference between behavioural intentions and real behaviour when studying discrimination. Respondents to the survey may be deceiving themselves about how they would actually react in an actual recruitment situation thus distorting the picture towards a more socially desirable outcome (Dalal and Hakel, 2016). This may also be true for research areas that in practice involve personal sacrifice or expenses that are not felt in survey experiments, for instance charity or support for infrastructure projects close to participants’ homes. Another explanation might be that the respondents find it difficult to generate real emotions through vignette studies (Parkinson and Manstead, 1993). This could be because respondents simply do not react as strongly to a vignette description compared to a tangible experience or because they simply do not make the sufficient effort to place themselves in the recruiting situation (Collett and Childs, 2011; Kiyonari, Tanida, and Yamagishi, 2000). This is a common threat to survey experiments and other vignette based studies that will cause
differences when compared to field experiments where the real life settings ensures (at least near) full levels of immersion.

Creating an adequate level of immersion in vignette studies is crucial (Aguinis and Bradley, 2014). This is extremely difficult for situations such as recruitment and thus ‘aversive racism’ (Cortina, 2008; Dovidio and Gaertner, 2004) may be less pronounced in a low-consequence survey experiment than in a high-consequence field experiment. This is indicated in the recent survey experiment of Belgian politicians discussed above (Baekgaard and George, 2018). Here, the authors find politicians to prefer candidates with Arabic names for administrative top management positions for which they are responsible, while real world data show that no person with an Arabic name occupies such a position. Consistent with our study and theories of implicit biases (Cortina, 2008), the politically correct answer is chosen when little is at stake, while conscious or implicit biases may shape real recruitment behaviour.

Our results also illustrate how a survey experiment design might affect participant response. In Study 2, it might be more obvious to the respondent that the separating factor between the two applicants is the ethnicity of the name. Even though this design forces the participant to decide between the two candidates, it also may invite social desirability pressures to come into effect. It may become obvious to the participant what is being studied, which again undermines the treatment (King et al., 2013). Surprisingly, the results in Study 3 seem even more deviant from the results of the field experiment despite the design where participants are only presented with one candidate. This difference might be caused by the difference in measurement as the participants in Study 3 were asked to rate the candidate on a series of questionnaire items. This result should invoke caution for researchers that wish to extrapolate suggested correlations between intentions and actual behaviour to studies of sensitive organizational topics.
Our own study is of course not without its limitations. While Study 1 was inspired by studies employing a between-subjects design, Studies 2 and 3 were inspired by studies using within subject designs. To the best of our knowledge, it has yet to be investigated if this difference in research design matters for the results of studies such as ours. We invite future research to take this possibility under consideration.

Finally, one may argue that the variances between the three studies are driven by the different samples. However, we consider that to be unlikely. Recent field experimental research suggests that public employers discriminate at least as much as those in the private sector (Villadsen and Wulff, 2018), which is in direct contrast to the results of Study 3. Further, the sample in Study 2 consists of a representative sample of managers that are involved in recruitment decisions on a regular basis. Still, we encourage future studies to investigate whether our results can be replicated when using the same sample of participants. Such a study could be designed similarly to Agerström and Rooth’s (2011) study where surveys are administered to the same participants that had taken part in a field experiment months before. In this sense, researchers could aim to have two or more studies with even fewer differences than we have in ours. A single-sample study would constitute a valuable addition to the results presented in this study. To be clear, our goal was to compare different methodological approaches as they are currently used in management and organization research. We contend that our survey experiments are not following best practices as little immersion is achieved, yet they reflect current practices in the field. Together, we believe that our study and a single-sample replication, as sketched above, would be make an important contribution to the ongoing debate on the advantages and drawbacks of different experimental approaches.

---

6 We thank an anonymous reviewer for suggesting that the mentioned design difference might confound our results.
Validity of research methods is a timely topic. Large-scale replication projects in psychology and biomedicine have revealed that a disturbingly low percentage of top journal studies can be successfully replicated (Marsman et al., 2017; Open Science Collaboration, 2015). Our study may offer some interesting opportunities for so-called quasi-replications where the robustness of previous findings are tested against different methods (Bettis, Helfat, and Shaver, 2016). We encourage future replications of studies of organizational topics to test how robust previous findings are if the research design is switched from a survey- to a field experiment or vice-versa.
Appendix A.1. Translated Example of an Application from the Field Experiment

Application for teaching positions at Amager private school

I am ready for teaching the pupils at Amager private school in social studies, history, religion and German while contributing to a prosperous educational environment. In addition, I am able to teach mathematics.

My motivation for applying for the position is that I will be able to apply my skills within my core subjects (German and History) and my knowledge about other course subjects. In addition, it motivates me that I will be able to develop my professional competency at your school while having a focus on the individual pupil. I hope that I will be able to make a difference both for the regular pupils and the challenged ones.

I am very motivated by new assignments and not afraid to meet challenges such as testing new teaching techniques. I have experience as a class teacher and thus experience with the responsibilities this entails. Having graduated recently in 2013, I am very interested in getting experience through courses and continuing education to ensure my continuing professional development. My potential pupils and colleagues can expect that I can contribute with the newest knowledge and that I am going to be a valued employee.

I see great importance in a solid collaboration with pupils, parents and teachers. Thus, it is important for me to have a positive attitude in my work. I look forward to showing you how I can contribute as your new teacher.

I hope that you wish to see me for an interview where we can discuss our common future at Amager private school.

Best wishes,

Mohammad Ashraf

Curriculum Vitae

Personal information:
Name: Mohammad Ashraf
Adresse: Nørregade 75, 1165 København K
Phone.: 25377167
Mail: mohammad.ashraf.pro@gmail.com
Date of birth: 15.03.1987

Education:
2013: Teaching education Aarhus – VIA UC
- Core subjects: Mathematics, German og history
- GPA: 8.7
2008: Ikast sports academy
- Coaching specialization
2007 High school graduate from Aarhus katedralskole
- GPA: 9

Teaching experience:
2013 – 2015: Husum school in Brenshøj
- At Husum School I have taught student in mathematics, German and history as well as social studies and religion. I have acted as class teacher and have arranged academic and social event for the students, e.g. company visits.

2011-2013: Private tutor
- I have worked as a private tutor for three children with math difficulties. This has provided me with experience in working with dyslectic students.

Other relevant experience:
2011-2015 Red Cross visitor
Volunteer work where I have visited elderly and feeble citizens.

2006 – 2008 Substitute at the assisted living facility in Thomasminde

Other:
I am good at IT and have good oral and written competencies especially in English.
I have played a fair share of volleyball and done mountain bike riding. At my school stay I was able to combine interest in sports with a coaching education.
Appendix A.2. Translated Example of Candidate Descriptions from Study 2

Which candidate would you prefer to invite for a job interview?

[Name] is 33 and has an educational background that fits the job description. The candidate has job experience from two former positions that are similar to the one applied for. Recommendations from former employers indicate that [Name] is capable of doing exactly the job required by the company. The candidate’s IT skills are very good and the candidate has experience with the systems that are used in the organization. The candidate does not seem like a good fit with the organizational culture. In the application, the candidate emphasizes personal characteristics that are not particularly compatible with the values important to the hiring organization. The candidate also seems like a bit of a lone wolf.

[Name] is 33 and has experience from an organization similar to your organization. In a recommendation, a former employer mentions that the candidate is a very positive, happy and conscientious worker. Thus, you are confident that the candidate would fit well with your organization’s culture. However, [Name] lacks some of the technical skills that you assess are important for the position. You assess that a substantial amount of training is necessary before the candidate is able to solve the required tasks. The candidate’s prior experience is hardly relevant to what is needed in the position and the candidate lacks experience with the relevant IT-system.

Appendix A.3. Translated Example of Candidate Description from Study 3

[NAME] is 38 years old. He has prior job experience working with assignments similar to those described in the job description. The candidate is familiar with the IT-system used by the hiring organization, which is important to be able to do the job. Therefore, you assess that only minimal training is necessary. The candidate holds a degree that fits the job well and is described as being hard-working and focused with an attention to detail. Still, you get the impression from the application that the candidate works best alone and does not seem like a team player. The candidate is married and has two children and spends his spare time at the local football association.
References


