Original Article

What Can We Learn From Factorial Surveys About Human Behavior?

A Validation Study Comparing Field and Survey Experiments on Discrimination

Knut Petzold[1] and Tobias Wolbring[2]

[1]Department of Sociology, Katholische Universität Eichstätt-Ingolstadt, Eichstätt, Germany
[2]School of Business and Economics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Nürnberg, Germany

Abstract: Factorial survey experiments are increasingly used in the social sciences to investigate behavioral intentions. The measurement of self-reported behavioral intentions with factorial survey experiments frequently assumes that the determinants of intended behavior affect actual behavior in a similar way. We critically investigate this fundamental assumption using the misdirected email technique. Student participants of a survey were randomly assigned to a field experiment or a survey experiment. The email informs the recipient about the reception of a scholarship with varying stakes (full-time vs. book) and recipient's names (German vs. Arabic). In the survey experiment, respondents saw an image of the same email. This validation design ensured a high level of correspondence between units, settings, and treatments across both studies. Results reveal that while the frequencies of self-reported intentions and actual behavior deviate, treatments show similar relative effects. Hence, although further research on this topic is needed, this study suggests that determinants of behavior might be inferred from behavioral intentions measured with survey experiments.

Keywords: factorial survey experiment, field experiment, behavioral validity, validation study, misdirected email technique

Factorial survey experiments and related methods are increasingly used in sociology and beyond (see Wallander, 2009) because combining experimental designs with survey methods provides several advantages, especially regarding causal inference and generalizability (see Auspurg & Hinz, 2015; Mutz, 2011). While Rossi (1979; Rossi & Anderson, 1982) had originally proposed the factorial survey approach in the late 1970s to analyze attitudes and normative judgments, the method is now increasingly used to investigate behavioral intentions, for instance, regarding hiring decisions (Di Stasio, 2014), family migration (Abraham, Auspurg, & Hinz, 2010), or bribing of academic staff (Graeff, Sattler, Mehlkop, & Sauer, 2014).

However, the assumption that determinants of behavioral intentions reported in hypothetical situations reveal information about influences on real-life behavior is potentially problematic (see Auspurg & Hinz, 2015, p. 11). The "behavioral validity" is therefore a key issue in survey experiments on behavioral intentions (Berger & Wolbring, 2015). Surprisingly, empirical evidence on the behavioral validity of survey experiments is still scarce and, moreover, rather ambiguous. Some studies indicate that the determinants that explain hypothetical choices made in survey experiments largely correspond to the determinants that explain actual behavior made in similar real-world situations (Hainmueller, Hangartner, & Yamamoto, 2015; Nisic & Auspurg, 2009; Raub & Buskens, 2008). Other authors conclude that though the proportion of actual behavior can hardly be determined by survey experiments, at least the relative effects of situational attributes on behavior point in the same direction (Diehl, Andorfer, Khoudja, & Krause, 2013; Eifler, 2010; Groß & Börensen, 2009). However, a roughly equal number of studies provide contradictory results, reporting large differences between hypothetical behavior and decisions made in real situations (Eifler, 2007; Findley, Laney, Nielson, & Sharman, 2017; Pager & Quillian, 2005).

Taken together, behavioral validation studies on factorial survey experiments provide mixed evidence; hence, further research is sorely needed. Two fundamental problems of many existing validation studies are that (a) estimates used as a behavioral benchmark might suffer from biases due to a weak research design and (b) the samples for the factorial survey and the behavioral benchmark systematically differ,

thus limiting comparability. These methodological flaws have likely contributed to the ambiguous state of research, since empirically observed differences between hypothetical and actual behavior might result from differences in sampling composition or biased estimates of the behavioral benchmark (cf. Eifler & Petzold, in press).

In contrast to former studies, we conducted a validation design using an unobtrusive field experiment to maximize the validity of our behavioral benchmark estimates. We secured the comparability of the samples in the field, surveyed experiments by randomization, and used identical operationalizations of treatments and covariates. By doing so, we contribute to the continuing small body of methodological literature on the behavioral validity of survey experiments.

The Behavioral Validity of Survey Experiments

When Do Self-Reported Intentions Predict Real-Life Behavior?

When survey experiments aim for the measurement of self-reported behavioral intentions under varying conditions, the prediction of real-life behavior might be limited for at least three reasons.

First, even in real-life situations, intentions do not always translate into behavior. Situational and personal resources and obstacles may represent high costs of performing a behavior (cf. Diekmann & Preisendörfer, 2003). The resulting low level of behavioral control makes the behavior in question less likely (Fishbein & Ajzen, 2010). In their meta-analysis, Armitage and Conner (2001) found an overall correlation between intentions and behavior to be only $r = .47$.

Second, the information that is taken into consideration when filling out a questionnaire can differ from the available information in real life. As a result, self-reported intentions may differ from intentions that are present in real life. As reality cannot be reconstructed in all its details with the help of vignettes (Hughes & Huby, 2004), self-reported intentions measured with vignettes are prone to a "hypothetical bias" (Ajzen, Brown, & Carvajal, 2004). The hypothetical bias can be reduced if perceived behavioral control of the hypothetical situation is brought in line with actual control in real life. The more realistic the respondent's induced expectations and evaluations in a situation are, the better should the intentions assessed in a questionnaire predict actual behavior (see Fishbein & Ajzen, 2010, p. 62). Since the evaluation of constraints to hypothetical behavior might be especially prone to bias in high-cost settings, we look upon a low-cost setting in this study where the intention–behavior relation is stronger.

Third, self-reported intentions may also be distorted by social desirability (Tourangeau & Yan, 2007). Respondents may deviate from their "true" responses as they strive to achieve social approval. This social desirability bias can thus be understood as a specific type of hypothetical bias where perceived normative expectations differ between hypothetical situations and real life. Even though the factorial survey technique attenuates the social desirability bias as compared to direct questioning (Armacost, Hosseini, Morris, & Rehbein, 1991), intentions may still represent desirability beliefs (Stocké, 2007).

What Makes a Good Validation Study?

The validity of experimental results is usually evaluated in terms of two major criteria (Shadish, Cook, & Campbell, 2002): While "internal validity" refers to the question of causal inference, "external validity" addresses the extrapolation of causal inferences to variations in units, settings, treatments, and outcomes (Campbell, 1957; Cronbach, 1982). From this perspective, "behavioral validity" is a specific type of external validity that particularly concerns inference about the extent to which a causal effect of a treatment variable on observed behavioral intentions can be generalized to other outcome measures, particularly to observations of real behavior.

As the generalizability of a causal effect may be generally threatened by interactions with units, treatments, settings, or outcomes (Shadish et al., 2002, pp. 83–102), testing the validity over variations specifically in the outcome variable requires exclusion of all other potential interactions. Hence, a good behavioral validation study is the replication of a factorial survey experiment aiming for maximum similarity with respect to units, setting, and treatments while varying the outcome by measuring actual behavior instead of self-assessed behavioral intentions.

To avoid interactions of an effect with units, the samples of the survey experiment and the respective study delivering the behavioral benchmark must not differ systematically. Pre-post designs, randomization, and ex-post stratification are ways to adjust for heterogeneities across study conditions (Morgan & Winship, 2015). To capture influences by variations in the setting, the decision situation of the replication study needs to be as similar as possible to the survey experimental scenario. This is not trivial, since the written vignette descriptions of a survey experiment cannot capture all details that may affect decisions in real life. Finally, observed effects may vary with differing treatments across the survey experiment and its replication study. In survey experiments, global categorizations are frequently applied as treatment variables, such as "men" or

"women", while in real-world situations, more specific treatment variables are used, such as specific names indicating male or female gender. To avoid lacking comparability of effect estimates due to differences in treatment operationalization, the specific treatment conditions should be constructed with as much similarity as possible.

In sum, the ideal design for behavioral validation of survey experimental results contains different measures of the outcome while holding constant units, treatments, and settings. In this study, we demonstrate a validation approach that tries to meet these criteria.

Research Design

We use an experiment to isolate the effect of research design on the difference between intended versus actual behavior. Comparable survey and field experiments were conducted, while subjects were randomly allocated to the two research design conditions.

Field Experiment

For the field experiment, we used the misdirected email technique (Stern & Faber, 1997; Vaes, Paladino, & Leyens, 2002). The basic idea is that subjects receive an important email which is obviously misdirected. Henceforth, they have to decide whether to inform the sender of the email about the mistake, to forward the email to the actual addressee of the message (if available), or to simply ignore/delete the email. The design has been particularly used to uncover discrimination against Arabs (Bushman & Bonacci, 2004; Tykocinski & Bareket-Bojmel, 2009).

The misdirected email technique appeared to us as an ideal method for a behavioral validation of survey experiments. First, in a concealed field experiment, unobtrusive measurement of the outcome is possible that might otherwise be biased due to social desirability in surveys. Second, assuming that all respondents received the email, nonresponse bias can be ruled out, thus securing sample comparability. Third, the low-cost setting of the study ensures sufficient behavioral control so that intended behavior can be actually executed in the field.

We followed the example by Bushman and Bonacci (2004) and sent misdirected emails about scholarship admission to university students. The email informed the receiver that the fictitious "Hans-Albert-Foundation" awards a full-time scholarship (€ 897 per month) or a book scholarship (€ 100 per month) either to a student with a German name ("Tobias Müller"/"Christoph Winter") or to a student with an Arabic name ("Rashid Yassir"/"Tarik El Morabet") (2 × 2 design). However, the recipient has to confirm acceptance of the scholarship within 1 month. Figure 1 gives an example for the emails sent.[1]

To increase the plausibility and credibility of the emails, messages were sent via the email account of one of the authors, who was assistant professor at the students' university during the study. The misdirected email did not contain the email address of the intended receiver of the message. However, since the actual receiver could not apply for the scholarship and since the salutation contained a different name, it was apparent to everybody that this email was misdirected. Several (concealed) pretests as well as students' reactions to the misdirected email in the actual study corroborated this assumption.

The outcome variable was measured on the basis of whether the false receiver highlighted the mistake. Overall, 37.9% of all subjects responded to our email and made us aware of the problem.

Survey Experiment

The survey experiment was set up in a very similar way as the field experiment. Respondents were presented a description of the hypothetical scenario: "Suppose you open your email account and see the following email that is obviously misdirected". We then showed the respondents a screenshot of the email and asked them about their anticipated reaction in this hypothetical situation on a rating scale, in manner similar to when intentions are measured with vignettes: "On a scale from 0 (= *very unlikely*) to 10 (= *very likely*), how likely will you respond to this email and highlight the misdirection of the email to the sender?"[2] With a mean value of 8.94 (SD = 2.12), the large majority of respondents stated that they would definitely respond to the email (66.5%).

As is usually done in factorial survey experiments, we used a within-design showing each respondent not only one but each of the four possible combinations of the two

[1] It must be noted that the specific names were uniquely presented together with the specific scholarships in order to avoid misinterpretations in the factorial survey. Otherwise due to within-design in the survey experiment (see Survey Experiment section), respondents might believe that one and the same person receives two scholarships at the same time. Though this decision leads to confounding names and scholarships in the treatment, it does not affect the results of our validation study because we chose the same design across both experimental types. We address this issue also in the section on estimation methods (see Footnote 4).

[2] While we explained in section "What makes a good validation study?" that comparability of measurement is important for a good validation study, we used different measures of the outcome variable on purpose. The reason is that it is standard practice in factorial surveys to use rating scales. Not following this standard practice and using binary outcome measures instead might seriously limit the transportability of our findings to actual survey experiments.

To:
rashid_yassir@gmail.com | tobias_mueller@gmail.com
christoph.winter@gmx.net | tarik_el_morabet@yahoo.com

Subject:
Admitted as fellow by the Hans-Albert-Förderfonds for students |
Book scholarship of the Hans-Albert-Förderfonds for students

Dear Mr **Yassir | Müller | El Morabet | Winter**,

You applied for a **full-time | book** scholarship for students, donated by the Hans-Albert-Förderfonds, at the Faculty of Social Sciences of the University of Mannheim.

We are pleased to inform you that you have been chosen for the scholarship for the academic year 2016 by virtue of the vote of our selection committee. **The amount of the scholarship consists of a monthly study fee of € 300.00 and monthly living expenses of up to € 597. In total, you receive € 897 per month. In addition, grants for health and long-term care insurance may be granted. | A monthly amount of € 100 will be paid in order to cover expenses for work and study materials.** The scholarship will be granted to you for a period of one year and can be renewed for two further years at maximum, on the basis of your annual interim reports.

We kindly ask you to inform us as soon as possible whether you will take advantage of the scholarship granted to you from 1st January 2016, but after the expiry of a week at the latest; start at a later date would have to be negotiated separately.

If you want to start with the scholarship, you will immediately receive a form for personal information, which we urgently need to release your scholarship. If you decide not to join the scholarship, it will be awarded to another applicant.

We are looking forward to welcoming you as a fellow of the new academic year at the opening session of the Hans-Albert-Förderfonds (invitation follows).

Best regards

Tobias Wolbring
Chairman of the Selection Committee

--
** Contact information**

Note: Varied treatment conditions indicated by bold font.

Figure 1. Text versions of the misdirected email. Varied treatment conditions indicated by bold font.

binary treatments (Arab name/German name; high/low stakes) in order to increase statistical power.

Sampling and Measures

As explained above, a major problem in most validation studies is that the compositions of the samples in the survey experiment and in the replication study differ. One key reason for this is nonresponse in surveys; another issue is simply drawing comparable samples for studying actual decision-making. As a consequence, it remains largely unclear whether differences between the survey experiment and the field experiment are due to design effects or due to differences in sample compositions (cf. Eifler & Petzold, in press).

To avoid this problem, we decided to conduct one joint survey among all participants of both studies. We contacted 702 students of the University of Mannheim who had enrolled in an experimental subject pool and asked them to participate in an online survey on "Living together in Germany".[3] As an incentive for participation, students could win one € 100 and ten € 10 Amazon vouchers. Two hundred and eighty students participated in the survey, resulting in a response rate of 39.9%. Note that this

[3] We want to thank Felix Bader for collecting the email addresses and managing the subject pool.

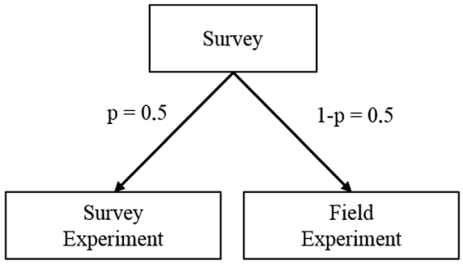

Figure 2. Research design.

convenience sample is not problematic for the internal validity of our validation study due to the experimental design we use.

Within the survey, respondents were randomly assigned to two experimental conditions, namely survey experiment and field experiment (see Figure 2). Both groups completed the general survey containing scales on attitudes toward foreigners and Muslims, and questions on sociodemographics. The first experimental group additionally completed the survey experiment, while the second experimental group received a misdirected email afterward.

It is worth highlighting that, besides securing comparability of the groups, conducting a joint survey among all subjects brings the further advantage of securing measurement equivalence with respect to covariates – another assumption that is usually made but not necessarily revealed by the researchers when conducting validation studies.

Table 1 shows that the assignment of participants was proportional and gained balanced allocations across the four experimental conditions in both survey experiment and field experiment. For all variables, group differences are small and not statistically significant, indicating that the random assignment to the two study conditions worked well.

Furthermore, since we used an everyday situation of email receipt and presented an image of the email just as it is usually displayed in Internet browsers, we claim a highly realistic vignette presentation which assures high construct validity. The varied treatments of receiver's name and scholarship amount are as similar as possible in both studies to avoid biases through interactions between treatments and effects.

Estimation Methods

We analyze the survey experiment and the field experiment in a first step separately and then pool the data to provide a more direct comparison between both designs. The outcome of the survey experiment is a variable ranging from

Table 1. Sample comparison across experimental conditions and designs

	Total		Field experiment		Survey experiment		
Experimental conditions	*M* (*SD*)/*N*	Range/%	*M* (*SD*)/*N*	Range/%	*M* (*SD*)/*N*	Range/%	t/χ^2 (*p*)
Arab/low amount			33	23.6%	140	25.0%	
Arab/high amount			35	25.0%	139	24.9%	
German/low amount			35	25.0%	140	25.0%	
German/high amount			37	26.4%	140	25.0%	2.796 (.424)
							(First vig. only)
Covariates							
Age	23.13 (3.69)	19–46	22.85 (3.32)	19–39	23.41 (4.02)	19–46	−1.263 (.207)
Sex							
Male	103	36.8%	48	34.3%	55	39.3%	
Female	177	63.2%	92	65.7%	85	60.7%	0.752 (0.386)
Religion							
Catholic	93	33.2%	47	33.6%	46	32.9%	
Protestant	83	29.6%	46	32.9%	37	26.4%	
Other Christian	16	5.7%	4	2.9%	12	8.6%	
Non-Christian	8	2.9%	3	2.1%	5	3.6%	
Nonreligious	80	28.6%	40	28.6%	40	28.6%	5.486 (0.241)
Origin							
Foreign	33	11.8%	12	8.6%	21	15.0%	
German	247	88.2%	128	91.4%	119	85.0%	2.782 (0.095)
Att. foreigners (support)	3.41 (0.79)	1–5	3.45 (0.81)	1–5	3.37 (0.79)	1.5–5.0	0.877 (0.381)
Att. foreigners (protection)	4.13 (0.82)	1–5	4.22 (0.72)	1.75–5.00	4.04 (0.89)	1–5	1.845 (0.066)
Att. Muslims	3.75 (0.78)	1–5	3.80 (0.76)	1.33–5.00	3.81 (0.80)	1–5	0.964 (0.335)
$N_{Persons}$	280		140		140		
$N_{Vignettes}$	559				559		

0 to 10 so that models for continuous or categorical outcomes can be applied. Since each participant assessed a number of misdirected emails, the data structure is hierarchical (Hox, Kreft, & Hermkens, 1991; Jasso, 2006). We capture this multilevel structure by using random intercept models (Snijders & Bosker, 2012).

In contrast to the survey experiment, the field experiment provides a simple binary outcome coded with 0 (no support) and 1 (support by highlighting misdirection) so that probit or logit models are appropriate. However, for the sake of comparability with the findings for the survey experiments, we estimate a linear probability model (for robustness checks, see Footnote 4). This is particularly recommended when interaction terms are included (see Best & Wolf, 2015), as is the case in the joint model. For both models, robust standard errors are reported due to potential heteroscedasticity.

To provide a more direct comparison between the survey experiment and the field experiment, we then estimate a joint model using the pooled data from both experiments and specifying a linear regression model. Since respondents of the survey experiment expressed their behavioral intentions for up to four hypothetical vignette emails (within-design), we again estimate random intercept models. Besides main effects of treatments (name and stake) and experimental design, the joint model contains multiplicative interaction terms of the latter and the former variables. We also included interaction terms between the experimental mode and all covariates in the joint model to control for remaining differences in sample compositions (see Morgan & Winship, 2015). Yet, the outcome variables in the two experiments are not directly comparable, since the measure for the field experiment is dichotomous, whereas respondents stated their intentions in the survey experiment on a ten-point scale. For the joint model, we thus decided to reduce information from the survey experiments by splitting the outcome variable and binary coding the measure for behavioral intention with 1 (= *very likely*) and 0 (= *below very likely*).[4]

Results

As displayed in Figure 3, behavioral intentions in the survey and observed behavior in the field differ substantially. While the vast majority (66.5%) of the respondents in the survey experiment claimed that they would definitely highlight the misdirection of the email, only a minority (37.9%) showed this behavior in the field experiment. Hence, it does not appear advisable to use the distribution of the outcome in the factorial survey to predict real-world behavior, as the measures of self-reported intentions may suffer from a hypothetical bias such as social desirability. However, whether this bias also affects treatment estimates is an empirical question.

Table 2 shows estimates from separate models for the survey experiment and the field experiment. In the survey experiment, we find no effect of recipient's name on respondents' intention to help, but a significantly higher inclination to highlight the misdirection of the email if a full-time scholarship instead of a book scholarship is provided. Results of the field experiment corroborate this finding. The probability to highlight the obvious misdirection of the email to the sender is not significantly influenced by the recipient's name in the field. However, it increases by 20% points if the email informs about the receipt of a full-time scholarship as compared to a book scholarship. In other

[4] In addition to the models reported, the estimation results of both experiments have undergone a number of robustness checks:
(i) Regarding the survey experiment, the results of nested linear random intercept models containing all vignettes and of nested OLS models containing only the first vignettes have been calculated (see Table A2 in the Appendix). By doing so, we ruled out learning effects as the estimation is based on the response to the first hypothetical email presented. In addition, since all treatment coefficients are quite similar, randomization succeeded in balancing covariates across treatment conditions. Moreover, since the distribution of the outcome variable of the survey experiment is heavily skewed (see Table A1 in the Appendix), generalized linear models and logit estimations for several binary versions of the outcome variable (middle of the scale, < 10) were applied, all corroborating our findings.
(ii) Regarding the field experiment, all results have been replicated using nested logit models with and without covariates to check randomization quality. For means of effect comparisons between models, we calculated average marginal effects (AMEs) as recommended by Mood (2010) and Karlson, Holm, and Breen (2012). The average marginal effects are almost identical to the non-standardized coefficients provided by the linear models (see Table A3 in the Appendix).
(iii) To explore how the dichotomization of the intention scale in the joint model affected our findings, we also estimated pooled logit models using different split thresholds (e.g., at the middle of the scale) and estimated linear models dividing the outcome measure for the survey experiment by ten (resulting in a variable ranging from 0 to 1 in ten 0.1 steps). Although effect strengths vary with choice of operationalization and model specification, our findings are robust regarding direction and significance of the effects in the joint models.
(iv) Numerous additional models have been estimated for both experiments with further covariates, including social desirability scales and several interactions between treatments, and between treatments and covariates. All checks revealed that the results are remarkably robust; this further confirms that randomization worked well at both stages of the experimental design.
(v) Since the Arab and German names were confounded with scholarship amount, we controlled for this in a further robustness check. Separate regressions for each experiment (see Table A5 in the Appendix) and pooled regressions including interaction terms between treatments and experimental type (see Table A6 in the Appendix) were reestimated using the detailed treatments (exact names) instead of the categorical treatments (Arab/German). The results corroborate our interpretation.

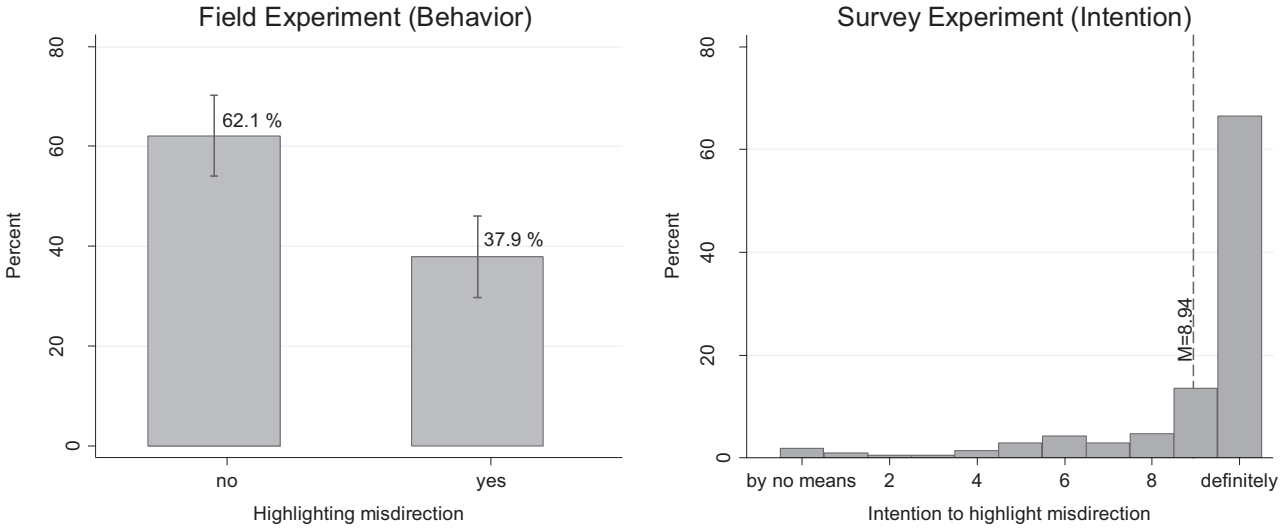

Figure 3. Outcomes of the field experiment and the survey experiment

Table 2. Estimations for the survey experiment and for the field experiment

	Support (Intention)	Support (Behavior)
	Survey experiment	Field experiment
Recipient's name (foreign, ref. native)	−0.006 (−0.13)	−0.045 (−0.50)
Amount of scholarship (high, ref. low)	0.308*** (3.90)	0.205* (2.30)
$N_{Vignettes}$	559	
$N_{Persons}$	140	140
$\sigma_{Persons}$	1.802	
R^2_{within}	0.059	
$R^2_{between}$	0.240	
$R^2_{overall}$	0.225	0.077
Wald χ^2/F	49.97***	1.20

Notes. Model survey experiment = Random intercept estimation, non-standardized beta coefficients, robust standard errors, *z*-values in parentheses; Model field experiment = OLS estimation, non-standardized beta coefficients, robust standard errors, *t*-values in parentheses; Covariates in both models = age, sex, religion, origin, attitudes toward foreigners and Muslims. *$p < .05$, **$p < .01$, ***$p < .001$.

words, just like respondents in the survey experiment, participants of the field experiment generally reacted to the stakes of the scholarships but not to the recipient's name.

To systematically test the robustness of the treatment effects across both experiments, we estimated joint models, including interaction terms between treatments and study-design (Table A4 in the Appendix). In these models, both treatment effects are stronger in the field experiment than in the survey experiment, which provides suggestive evidence for a social desirability bias in survey experiments. However, the interaction of experimental design neither with the recipient's name nor with the type of scholarship is statistically significant. Thus, although this study does not have sufficient statistical power to determine whether the clearly existing differences in effect size are systematic or simply due to chance, we conclude that – despite deviations through social desirability – the survey experiment and the field experiment lead to quite similar findings, at least with respect to direction, relative strength, and statistical significance of treatment effects.

Discussion

Even though a growing number of survey experiments measure behavioral intentions, there is mixed evidence regarding the behavioral validity of survey experiments. Since many validation studies may suffer from methodological flaws such as differing research designs and samples, their results are limited. Against this background, we designed a study that offers a high degree of comparability and conducted an unobtrusive field experiment that serves as the behavioral benchmark for a survey experiment on discrimination in everyday life. We used the misdirected email technique and measured whether false receivers of a misdirected email highlighted the mistake. We avoided interactions of the effects with units, settings, and treatments by a random assignment of the participants to field and survey experiments and by the identical operationalization of treatments and covariates (identical email).[5]

We found large differences between the frequencies of behavioral intentions and actual behavior. Hence, a first lesson from this study is not to generalize the distributions of self-reported behavioral intentions to distributions of actual behavior in real-world situations. However, like in

[5] For a similar validation design, but based on propensity score matching and using a status experiment, see Eifler and Petzold (in press).

other studies (Eifler, 2010; Groß & Börensen, 2009), the estimates of the treatment effects are quite similar across the field experiment and the survey experiment in terms of direction, relative effect size, and statistical significance. Thus, a second lesson from this study is that researchers, when thinking about determinants of real-life behavior, should solely focus on directions and relative effect sizes in survey experiments instead of relative frequencies of intended actions.

Nonetheless, considering the overreported, less varying intentions for prosociality, our results from the factorial survey indicate that they may still be biased due to social desirability. On the one hand, respondents might perceive a lack of anonymity and the predefined answer categories might communicate a social expectation to report a redirection of the email in the survey experiment. On the other hand, we varied only two dimensions in the survey experiment for the sake of comparability with the field experiment. Hence, all possible combinations were rated by individual respondents. This setup increases the risk that respondents discover the substantive focus of the study and react in a socially desirable way. We addressed this issue in our robustness checks and found no intention toward ethnic discrimination by estimating the effects of the first rated vignettes only, where respondents were unlikely to detect the varied levels. However, the less complex design using the full factorial may still have increased social desirability bias. This problem would probably not occur in more common applications of factorial survey experiments, varying more dimensions and presenting only a small subset of the vignette universe to the respondents in order to systematically counter the tendency toward socially desirable response behavior (see Auspurg, Hinz, Liebig, & Sauer, 2015; Mutz, 2011), especially as compared to techniques of direct questioning (see Armacost et al., 1991).

We found evidence of discrimination against Arabs neither in the field experiment nor in the survey experiment, so that the results appear to be in line with each other. One explanation for this finding is that there is indeed no ethnic discrimination in the field because of the rather small and very homogenous sample of a student population from a city with a large Muslim minority population. Yet, there might be alternative explanations for the lacking ethnicity effect, such as insufficient statistical power to detect small differences or the chosen low-cost setting which eases norm compliance. Some of the emails in the field experiment might also have been classified as spam (automatically or by the respondent), resulting in measurement error of the outcome variable in the field experiment. Since the reasons can differ between the survey experiment and the field experiment, the finding that ethnicity does not affect outcomes in the experiments should be interpreted with some caution.

This leads to the more general question of the generalizability of our results to other settings. Taking the former evidence on the comparability of survey and field experiments into consideration, there are three types of results: clear correspondence, large differences, and similar treatment effects but differing outcome distributions. Some studies report a clear correspondence between both experimental modes. This has been shown, for example, for voting about the naturalization of immigrants (Hainmueller et al., 2015), for making a residential choice (Nisic & Auspurg, 2009), or for buying a used car (Raub & Buskens, 2008). In most of these situations, few normative expectations seem to exist, so they appear to be hardly threatened by a social desirability bias.

Other studies reveal large differences. For instance, inclinations toward theft by finding (Eifler, 2007), toward hiring former offenders (Pager & Quillian, 2005), or toward supporting anonymous business incorporations (Findley et al., 2017) could not be replicated at all with vignettes. The lack of correspondence seems especially evident when sensitive topics are investigated, making the survey prone to socially desirable responses, while in the field experiment, norm-compliant behavior is canceled out by costly restrictions.

As in our study, a third group of studies detected similar treatment effects but differing outcome distributions. Respective results were observed for behavior at a traffic light (Groß & Börensen, 2009), for prosocial behavior when redirecting a lost letter (Eifler, 2010), and for ethnic discrimination in shared housing among university students (Diehl et al., 2013). In those decision problems, socially desirable response behavior seems to be consistent with socially desired realized behavior due to relatively low costs of real behavior.

Hence, survey experiments may provide valid measures of behavioral intentions when there exist either few normative expectations toward a specific behavior in a situation (no social desirability) or low costs of performing norm-compliant behavior in the real world, resulting in similar socially desirable actions in the field and responses in the survey. In turn, survey experiments may not provide valid measures of behavioral intentions when norm-complying survey responses are convenient, while norm-compliant behavior implies high costs of action. Since our study was conducted using a specific low-cost setting of prosocial behavior in an everyday life situation, it provides further evidence for this conclusion.

Finally, we want to highlight that – due to the experimental comparison between the two experiments and identical treatment operationalizations – the results can be considered being highly internally valid. In our view, it is worth pursuing this approach in future validation studies to fix any differences in samples, units, and treatments and only vary the outcome. A number of other high- and low-cost settings, treatments, and samples should be investigated

to further clarify the circumstances under which behavioral intentions measured in survey experiments are generalizable to performed behavior in real life.

References

Abraham, M., Auspurg, K., & Hinz, T. (2010). Migration decisions within dual-earner partnerships: A test of bargaining theory. *Journal of Marriage and Family, 72*, 876–892. https://doi.org/10.1111/j.1741-3737.2010.00736.x

Ajzen, I., Brown, T. C., & Carvajal, F. (2004). Explaining the discrepancy between intentions and actions: The case of hypothetical bias in contingent valuation. *Personality and Social Psychology Bulletin, 30*, 1108–1121. https://doi.org/10.1177/0146167204264079

Armacost, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An empirical comparison of direct questioning, scenario, and randomized response methods for obtaining sensitive business information. *Decision Sciences, 22*, 1073–1099. https://doi.org/10.1111/j.1540-5915.1991.tb01907.x

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology, 40*, 471–499. https://doi.org/10.1348/014466601164939

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. London, UK/Thousand Oaks, CA: Sage Publications.

Auspurg, K., Hinz, T., Liebig, S., & Sauer, C. (2015). The factorial survey as a method for measuring sensitive issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving survey methods: Lessons from recent research* (pp. 137–150). New York, NY/Hove, UK: Routledge.

Berger, R., & Wolbring, T. (2015). Kontrafaktische Kausalität und eine Typologie sozialwissenschaftlicher Experimente [The counterfactual approach to causal inference and a typology of social science experiments]. In M. Keuschnigg & T. Wolbring (Eds.), *Experimente in den Sozialwissenschaften. Soziale Welt Sonderband 22* (pp. 34–52). Baden-Baden, Germany: Nomos.

Best, H., & Wolf, C. (2015). Logistic regression. In H. Best & C. Wolf (Eds.), *The Sage handbook of regression analysis and causal inference* (pp. 153–172). London, UK: Sage.

Bushman, B. J., & Bonacci, A. M. (2004). You've got mail: Using e-mail to examine the effect of prejudiced attitudes on discrimination against Arabs. *Journal of Experimental Social Psychology, 40*, 753–759. https://doi.org/10.1016/j.jesp.2004.02.001

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297–312. https://doi.org/10.1037/h0040950

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Di Stasio, V. (2014). Education as a signal of trainability: Results from a vignette study with Italian employers. *European Sociological Review, 30*, 796–809. https://doi.org/10.1093/esr/jcu074

Diehl, C., Andorfer, V. A., Khoudja, Y., & Krause, K. (2013). Not in my kitchen? Ethnic discrimination and discrimination intentions in shared housing among university students in Germany. *Journal of Ethnic and Migration Studies, 39*, 1679–1697. https://doi.org/10.1080/1369183X.2013.833705

Diekmann, A., & Preisendörfer, P. (2003). Green and greenback: The behavioral effects of environmental attitudes in low-cost and high-cost situations. *Rationality and Society, 15*, 441–472. https://doi.org/10.1177/1043463103154002

Eifler, S. (2007). Evaluating the validity of self-reported deviant behavior using vignette analyses. *Quality and Quantity, 41*, 303–318. https://doi.org/10.1007/s11135-007-9093-3

Eifler, S. (2010). Validity of a factorial survey approach to the analysis of criminal behavior. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 139–146. https://doi.org/10.1027/1614-2241/a000015

Eifler, S., & Petzold, K. (in press). Validity aspects of vignette experiments. Expected 'What-If' differences between reported and actual behavior. In P. J. Lavrakas, E. D. de Leeuw, A. L. Holbrook, & C. Kennedy (Eds.), *Experimental methods in survey research: Techniques that combine random sampling with random assignment*. Hoboken, NJ: John Wiley & Sons.

Findley, M. G., Laney, B., Nielson, D. L., & Sharman, J. C. (2017). External validity in parallel global field and survey experiments on anonymous incorporation. *The Journal of Politics, 79*, 856–872. https://doi.org/10.1086/690615

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior. The reasoned action approach*. Hove, UK: Psychology Press.

Graeff, P., Sattler, S., Mehlkop, G., & Sauer, C. (2014). Incentives and inhibitors of abusing academic positions: Analysing university students' decision about bribing academic staff. *European Sociological Review, 30*, 230–241. https://doi.org/10.1093/esr/jct036

Groß, J., & Börensen, C. (2009). Wie valide sind Verhaltensmessungen mittels Vignetten? Ein methodischer Vergleich von faktoriellem Survey und Verhaltensbeobachtung [How valid are measures of behavior using vignettes? A comparison of factorial survey and observed behavior]. In P. Kriwy & C. Gross (Eds.), *Klein aber fein! Quantitative Sozialforschung mit kleinen Fallzahlen* (pp. 149–178). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences, 112*, 2395–2400. https://doi.org/10.1073/pnas.1416587112

Hox, J. J., Kreft, I. G., & Hermkens, P. L. J. (1991). The analysis of factorial surveys. *Sociological Methods and Research, 19*, 439–510. https://doi.org/10.1177/0049124191019004003

Hughes, R., & Huby, M. (2004). The construction and interpretation of vignettes in social research. *Social Work & Social Sciences Review, 11*, 36–51.

Jasso, G. (2006). Factorial survey methods for studying beliefs and judgments. *Sociological Methods and Research, 34*, 334–423. https://doi.org/10.1177/0049124105283121

Karlson, K. B., Holm, A., & Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit: A new method. *Sociological Methodology, 42*, 286–313. https://doi.org/10.1177/0081175012444861

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*, 67–82. https://doi.org/10.1093/esr/jcp006

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge, UK: Cambridge University Press.

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton, NJ: Princeton University Press.

Nisic, N., & Auspurg, K. (2009). Faktorieller Survey und Klassische Bevölkerungsumfrage im Vergleich – Validität, Grenzen und Möglichkeiten beider Ansätze [Comparing factorial survey and classical survey. Validity, limitations and potentials of both approaches]. In P. Kriwy & C. Gross (Eds.), *Klein aber fein! Quantitative Sozialforschung mit kleinen Fallzahlen* (pp. 211–245). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Pager, D., & Quillian, L. (2005). Walking the talk? What employers say versus what they do. *American Sociological Review, 70*, 355–380. https://doi.org/10.1177/000312240507000301

Raub, W., & Buskens, V. (2008). Theory and empirical research in analytical sociology: The case of cooperation in problematic social situations. *Analyse & Kritik, 30*, 689–722. https://doi.org/10.1515/auk-2008-0218

Rossi, P. H. (1979). Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman, & P. H. Rossi (Eds.), *Qualitative and quantitative social research: Papers in honor of Paul F. Lazarsfeld* (pp. 176–186). New York, NY: Free Press.

Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach: An introduction. In P. H. Rossi & S. L. Nock (Eds.), *Measuring Social Judgments. The Factorial Approach* (pp. 15–67). Beverly Hills, CA: Sage Publications.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA/New York, NY: Houghton Mifflin.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Los Angeles, CA: Sage.

Stern, S. E., & Faber, J. E. (1997). The lost e-mail method: Milgram's lost-letter technique in the age of the internet. *Behavior Research Methods, Instruments, & Computers, 29*, 260–263. https://doi.org/10.3758/BF03204823

Stocké, V. (2007). Determinants and consequences of survey respondents' social desirability beliefs about racial attitudes. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*, 125–138. https://doi.org/10.1027/1614-2241.3.3.125

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859–883.

Tykocinski, O. E., & Bareket-Bojmel, L. (2009). The lost e-mail technique: Use of an implicit measure to assess discriminatory attitudes toward two minority groups in Israel. *Journal of Applied Social Psychology, 39*, 62–81. https://doi.org/10.1111/j.1559-1816.2008.00429.x

Vaes, J., Paladino, M.-P., & Leyens, J.-P. (2002). The lost e-mail: Prosocial reactions induced by uniquely human emotions. *British Journal of Social Psychology, 41*, 521–534. https://doi.org/10.1348/014466602321149867

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research, 38*, 505–520. https://doi.org/10.1016/j.ssresearch.2009.03.004

Received April 29, 2018
Revision received July 30, 2018
Accepted August 30, 2018
Published online December 12, 2018

Tobias Wolbring
School of Business and Economics
FAU Erlangen-Nürnberg (FAU)
Findelgasse 7/9
90402 Nürnberg
Germany
tobias.wolbring@fau.de

Knut Petzold
Department of Sociology
Katholische Universität Eichstätt-Ingolstadt
Ostenstraße 26
85072 Eichstätt
Germany
knut.petzold@ku.de

Knut Petzold is sociologist with expertise in experimental methods in social science research. His current research interests include specific applications in research on higher education and work, international mobility and migration, and social inequality.

Tobias Wolbring holds the chair of empirical economic sociology at Friedrich-Alexander-University Erlangen-Nürnberg. Current research interests comprise economic sociology, (higher) education, and causal inference.

Appendix

Table A1. Distribution of outcomes in field experiment and survey experiment

Field Experiment: Support (Behavior)	
Support (= 1)	53 (37.86%)
No support (= 0)	87 (62.14%)
Total	140 (100.00%)
Survey Experiment: Support (Intention)	
Range	0–10
Mean	8.945
Standard deviation	2.123
Variance	4.511
Skewness	−2.539
Kurtosis	9.335
Total	559
Support (Int. Split)	
Support (= 10)	372 (33.45%)
No support (= 0 < 10)	187 (66.55%)
Total	559 (100.00%)

Table A2. Estimation results of the survey experiment

	Support (Intention)			Support (Intention)/First vignette		
	Model 1a	Model 2a	Model 3a	Model 1b	Model 2b	Model 3b
Recipient's name (foreign, ref. native)	−0.006 (−0.13)	−0.006 (−0.13)	−0.006 (−0.13)	0.201 (0.54)	0.185 (0.48)	0.081 (0.24)
Amount of scholarship (high, ref. low)	0.308*** (3.94)	0.308*** (3.91)	0.308*** (3.90)	0.555 (1.56)	0.616 (1.69)	0.749* (2.26)
$N_{Vignettes}$	559	559	559			
$N_{Persons}$	140	140	140	140	140	140
$\sigma_{Persons}$	2.005	2.003	1.802			
R^2_{within}	0.059	0.059	0.059			
$R^2_{between}$	0.002	0.054	0.240			
$R^2_{overall}$	0.054	0.054	0.225	0.021	0.081	0.285
Wald χ^2/F	15.54***	25.00**	49.97***	1.43	1.72	3.00

Notes. Model 1a, Model 2a, Model 3a = Random intercept estimation, non-standardized beta coefficients, robust standard errors, *z*-values in parentheses; Model 1b, Model 2b, Model 3b = OLS estimation, non-standardized beta coefficients, robust standard errors, *t*-values in parentheses; Model 1a, 1b without covariates; Model 2a, 2b covariates: age, sex, religion, origin; Model 3a, 3b covariates: age, sex, religion, origin, attitudes toward foreigners and Muslims. *$p < .05$, **$p < .01$, ***$p < .001$.

Table A3. Estimation results for the field experiment

	Support (Behavior)			Support (Behavior)		
	Model 1a	Model 2a	Model 3a	Model 1b	Model 2b	Model 3b
Recipient's name (foreign, ref. native)	−0.064 (−0.80)	−0.065 (−0.80)	−0.048 (−0.58)	−0.064 (−0.79)	−0.066 (−0.76)	−0.045 (−0.50)
Amount of scholarship (high, ref. low)	0.193* (2.41)	0.190* (2.33)	0.205* (2.44)	0.191* (2.23)	0.190* (2.23)	0.205* (2.30)
N	140	140	140	140	140	140
(Pseudo) R^2	0.033	0.046	0.061	0.044	0.060	0.077
LR χ^2/F	6.23*	8.58	11.27	3.26*	1.01	1.20

Notes. Model 1a, Model 2a, Model 3a = Logit estimation, average marginal effects, robust standard errors, *z*-values in parentheses; Model 1b, Model 2b, Model 3b = OLS estimation, non-standardized beta coefficients, robust standard errors, *t*-values in parentheses; Model 1a, 1b without covariates; Model 2a, 2b covariates: age, sex, religion, origin; Model 3a, 3b covariates: age, sex, religion, origin attitudes toward foreigners and Muslims. *$p < .05$.

Table A4. Estimation results of joint models

	Survey experiment and field experiment	Survey experiment and field experiment (First vignette only)
	Support (Int. Split) and Support (Behavior)	Support (Int. Split) and Support (Behavior)
Recipient's name (Arab, ref. German)	−0.045 (0.087)	−0.045 (0.089)
Amount of scholarship (high, ref. low)	0.204* (0.086)	0.204* (0.088)
Survey Experiment × Arab Name	0.045 (0.087)	0.003 (0.120)
Survey Experiment × High Scholarship	−0.126 (0.089)	−0.039 (0.119)
$N_{Vignettes}$	559	140
$N_{Persons\ survey\ experiment}$	140	140
$N_{Persons\ field\ experiment}$	140	140
$\sigma_{Persons}$	0.419	
R^2_{within}	0.054	
$R^2_{between}$	0.223	
$R^2_{overall}$	0.231	0.220
Wald χ^2/F	190.54***	6.75***

Notes. Joint model = Random intercept estimation, non-standardized beta coefficients, robust standard errors, *z*-values in parentheses; Joint model (First vignette) = OLS estimation, non-standardized beta coefficients, robust standard errors, *t*-values in parentheses; Covariates in both models: attitudes toward foreigners and Muslims; Conditional effects for field experiment in both models. *$p < .05$, ***$p < .001$.

Table A5. Estimations for the survey experiment and for the field experiment (detailed treatment conditions)

	Support (Intention)	Support (Behavior)
	Survey experiment	Field experiment
Tarik El Morabet/€100 (reference)	–	–
Rashid Yassir/€897	0.321** (3.21)	0.180 (1.43)
Tobias Müller/€100	0.007 (0.09)	−0.069 (−0.57)
Christoph Winter/€897	0.302** (3.25)	0.157 (1.24)
$N_{Vignettes}$	559	
$N_{Persons}$	140	140
$\sigma_{Persons}$	1.802	
R^2_{within}	0.059	
$R^2_{between}$	0.240	
$R^2_{overall}$	0.225	0.077
Wald χ^2/F	54.51***	1.17

Notes. Model survey experiment = Random intercept estimation, non-standardized beta coefficients, robust standard errors, *z*-values in parentheses; Model field experiment = OLS estimation, non-standardized b coefficients, robust standard errors, *t*-values in parentheses; Covariates in both models: age, sex, religion, origin, attitudes towards foreigners and Muslims. $^{**}p < .01$, $^{***}p < .001$.

Table A6. Estimation results of joint models (detailed treatment conditions)

	Survey experiment and field experiment	Survey experiment and field experiment (First vignette only)
	Support (Int. Split) & Support (Behavior)	Support (Int. Split) & Support (Behavior)
Tarik El Morabet/€100 (reference)	–	–
Rashid Yassir/€897	0.180 (1.48)	0.180 (1.43)
Tobias Müller/€100	−0.069 (−0.59)	−0.069 (−0.57)
Christoph Winter/€897	−0.157 (1.28)	−0.157 (1.24)
Survey experiment × Tarik El Morabet/€100 (difference in reference)	−0.355 (−0.73)	−0.257 (−0.51)
Survey experiment × Rashid Yassir/€897	−0.094 (−0.75)	−0.044 (−0.26)
Survey experiment × Tobias Müller/€100	0.076 (0.64)	0.001 (0.01)
Survey experiment × Christoph Winter/€897	−0.077 (−0.62)	−0.033 (−0.22)
$N_{Vignettes}$	559	140
$N_{Persons\ survey\ experiment}$	140	140
$N_{Persons\ field\ experiment}$	140	140
$\sigma_{Persons}$	0.419	
R^2_{within}	0.055	
$R^2_{between}$	0.223	
$R^2_{overall}$	0.231	0.220
Wald χ^2/F	196.12***	6.43***

Notes. Joint model = Random intercept estimation, non-standardized beta coefficients, robust standard errors, *z*-values in parentheses; Joint model (First vignette) = OLS estimation, non-standardized beta coefficients, robust standard errors, *t*-values in parentheses; Covariates in both models: attitudes toward foreigners and Muslims; Conditional effects for field experiment in both models. $^{***}p < .001$.

