Validating vignette and conjoint survey experiments against real-world behavior

Jens Hainmueller[a], Dominik Hangartner[b,c], and Teppei Yamamoto[d,1]

[a]Department of Political Science and Graduate School of Business, Stanford University, Stanford, CA 94305-6044; [b]Department of Methodology, London School of Economics, London WC2A 2AE, United Kingdom; [c]Institute of Political Science, University of Zurich, 8050 Zurich, Switzerland; and [d]Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited* by David D. Laitin, Stanford University, Stanford, CA, and approved December 30, 2014 (received for review August 29, 2014)

Survey experiments, like vignette and conjoint analyses, are widely used in the social sciences to elicit stated preferences and study how humans make multidimensional choices. However, there is a paucity of research on the external validity of these methods that examines whether the determinants that explain hypothetical choices made by survey respondents match the determinants that explain what subjects actually do when making similar choices in real-world situations. This study compares results from conjoint and vignette analyses on which immigrant attributes generate support for naturalization with closely corresponding behavioral data from a natural experiment in Switzerland, where some municipalities used referendums to decide on the citizenship applications of foreign residents. Using a representative sample from the same population and the official descriptions of applicant characteristics that voters received before each referendum as a behavioral benchmark, we find that the effects of the applicant attributes estimated from the survey experiments perform remarkably well in recovering the effects of the same attributes in the behavioral benchmark. We also find important differences in the relative performances of the different designs. Overall, the paired conjoint design, where respondents evaluate two immigrants side by side, comes closest to the behavioral benchmark; on average, its estimates are within 2% percentage points of the effects in the behavioral benchmark.

stated preferences | survey methodology | public opinion | conjoint | vignette

Survey experiments, such as conjoint analysis (1, 2) and vignette factorial surveys (3, 4), are widely used in many areas of social science to elucidate how humans make multidimensional choices and evaluate objects (e.g., people, social situations, and products). Such stated preference experiments typically ask respondents to choose from or rate multiple hypothetical descriptions of objects (often called profiles or vignettes) that vary along different attributes that are presumed to be important determinants of the choice or rating. The values of the attributes are randomly varied across respondents and tasks, allowing the researcher to estimate the relative importance of each attribute for the resulting choice or rating.

Proponents of stated preference experiments often argue that these experimental designs are capable of narrowing or even closing the gap between the survey and the real world, because they mimic real decision tasks (5–7). Viewed from this perspective, survey experiments provide an effective, low-cost, and widely applicable tool to study human preferences and decision-making. However, critics argue that such experiments fundamentally lack external validity and do not accurately capture real-world decision-making. It is known that survey self-reports are prone to various sources of response bias, such as hypothetical bias, social desirability bias, acquiescence bias, satisficing, and other cognitive biases that might seriously undermine the validity of survey experimental measures (8, 9). These biases can lead respondents to behave quite differently when they make choices in survey experiments compared with similar choices in the real world. After all, talk is cheap, and hypothetical choices carry no real costs or consequences—so why would respondents take the decision task seriously or be able to correctly predict how they would approach the task in the real world (10, 11)? Viewed from this perspective, stated preference experiments only allow for inferences about what respondents say that they would do but not about what they would actually do.

Despite the fundamental importance of external validity for the accumulation of knowledge about human behavior in the social sciences, there has been surprisingly little effort to examine how well stated preference experiments capture real-world decisions. In fact, to the best of our knowledge, our study is the first to externally validate two of the most commonly used designs for stated preference experiments—vignette and conjoint analyses—in a social science context. By external validation, we mean a comparison that investigates how well the estimated effects of the profile attributes on the hypothetical choice in the survey experiment recover the true effects of the same profile attributes in a behavioral benchmark, where humans make similar choices under real-world conditions. Our validation analysis, therefore, does not aim at the question of pure measurement, another important dimension of external validity in survey research that has been extensively examined (12, 13). We, instead, focus on the external validity of the estimated causal effects and examine whether the inferences that one would draw from a survey experiment about the relative importance of the attributes for explaining stated choices match the revealed relative importance of these attributes for similar actual choices. [We are not aware of any study that externally validates vignette analysis against a behavioral benchmark. For conjoint analysis, there have been only a few attempts at external validation in marketing and transportation (14, 15), but these studies typically only

Significance

Little evidence exists on whether preferences about hypothetical choices measured in a survey experiment are driven by the same structural determinants of the actual choices made in the real world. This study answers this question using a natural experiment as a behavioral benchmark. Comparing the results from conjoint and vignette experiments on which attributes of hypothetical immigrants generate support for naturalization with the outcomes of closely corresponding referendums in Switzerland, we find that the effects estimated from the surveys match the effects of the same attributes in the behavioral benchmark remarkably well. We also find that seemingly subtle differences in survey designs can produce significant differences in performance. Overall, the paired conjoint design performs the best.

Author contributions: J.H., D.H., and T.Y. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: teppei@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1416587112/-/DCSupplemental.

SOCIAL SCIENCES

compare whether market shares of products estimated from stated preference data predict actual market shares. However, because the benchmarks are limited to aggregate market shares and do not include data on behavioral choices, they cannot compare the effects of the attributes to see if the reasons that explain the hypothetical choices are the same as the reasons that explain the behavioral choices.] Investigating the external validity of the causal effects is of crucial importance given that the causal effects are typically the key quantity of interest in survey experiments.

In particular, we ask (*i*) whether any survey experimental design comes close to the behavioral benchmark and (*ii*) if there is important variation in the relative performance of the various designs. Included in our horserace are the most commonly used survey experimental designs, including vignettes with single and paired profiles, conjoints with single and paired profiles, and a paired conjoint design with forced choice.

Our external validation test takes advantage of a unique behavioral benchmark provided by data from a natural experiment in Switzerland, where some municipalities used referendums to vote on the naturalization applications of immigrants. In such referendums, voters received a voting leaflet with a short description of the applicant, including information about his or her attributes, such as age, sex, education, origin, language skills, and integration status. Voters then cast a secret ballot to accept or reject individual applicants one at a time, and applicants that received more yes than no votes received Swiss citizenship (16). *SI Appendix* provides details of the referendum process.

These data provide an ideal behavioral benchmark to evaluate stated preference experiments, because they closely resemble a real-world vignette experiment. Voters decided over thousands of immigrants with varying characteristics in a real-world setting, allowing us to causally identify how much each particular attribute affected the probability of being accepted or rejected by voters. These voting data yield an accurate measure of the revealed preferences of the voters given that the referendums used secret ballots and the stakes were significantly high (on naturalization, immigrants acquire the same rights as existing members of the local citizenry, including the right to vote and permanently stay in the country). Moreover, unlike many other real-world choice situations, in the referendums, the information environment and choice attributes are sufficiently constrained, such that they can be accurately mimicked in a survey experimental design. In other words, because we know which applicant's information voters had at their disposal when voting on the applicant's naturalization request, we can include precisely the same attributes in the behavioral benchmark regression and rule out omitted variable bias (i.e., the possibility that the decisions are driven by other unobserved factors that might have influenced the voting decision; ref. 16 has a discussion of this assumption). This absence of omitted variable bias is a key requirement for a valid benchmark that fails in many other real-world settings, where it is typically difficult to accurately assess the importance of the attributes for the resulting choice (for example, we might be able to observe whether voters elect a candidate or customers purchase a product, but in most instances, we cannot determine which attributes of the candidate or product influenced the choice, let alone by how much.).

There are at least two reasons why our study provides a particularly difficult test for showing the external validity of stated preference experiments. First, our comparison is out of sample, because the use of naturalization referendums ended in 2003, and our survey experiment was administered in 2014, which implies a gap of more than 10 y between the survey and behavioral data. Evidence from other survey data collected throughout this time period suggests that public attitudes toward immigration remained fairly stable over this time period in the municipalities under study (details in *SI Appendix*). However, the test is more difficult compared with a scenario where the data would be collected at the same point in time. Second, the naturalization of immigrants is a politically sensitive issue in Switzerland. In particular, right-wing parties have repeatedly mobilized against "mass naturalizations" of immigrants with campaign posters that portray the hands of foreigners snatching Swiss passports. It, therefore, raises the specter of potentially strong social desirability bias (17) if, for example, respondents in the survey pretend that they would not discriminate against immigrants from certain origins, such as Turkey and Yugoslavia, to seem politically correct to the researcher. In the actual naturalization referendums, where votes were cast with secret ballots, we, indeed, see a strong origin-based discrimination against such applicants (16).

Experimental Design and Data

Just as in the real-world referendums, in our experiment, respondents are presented with profiles of immigrants and then asked to decide on their application for naturalization. The immigrant profiles vary on seven attributes, including sex, country of origin, age, years since arrival in Switzerland, education, language skills, and integration status. Each attribute can take on various values, which are randomly chosen to form the immigrant profiles. *SI Appendix* provides a full list of attribute values. This list of attributes closely matches the list of attributes that voters saw on the voting leaflets distributed for the referendums. The attributes are presented in the same order as on the original leaflets.

Each respondent is randomly assigned to one of five different designs and asked to complete 10 choice tasks, which are presented on separate screens (details of the designs are in *SI Appendix*). The first design is a single-profile vignette design, where a single immigrant profile is presented in the form of a short paragraph that describes the applicant with the attributes listed in the text, and then, respondents are asked to accept or reject the applicant. This design is close to the format of the actual voting leaflets used in the referendums, where voters also received short text descriptions of each applicant and voted on each applicant one at a time. Vignettes with single profiles are also perhaps the most widely used factorial survey design in the social sciences (4).

The second design is a paired profiles vignette, which is similar to the single-profile vignette, except that two immigrant vignettes are presented one below the other, and then, respondents are asked to accept or reject each of the two applicants. The idea in this condition is that respondents are implicitly encouraged to compare the two applicants, and this encouragement to compare might increase survey engagement.

The third design is a single-profile conjoint, where one immigrant profile is presented in a table that resembles a curriculum vitae with two columns. The first column lists the names of the attributes, and the second column lists the attribute values. Respondents are again asked to accept or reject the applicant. This conjoint design is dissimilar to the format of the voting leaflets, but its potential advantage is that the applicant information is more accessible to respondents in a tabular form compared with the text descriptions used in the vignettes and the leaflets.

The fourth design is a paired profiles conjoint, which is similar to the single-profile conjoint, except that two immigrant profiles are presented next to each other in the conjoint table. Respondents are asked to accept or reject each of the two applicants. The potential advantage of this design is that it makes it easy for respondents to compare the two applicants on each attribute. The paired design is widely used for conjoint analysis in marketing (18).

The fifth design is equivalent to the paired profiles conjoint, except that respondents are forced to choose which of the two immigrant profiles they prefer for naturalization. The forced choice design is popular, because it might encourage respondents to more carefully consider the information about the profiles and increase their engagement with the task. However, this design is perhaps furthest away from the actual referendums, which did not entail a forced choice and therefore, did not constrain the unconditional probability of accepting an applicant to exactly one-half.

Our data consist of a sample of 1,979 Swiss citizens who we randomly sampled from the voting age population of the municipalities that used naturalization referendums before 2004. We recruited respondents by telephone using interviewers from a survey company. Respondents subsequently completed our survey online. Our sample is, therefore, a probability sample of

the target population, and our respondents are not routine survey-takers, in contrast to some survey experimental studies that rely on respondents recruited from opt-in internet panels (19).

SI Appendix contains details of the survey sample. The survey sample closely matches the demographic composition of the voter population in the municipalities as measured by the Swiss postreferendum study (the best available data on the Swiss voting population), including the margins for age, sex, political interest, political participation, education, and employment. To match as closely as possible the target population of voters that participated in naturalization referendums before 2004, we restricted the analysis to those voters who report in our survey that they participated in naturalization referendums and are 30 y of age or older. Note that, of those 30 y old and older, about 34% report that they did vote in naturalization referendums, which closely approximates the typical turnout for the naturalization referendums before 2004. We also correct for any small remaining imbalances using entropy-balancing weights (20) that adjust the sample data to exactly fit the respective demographic margins measured in the Swiss postreferendum study. Results are very similar without this reweighting (*SI Appendix*).

After the completion of our main experiment, we also conducted a similar experiment on a sample of Swiss students as well as staff of a large public university in Zurich. The participants were recruited through an email sent out to all students and employees. The only major difference between our main and student experiments is that the latter only included the paired profiles conjoint design with forced choice. A primary purpose of the student experiment was to examine whether the results in the main experiment could also be replicated on a separate sample representing a very different population of Swiss respondents.

Results

We assess the results of our experiments from two different perspectives. First, do the survey results and behavioral benchmark match qualitatively (i.e., are the overall conclusions about the relative importance of the attributes similar in both the survey and behavioral data?). Second, we examine whether the survey results and behavioral benchmark match quantitatively (i.e., how close do the attribute effects match in the survey and behavioral data?).

Fig. 1, column 1 (enclosed in a gold box) shows the effects of the applicant attributes on the rejection probability in the behavioral benchmark. The plot shows the point estimates and their 95% confidence intervals from a linear model fitted by ordinary least squares, where we regress the rejection rate on sets of dummy variables for the applicant attributes. We omit one level for each attribute that serves as the reference category (shown with the dots without confidence intervals). The regression estimates are also shown in *SI Appendix*, Table S3. In the behavioral data, the applicant's country of origin has by far the greatest effect on the rejection probability. In particular, applicants from Turkey and Yugoslavia (we use the term Yugoslavia here as a shorthand for applicants from Bosnia and Herzegovina, Croatia, and the former Yugoslavia.) are about 15–19 percentage points more likely to be rejected compared with observably similar applicants from The Netherlands (the reference category). In contrast, applicants from other European countries are no more likely to be rejected than applicants from The Netherlands, with the possible exception of German applicants, who are slightly more likely (3 percentage points; $P \approx 0.26$) to be rejected. A key question for the benchmarking is, thus, whether the survey results can replicate the massive penalty for Turkish and Yugoslavian applicants that constitutes the most dominant feature driving the rejection of applicants. The origin attribute is also the one that presumably carries the strongest social desirability connotations given that origin-based discrimination is prohibited by the antidiscrimination clause in the Swiss constitution (16).

Apart from origin, we also see that applicants with high levels of education are about 3 percentage points less likely to be rejected compared with observably similar applicants with low levels of education. Natives also slightly prefer immigrants that are so well-integrated that they are essentially indistinguishable from a Swiss native compared with those familiar with Swiss traditions. However, these effects are much smaller in magnitude than the origin effects. The findings also suggest that effects for sex, age, and years of arrival are close to zero and generally statistically insignificant at conventional levels.

How close do the stated preference experiments capture the patterns in the behavioral benchmark? Fig. 1, columns 2–7 shows the estimated effects in each survey experimental condition. Strikingly, although there is some important variation in the relative performance of the different designs, overall, the stated preference experiments match the behavioral benchmark rather well, with the important exception of the student sample.

The paired conjoint design (Fig. 1, column 2) comes the closest overall. It almost exactly reproduces the magnitude of the origin penalty for applicants from Turkey and Yugoslavia and also replicates the slight penalty for German applicants fairly closely. Moreover, the estimates are also remarkably close to the benchmark for the applicant's sex, age, and education. The only

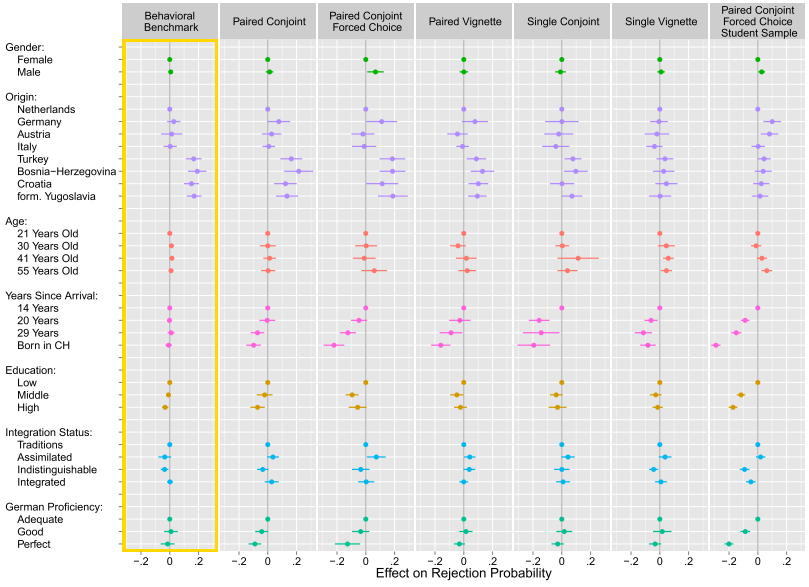

Fig. 1. Effects of applicant attributes on opposition to naturalization request: behavioral benchmark vs. stated preference experiments. The figure shows point estimates (dots) and corresponding cluster-robust 95% confidence intervals (horizontal lines) from ordinary least squares regressions. The dots on the zero line without confidence intervals denote the reference category for each applicant attribute. CH, Switzerland.

systematic differences are that natives are less likely to reject applicants born in Switzerland or in the country for 29 y (compared with 14 y) as well as applicants that have perfect (as opposed to adequate) German proficiency. Applicants assimilated into Switzerland (as opposed to familiar with Swiss traditions) also receive a small penalty compared with the benchmark. However, even for these attributes, the estimates do not deviate very strongly. Overall, the paired conjoint design captures the general patterns of the behavioral benchmark remarkably well. As in the benchmark, a massive origin penalty for Turkish and Yugoslavian applicants emerges as a clear conclusion, whereas the other attributes are generally found to play minor roles.

The other designs also perform rather well for our main survey sample. The paired conjoint design with forced choice (Fig. 1, column 3) captures the massive origin disadvantage for Turkish and Yugoslavian applicants very well, although it slightly overestimates the penalty for German applicants. It also matches well on most other applicant characteristics, except the substantial overestimation of the bonus for longer residency (21 percentage points for applicants born in Switzerland). The discrepancies that are found in the paired conjoint without forced choice (penalty for being assimilated and bonus for perfect German proficiency) are also present and somewhat amplified under the forced choice design. Overall, however, the results still match the patterns in the behavioral benchmark well, with the strengths of origin effects emerging as a clear central feature. This performance is remarkable given that this design is the one that is conceptually most different from the actual referendums.

The paired vignette design (Fig. 1, column 4) performs similarly to the preceding two designs. It captures the massive origin disadvantage for Turkish and Yugoslavian applicants, although the estimates are somewhat smaller and differ from the behavioral benchmark by 5–8 percentage points. It also matches well on all other applicant characteristics, except the years since arrival, where it overestimates and suggests a positive effect for longer residency. The size of this overestimation, however, is smaller than in the forced choice paired conjoint design (15 percentage points). Overall, the match is again quite good, although the strong origin effects perhaps come out less clearly as the dominant finding than in the preceding two designs.

The single-profile conditions, both conjoints (Fig. 1, column 5) and vignettes (Fig. 1, column 6), also perform fairly well overall, with the signs of estimated effects mostly agreeing with the behavioral benchmark when they are substantively different from zero. However, both designs vastly underestimate the penalty for applicants from Turkey and Yugoslavia. In fact, according to the single-conjoint design, Croatian applicants are just as likely to be rejected as observably identical applicants from The Netherlands, Germany, and Austria. This underestimation of the origin penalty is even stronger in the single-vignette design, where none of the origin effects are statistically distinguishable from zero at conventional levels. This finding is astonishing, because not only is the single vignette perhaps the most widely used design in the social sciences but also, the format of the leaflets used in the actual referendums most closely resembled the single vignettes.

Finally, the results from our follow-up experiment on the student sample (Fig. 1, column 7) provide an important lesson for survey experimental research. Despite the fact that the design used was identical to the forced choice paired conjoint design, the estimated effects of the attributes are far from the behavioral benchmark or any of the results on our main sample. In the student sample, German and Austrian applicants are estimated to receive a sizable penalty compared with Dutch applicants (10 and 8 percentage points, respectively), whereas applicants from Turkey or Yugoslavia receive no such penalty. Moreover, other attributes, such as years since arrival, education, and German proficiency, are estimated to have much larger effects on the probability of rejection than in the benchmark. The poor performance of our student experiment suggests that it is essential to match the characteristics of a survey sample to the target population as closely as possible for the survey experiment to generate externally valid conclusions about real-world behavior. This finding contrasts with other work that has found that results from survey experiments on convenience samples, like Amazon.com's Mechanical Turk, replicate results from survey experiments on representative probability samples (19). Our comparison is between survey experiments and real-world behavior.

Now we turn to a more systematic, quantitative assessment of our designs. Table 1 reports various performance measures for each design. Table 1, columns 1–3 display the mean, median, and maximum of the absolute differences from the behavioral benchmarks across the 21 attribute effects (the estimated differences are shown in *SI Appendix*, Fig. S10). On these metrics, the paired conjoint design is again the clear top performer. The mean and median differences from the benchmarks are only 2 and 1 percentage points, respectively, and the maximum difference is only 9 percentage points. The paired vignette emerges as the close second, with mean and median deviations of 3 and 2 percentage points, respectively, and a maximum difference of 15 percentage points. The other three designs for our main survey sample—paired conjoint with forced choice, single conjoint, and vignette—perform worse than the top two designs. Finally, the forced choice paired conjoint on the student sample is clearly the worst performer, missing the benchmark by no less than 28 percentage points at its worst.

Table 1, column 4 shows the total number of differences from the benchmark estimates that individually are statistically significantly different from zero at the 0.05 level for each design. Table 1, column 5 presents the same metric but with Bonferroni correction for multiplicity. On this criterion, the paired conjoint and vignette designs tie for first place, where only 4 of 21 differences are statistically distinguishable from zero without multiplicity correction and just 1 of 21 differences is statistically distinguishable from zero with correction. The paired conjoint design with forced choice and the single-conjoint design come next and perform similarly. Remarkably, the single-vignette design turns out to be the worst performer among the designs tested on our main sample. Again, the student sample performs by far the worst, with as many as two-thirds of 21 estimated effects significantly different from the benchmark values.

Table 1, column 6 presents an F statistic for the hypothesis test against the joint null of no difference between the effects in the behavioral benchmark and each survey design. Again, the paired conjoint design is the top performer, with a relatively small F value [$F(21, 1791) \approx 2.55$]. The paired vignette, single-conjoint, and vignette designs perform worse but not by large margins. Interestingly, the paired conjoint design with forced choice is the clear worst performer among our main designs on this test. This subpar performance is largely because of the one big mistake that it makes in overestimating the residency effect, to which the F statistic is sensitive by design. Finally, the student sample again performs terribly on this metric, with the F value more than 10 times as large as in the paired conjoint design.

Table 1, columns 8 and 9 shows metrics that are designed to capture the relative predictive performance. Here, we first obtain the predicted rejection probabilities for all actual applicant profiles in the behavioral data for each survey design by multiplying their observed attribute levels by the estimated regression coefficients for the design ($\hat{Y}$). We then calculate the bivariate correlation between the observed shares of rejection votes and the predicted rejection probabilities. Finally, we calculate the correlation between the observed and fitted rejection vote shares in the behavioral data as the benchmark. Thus, the question we ask is how well can the attribute effects estimated in the survey experiments generate inferences about the relative likelihood of rejection between the observed applicants compared with the actual attributes?

Table 1, column 7 presents the correlation coefficients calculated by the above procedure along with the correlation in the behavioral benchmark, and Table 1, column 8 directly compares the predicted rejection probabilities in the survey ($\hat{Y}_s$) against the fitted rejection rates in the behavioral regression ($\hat{Y}_b$) by calculating the correlation between the two. The results again reveal the remarkable performance of the paired conjoint design.

Table 1. Differences in effects of applicant attributes: survey vs. behavioral estimates

Design	Absolute differences			Significant differences		Joint F test	Cor$(Y,\hat{Y})$	Cor$(\hat{Y}_b,\hat{Y}_s)$
	Mean	Median	Maximum	Raw	Adjusted			
Paired conjoint	0.02	0.01	0.09	4/21	1/21	2.55	0.44	0.75
Paired conjoint, forced choice	0.04	0.02	0.21	6/21	3/21	10.33	0.34	0.58
Paired vignette	0.03	0.02	0.15	4/21	1/21	3.52	0.29	0.49
Single conjoint	0.05	0.03	0.19	7/21	2/21	3.91	0.29	0.49
Single vignette	0.04	0.03	0.17	9/21	4/21	3.64	0.26	0.44
Paired conjoint, forced choice (students)	0.07	0.06	0.28	14/21	11/21	26.69	0.13	0.23
Behavioral							0.58	

This table reports performance measures for each survey design. Columns 1–3 display the mean, median, and maximum of the absolute differences from the behavioral benchmark across the 21 attribute effects. Column 4 shows the total number of differences from the benchmark estimates that are statistically different from zero at the 5% significance level. Column 5 presents the same metric but with Bonferroni correction for multiple comparisons. Column 6 presents an F statistic for the hypothesis test against the joint null of no difference between the effects in the behavioral benchmark and each survey design. Column 7 presents the correlation between observed shares of rejection votes and the predicted rejection probabilities based on the survey estimates. Column 8 presents the correlation between the predicted rejection probabilities based on the survey estimates and the fitted rejection rates in the behavioral regression.

Although the predicted rejection rates in the behavioral data themselves are correlated with observed rejection rates at about 0.58, this correlation only drops to 0.44 when we use the attribute effects estimated in the paired conjoint experiment instead of the estimates directly based on the actual attributes of the applicants. This prediction translates into a correlation as large as 0.75 between the behavioral and survey-based predicted values for the paired conjoint design. Based on these correlations, the paired conjoint design with forced choice comes out in second place and clearly is above the rest of the designs. The paired vignette and single conjoint tie for third place. The single vignette performs worse than any of the other designs tested on our main representative sample. Finally, predictions from the student sample perform poorly, with correlations of only 0.13 and 0.23 with the observed rejection rates and behavioral predictions, respectively.

Although our focus for the external validation is on the match between the estimated causal effects of the attributes in the survey experiments and the behavioral benchmark, it is worth pointing out that the survey experiments generally do rather poorly in predicting the absolute levels of rejection rates observed in the actual referendums. The paired conjoint design, for example, predicts about 21% of the actual applicants to be rejected citizenship. In contrast, the observed rejection rate in the actual referendums turns out to be 37%. This difference is no smaller in any of the survey designs that we tested (details in *SI Appendix*). (Ironically, the two forced choice paired conjoint conditions—the designs that fix the unconditional rejection rate at exactly 50% by construction—come closest in terms of estimating the average behavioral rejection probability.) This finding is not so surprising given the mixed evidence on the reliability of survey-based preference measures. Indeed, past studies have found that surveys often fail to accurately measure the absolute levels of preferences for certain types of objects and behavior. For example, it is well-known in the literature on the contingent valuation method (21) that willingness to pay for public goods is often highly unreliable as a measure of the actual amount of dollars that respondents would pay in the real world. Likewise, public opinion surveys are consistently found to overpredict the actual level of voter turnout in national elections (22), although they tend to perform well for predicting certain other types of aggregate-level behavioral outcomes [e.g., election results (23)]. What is remarkable in our validation results, then, is the finding that some of the tested survey designs perform exceedingly well in recovering the structural effects of individual attributes, despite failing to match the absolute levels of support.

Discussion on Mechanisms

Why do some survey designs perform significantly better than others in reproducing real-world attribute effects? Specifically, why do paired designs produce more accurate estimates than single-profile designs? Although our study was not designed to draw definitive conclusions about causal mechanisms, the available evidence strongly suggests respondent engagement as a key mechanism. That is, it is likely that respondents in the paired conditions were more engaged in the survey and therefore less prone to questionnaire satisficing.

Less motivated respondents have a tendency to look for cues to provide reasonable answers that are easy to select with little thought to avoid the cognitive work required for optimal question answering (9). Such satisficing behavior manifests itself in nondifferentiation (giving the same answer to a battery of similar questions) and acquiescence response bias (the tendency to agree, regardless of the question content) (24). In our context, a satisficer might simply accept all applicant profiles that he or she is asked to evaluate, regardless of the applicant characteristics. Fig. 2 plots the fraction of respondents who exhibit this response pattern in each design (excluding the forced choice designs, which require that one-half of the respondents are rejected). The paired conjoint shows the lowest level of satisficing, with 56% of respondents accepting all of their applicants, followed by the paired vignette with 63%. The level of satisficing is much higher in the single-profile designs, with 70% and 72% of respondents accepting all applicants in the single-conjoint and single-vignette conditions, respectively. Note that these differences are driven by a pure design effect, because both

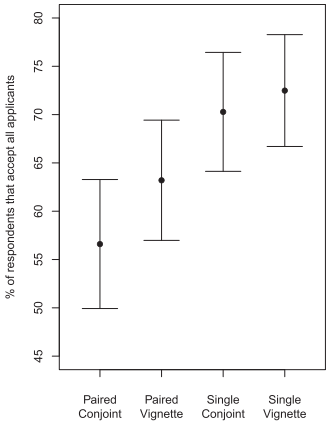

Fig. 2. Acquiescence and nondifferentiation in different survey designs. The figure shows the proportion of respondents who accept all applicants with corresponding 95% confidence intervals.

the applicant characteristics and the respondents are randomly assigned and therefore similar in expectation in all conditions. This finding is highly consistent with the idea that the paired designs induced a higher motivation to seriously engage with the decision tasks and evaluate information about the profiles more carefully compared with the single-profile designs.

Our conjecture that respondent engagement plays a key role in explaining design effects is further bolstered by some of the patterns that we observe in the estimated attribute effects. Note that the effects of countries of origin—the main real structural effects of immigrant attributes as identified in the behavioral benchmark—are largest in magnitude in Fig. 1, column 1 and then become smaller almost monotonically as we move to less well-performing designs in Fig. 1. Indeed, the sizes of these effects decrease almost exactly in proportion to the rate of satisficing reported in Fig. 2. Because nondifferentially accepting all applicants will mechanically shrink the effect of any attribute toward zero, this finding suggests that better-performing designs are able to recover the structural attribute effects more accurately by increasing the overall level of survey engagement and thus decreasing the amount of noise caused by respondents who are merely satisficing.

Finally, the data on actual and perceived response times provide yet another piece of evidence that respondents were more engaged in the paired conditions. Although respondents in the paired conditions spent about 60% more time on the tasks to decide on the applicants than respondents in the single-profile conditions, these groups show no differences when asked about dissatisfaction with the length and difficulty of the survey. Details on this finding are reported in *SI Appendix*.

Conclusion

Taking advantage of a unique behavioral benchmark of voting in secret ballot naturalization referendums in Switzerland, our study provides an external validation test of vignette and conjoint analyses that compares whether the relative importance of attributes for explaining the hypothetical choices in survey experiments matches the relative importance of the same attributes for actual choices in the real world.

Our main finding is that the stated preference experiments, which simulated the naturalization referendums in the survey, perform remarkably well in capturing the structural effects of attributes that drive voting behavior in the actual referendums. In particular, the paired conjoint design comes closest to the behavioral benchmark. It precisely recovers the qualitative pattern of the actual naturalization referendums, with its dominant effects of origin, and it also performs best according to various quantitative measures of performance based on absolute distances and correlations. The superior performance of the paired conjoint is quite striking given that this design is fairly dissimilar to the format of the leaflets that were used in the actual referendums. Relatedly, we find that the paired designs, in general, outperform the single-profile designs, and the evidence suggests that the paired designs induce more engagement and less satisficing among respondents. The single-vignette design, although the most similar to the format of the actual referendums, performs rather poorly compared with the other designs. This finding is important because, of the methods we tested, the single-vignette design is probably the most widely used method in the social sciences. Finally, although the paired conjoint forced choice design performs fairly well when administered in our main survey to a probability sample of the target population, the design fared poorly when replicated on a convenience sample of students.

Taken together, our findings suggest, to maximize external validity about real-world causal effects, that survey samples need to be carefully chosen to match the target population and that survey experimental designs need to be carefully crafted to motivate respondents to seriously engage with hypothetical choice tasks to mimic the incentives that they face when making the same choices in the real world. The results indicate that merely matching the appearance of decision tasks is insufficient; the effect of better survey engagement seems to eclipse the impact of superficial similarity in questionnaires. Our result also reinforces the importance of targeting the right population in sampling survey respondents.

How generalizable are the results from our external validation test? There are some worries. Even the best performing stated preference experiments fail to accurately predict the absolute levels of preference for accepting applicants for naturalization, a finding consistent with the past evidence on the difficulty of survey measurement. Furthermore, it is important to emphasize that stated preference experiments might exhibit lower external validity in other contexts. However, given that we test a hot-button issue that is likely to invoke some social desirability bias, and that there was a ten-year gap between the behavioral and the survey data, our results make us cautiously confident in the external validity of the stated preference experiments. Thus, our test is a useful step in assessing the validity of survey techniques to measure real-world behavior, and in showing the conditions under which we should have confidence in survey results.

ACKNOWLEDGMENTS. We thank Ryan Bakker and participants of the 31st Society for Political Methodology Annual Summer Meeting for their comments and suggestions and Stefan Schütz and Joël Wicki for excellent research assistance.

1. Green PE, Rao VR (1971) Conjoint measurement for quantifying judgmental data. *J Mark Res* 8(3):355–363.
2. Hainmueller J, Hopkins DJ, Yamamoto T (2014) Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Polit Anal* 22(1):1–30.
3. Rossi PH, Anderson AB (1982) *The Factorial Survey Approach: An Introduction* (Sage Publications, Beverly Hills, CA), pp 15–67.
4. Wallander L (2009) 25 years of factorial surveys in sociology: A review. *Soc Sci Res* 38(3):505–520.
5. Alexander CS, Becker HJ (1978) The use of vignettes in survey research. *Public Opin Q* 42(1):93–104.
6. Arnold HJ, Feldman DC (1981) Social desirability response bias in self-report choice situations. *Acad Manage J* 24(2):377–385.
7. Louviere JJ, Hensher DA, Swait JD (2000) *Stated Choice Methods: Analysis and Applications* (Cambridge Univ Press, Cambridge, United Kingdom).
8. Schwarz N (1999) Self-reports: How the questions shape the answers. *Am Psychol* 54(2):93–105.
9. Krosnick JA, Judd CM, Wittenbrink B (2014) *The Handbook of Attitudes*, eds Albarracin D, Johnson BT, Zanna MP (Psychology Press, New York), pp 21–76.
10. Bertrand M, Mullainathan S (2001) Do people mean what they say? implications for subjective survey data. *Am Econ Rev* 91(2):67–72.
11. Neill HR, Cummings RG, Ganderton PT, Harrison GW, McGuckin T (1994) Hypothetical surveys and real economic commitments. *Land Econ* 70(2):145–154.
12. Bernstein R, Chadha A, Montjoy R (2001) Overreporting voting: Why it happens and why it matters. *Public Opin Q* 65(1):22–44.
13. Benítez-Silva H, Buchinsky M, Man Chan H, Cheidvasser S, Rust J (2004) How large is the bias in self-reported disability? *J Appl Econ* 19(6):649–670.
14. Louviere JJ (1988) Conjoint analysis modelling of stated preferences: A review of theory, methods, recent developments and external validity. *J Transp Econ Policy* 22(1):93–119.
15. Natter M, Feurstein M (2002) Real world performance of choice-based conjoint models. *Eur J Oper Res* 137(2):448–458.
16. Hainmueller J, Hangartner D (2013) Who gets a swiss passport? a natural experiment in immigrant discrimination. *Am Polit Sci Rev* 107(1):159–187.
17. Tourangeau R, Yan T (2007) Sensitive questions in surveys. *Psychol Bull* 133(5):859–883.
18. Green PE, Krieger AM, Wind Y (2001) Thirty years of conjoint analysis: Reflections and prospects. *Interfaces* 31(3 Suppl):S56–S73.
19. Berinsky AJ, Huber GA, Lenz GS (2012) Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Polit Anal* 20(1):25–46.
20. Hainmueller J (2011) Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit Anal* 20(1):25–46.
21. Diamond PA, Hausman JA (1994) Contingent valuation: Is some number better than no number? *J Econ Perspect* 8(4):45–64.
22. Anderson BA, Silver BD (1986) Measurement and mismeasurement of the validity of the self-reported vote. *Am J Pol Sci* 30(4):771–785.
23. Gelman A, King G (1993) Why are american presidential election campaign polls so variable when votes are so predictable? *Br J Polit Sci* 23(4):409–451.
24. Holbrook AL, Green MC, Krosnick JA (2003) Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opin Q* 67(1):79–125.