Validating vignette and conjoint survey experiments against real-world behavior

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1073/pnas.1416587112">http://dx.doi.org/10.1073/pnas.1416587112</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>National Academy of Sciences (U.S.)</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Thu Feb 07 00:35:50 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/98031">http://hdl.handle.net/1721.1/98031</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Validating vignette and conjoint survey experiments against real-world behavior

Jens Hainmueller, Domink Hangartner, and Teppie Yamamoto

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.

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.
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 behavioral 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).

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 the referendum 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 (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 years 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, dissimilar Turkey and Yugoslavia, the voting leaflets, but its potential 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 details of the 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 close 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 the attribute effects match in both the survey and behavioral benchmark. Moreover, the estimates are also remarkably close to the behavioral benchmark for the applicant’s country of origin. 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 more 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
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 4) turns out to be substantially better, both in terms of 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 (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 benchmark well, with the estimated individual 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 benchmarks.

The paired vignette design (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 they are substantively different from zero. However, both designs very much underestimate the penalty for applicants born in Turkey and Yugoslavia. In fact, the size of this origin disadvantage, for Turkey, 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 3) and vignettes (Fig. 1, column 2), 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 very much underestimate the penalty for applicants born in Turkey and Yugoslavia. In fact, the size of this origin disadvantage, for Turkey, is 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. It shows significantly different from the behavioral benchmark estimates that individually are statistically significant from zero at the 0.05 level for each design.

Table 1, column 4 shows the total number of differences from the benchmarks. As our main sampling size is large, the maximum difference is only 9 percentage points. The paired vignette emerges as the close second, with mean and median overestimated differences 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 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 comes very close to performing significantly different from the benchmark. The single-vignette design turns out to be the clear worst performer among the designs tested on our main sample. It shows significantly different from the behavioral benchmark estimates that individually are statistically significant from zero at the 0.05 level for each design. As our main sampling size is large, the maximum difference is only 9 percentage points. The paired vignette emerges as the close second, with mean and median overestimated differences 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, 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 $(\hat{Y}_D)$. 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 $(\hat{Y}_D)$ against the fitted rejection rates in the behavioral regression $(Y)$ by calculating the correlation between the two. The results again reveal the remarkable performance of the paired conjoint design.
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 vignette 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...
The main real structural effects are... 

Interfaces

(Cambridge Univ Press, Cambridge, United Kingdom).

105.

119.

The Handbook of Attitudes

46.

385.

22(1):1

We thank Ryan Bakker and participants of the 31st

30.

72.

520.

Acad Manage J

Hainmueller et al.

104.

Polit

In particular, the paired conjoint design comes closest to the attributes that drive voting behavior in the actual referendums.

perform remarkably well in capturing the structural effects of 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 Joëll Wicki for excellent research assistance.


