

# Validating vignette and conjoint survey experiments against real-world behavior

Jens Hainmueller<sup>a</sup>, Dominik Hangartner<sup>b,c</sup>, and Teppei Yamamoto<sup>d,1</sup>

<sup>a</sup>Department of Political Science and Graduate School of Business, Stanford University, Stanford, CA 94305-6044; <sup>b</sup>Department of Methodology, London School of Economics, London WC2A 2AE, United Kingdom; <sup>c</sup>Institute of Political Science, University of Zurich, 8050 Zurich, Switzerland; and <sup>d</sup>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.

<sup>1</sup>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](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1416587112/-DCSupplemental).

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 were 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?).

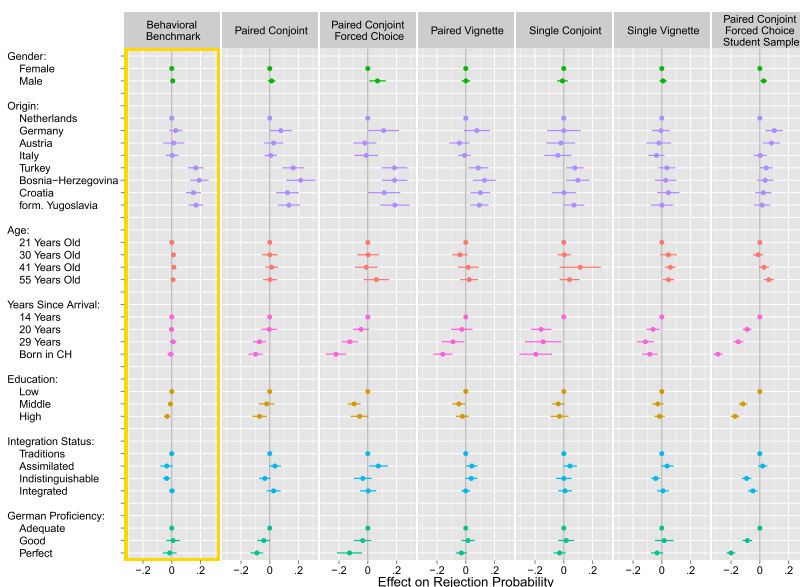
Fig. 1, column 1 (enclosed in a gold box) shows the effects of the applicant attributes on the rejection probability in the behavioral benchmark. The plot shows the point estimates and their 95% confidence intervals from a linear model fitted by ordinary least squares, where we regress the rejection 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 <i>F</i> 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.

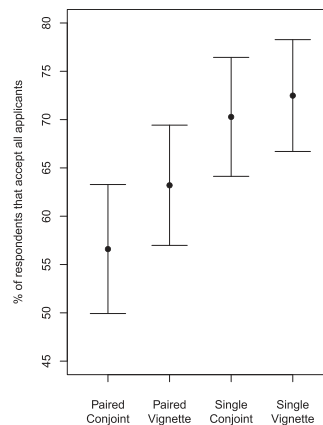
Although our focus for the external validation is on the match between the estimated causal effects of the attributes in the survey experiments and the behavioral benchmark, it is worth pointing out that the survey experiments generally do rather poorly in predicting the absolute levels of rejection 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.

- Green PE, Rao VR (1971) Conjoint measurement for quantifying judgmental data. *J Mark Res* 8(3):355–363.
- 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.
- Rossi PH, Anderson AB (1982) *The Factorial Survey Approach: An Introduction* (Sage Publications, Beverly Hills, CA), pp 15–67.
- Wallander L (2009) 25 years of factorial surveys in sociology: A review. *Soc Sci Res* 38(3):505–520.
- Alexander CS, Becker HJ (1978) The use of vignettes in survey research. *Public Opin Q* 42(1):93–104.
- Arnold HJ, Feldman DC (1981) Social desirability response bias in self-report choice situations. *Acad Manage J* 24(2):377–385.
- Louviere JJ, Hensher DA, Swait JD (2000) *Stated Choice Methods: Analysis and Applications* (Cambridge Univ Press, Cambridge, United Kingdom).
- Schwarz N (1999) Self-reports: How the questions shape the answers. *Am Psychol* 54(2):93–105.
- Krosnick JA, Judd CM, Wittenbrink B (2014) *The Handbook of Attitudes*, eds Albarracín D, Johnson BT, Zanna MP (Psychology Press, New York), pp 21–76.
- Bertrand M, Mullainathan S (2001) Do people mean what they say? implications for subjective survey data. *Am Econ Rev* 91(2):67–72.
- Neill HR, Cummings RG, Ganderton PT, Harrison GW, McGuckin T (1994) Hypothetical surveys and real economic commitments. *Land Econ* 70(2):145–154.
- Bernstein R, Chadha A, Montjoy R (2001) Overreporting voting: Why it happens and why it matters. *Public Opin Q* 65(1):22–44.
- 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.
- 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.
- Natter M, Feurstein M (2002) Real world performance of choice-based conjoint models. *Eur J Oper Res* 137(2):448–458.
- Hainmueller J, Hangartner D (2013) Who gets a swiss passport? a natural experiment in immigrant discrimination. *Am Polit Sci Rev* 107(1):159–187.
- Tourangeau R, Yan T (2007) Sensitive questions in surveys. *Psychol Bull* 133(5):859–883.
- Green PE, Krieger AM, Wind Y (2001) Thirty years of conjoint analysis: Reflections and prospects. *Interfaces* 31(3 Suppl):556–573.
- 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.
- 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.
- Diamond PA, Hausman JA (1994) Contingent valuation: Is some number better than no number? *J Econ Perspect* 8(4):45–64.
- Anderson BA, Silver BD (1986) Measurement and mismeasurement of the validity of the self-reported vote. *Am J Pol Sci* 30(4):771–785.
- 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.
- 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.

# SUPPORTING INFORMATION

“Do Survey Experiments Capture Real-World Behavior? External Validation of Conjoint and Vignette Analyses with a Natural Experiment”

Jens Hainmueller\*, Dominik Hangartner<sup>† ‡</sup>, Teppei Yamamoto<sup>§</sup>

*\*Department of Political Science and Graduate School of Business, Stanford University*

*†Department of Methodology, London School of Economics*

*‡Institute of Political Science, University of Zurich*

*§Department of Political Science, Massachusetts Institute of Technology*

## Contents

<b>S0 Introduction</b>	<b>S2</b>
<b>S1 Behavioral Benchmark: Naturalization Referendums</b>	<b>S2</b>
<b>S2 Stability of Immigration Preferences</b>	<b>S3</b>
<b>S3 Survey Design and Sample</b>	<b>S5</b>
<b>S4 Experimental Design</b>	<b>S7</b>
<b>S5 Additional Results</b>	<b>S13</b>
<b>S6 Survey Engagement</b>	<b>S19</b>

## S0 Introduction

This Supporting Information is structured as follows: In the first section we provide more background information about the naturalization referendums. The second section presents evidence suggesting that immigration-related preferences remained fairly stable from the time when the use of naturalization referendums ended and the time when we fielded our survey. The third section provides details about the survey sample. The fourth section provides details about the experimental design. The fifth section reports additional results and robustness checks for the main analysis. The last section reports additional results about the survey engagement in the different experimental designs.

## S1 Behavioral Benchmark: Naturalization Referendums

In Switzerland, each municipality autonomously decides on the ordinary naturalization applications of its foreign residents who seek Swiss citizenship (for more details on the Swiss naturalization procedure, see [1]). We focus on the group of municipalities that until 2003 used referendums with closed ballots to decide on naturalization requests. A typical naturalization referendum involved two stages. Local voters first received in the mail the ballot and an official voting leaflet that explained the pending naturalization request with a detailed description of each immigrant applicant including information about his or her age, gender, education, origin, language skills and or integration status. Figure S1 shows an anonymized example of a typical voting leaflet. Figure S2 provides an English translation. Voters then cast a secret ballot on each individual request and applicants with a majority of “yes” votes were granted Swiss citizenship.

Figure S1: Sample leaflet sent out to voters (names blacked out)

### Traktandum 2

Beschlussfassung über das Bürgerrechtsgesuch des italienischen Staatsangehörigen  
[redacted] geb. 30. November 1962 in Schwyz, wohnhaft 6440 Brunnen,  
[redacted].

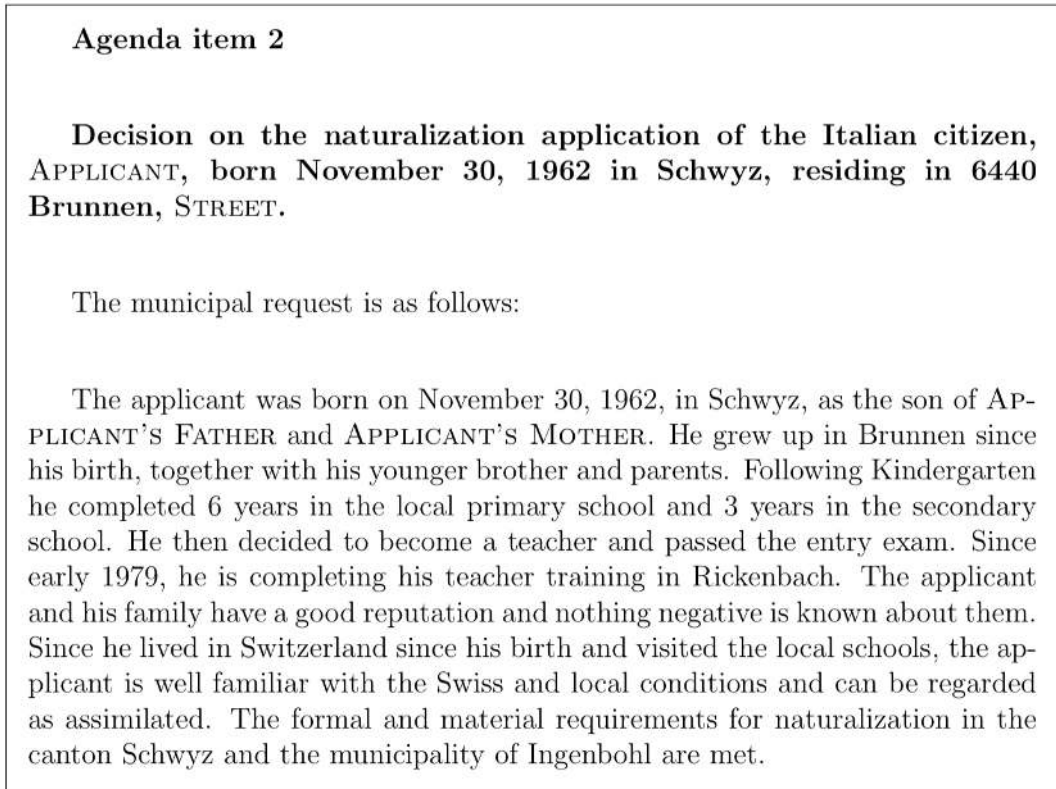
Der gemeinderätliche Antrag lautet:

Der Gesuchsteller wurde am 30. November 1962 in Schwyz geboren, als Sohn des [redacted] und der [redacted]. In Brunnen, wo er seit der Geburt zusammen mit einem jüngeren Bruder bei den Eltern aufwächst, besuchte er nach dem Kindergarten 6 Jahre die Primar- und 3 Jahre die Sekundarschule. Nachdem er sich für den Lehrerberuf entschlossen und die Aufnahmeprüfung bestanden hat, absolviert er seit dem Frühjahr 1979 das Lehrerseminar in Rickenbach. Der Gesuchsteller und seine Angehörigen geniessen einen unbescholtenen Leumund und über sie ist nichts Nachteiliges bekannt. Durch den immerwährenden Aufenthalt in der Schweiz und dem hiesigen Schulbesuch ist der Gesuchsteller mit den schweizerischen und örtlichen Verhältnissen vollends vertraut und kann als assimiliert bezeichnet werden. Die formellen und materiellen Voraussetzungen für die Einbürgerung im Kanton Schwyz und in der Gemeinde Ingenbohl sind gegeben.

32



Figure S2: Sample leaflet sent out to voters (English translation)



We use a subset of the data compiled by Hainmueller and Hangartner ([1]) that contains applicant characteristics and voting outcomes for 1,503 recorded naturalization referendums held between 1970 and 2003 in the 44 Swiss municipalities that used secret ballot referendums with voting leaflets.<sup>1</sup> The majority of the data consists of naturalization referendums held between 2000 and 2003. The behavioral data is recoded to match the survey attributes discussed below. We use these data to examine how applicant characteristics affect the outcome of naturalization referendums and thereby form the behavioral benchmark that we try to replicate with different survey experimental designs.

## S2 Stability of Immigration Preferences

The use of naturalization referendums ended in 2003, whereas our survey was administered in 2014. We use two different data sets to examine if voters' preferences regarding immigration might have changed between these years.

First, we use annual panel data from the Swiss Household Panel (SHP)<sup>2</sup>, to track changes in attitudes towards immigrants. The only immigration-related question in the SHP reads as follows: "Are you in favour of Switzerland offering foreigners the same opportunities as those

---

<sup>1</sup>The 44 municipalities are: Altdorf, Altendorf, Arth, Beckenried, Bühler, Buochs, Chur, Dallenwil, Davos, Einsiedeln, Emmen, Ennetmoos, Feusisberg, Freienbach, Gais, Galgenen, Gersau, Heiden, Hergiswil, Ingenbohl, Küsnacht, Lachen, Malers, Morschach, Oberiberg, Reichenburg, Rothenthurm, Schübelbach, Schwyz, Speicher, St. Margrethen, Stans, Stansstad, Steinen, Teufen, Trogen, Tuggen, Unteriberg, Urnäsch, Walzenhausen, Wangen, Weggis, Wolfenschiessen, and Wollerau.

<sup>2</sup>The data is hosted at <http://forscenter.ch/fr/our-surveys/swiss-household-panel/>.

offered to Swiss citizens, or in favour of Switzerland offering Swiss citizens better opportunities?”. Answers were recorded on a three point scale as (1) foreigners and Swiss citizens should be offered *equal opportunities*, (0) *neither* or (-1) Swiss citizens should be offered *better opportunities*. We use the subset of, on average,  $N = 1,395$  respondents per wave that reside in cantons that contain at least one target municipality. Figure S3 presents the SHP results. The trends over the years 1999 – 2009<sup>3</sup> are remarkably stable.

Figure S3: Stability of attitudes towards immigrants over time; Swiss Household Panel

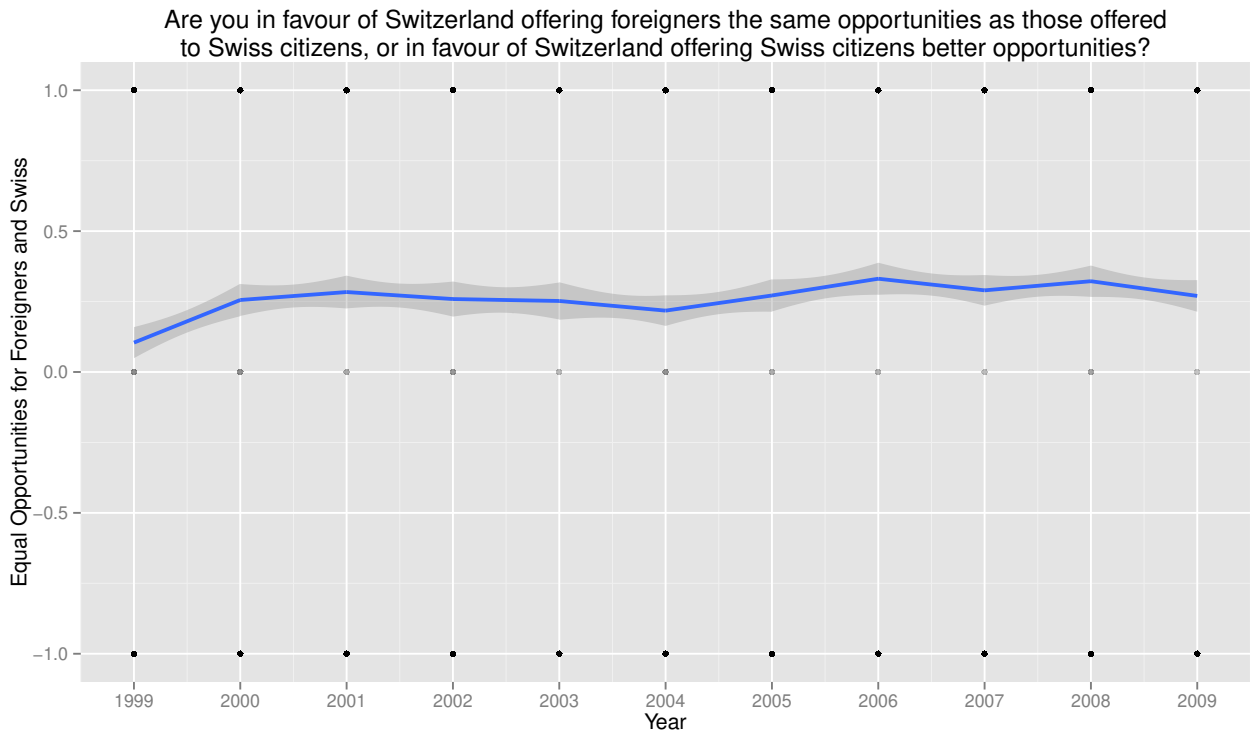


Figure shows year-to-year averages and corresponding 95 % confidence intervals. The variable measures attitudes towards immigrants over the years 1999–2009. Data: Swiss Household Panel, focusing on respondents from cantons that contain at least one target municipality. The sample size consists of, on average,  $N = 1395$  respondents per year.

Second, we use the VOX survey<sup>4</sup>, a cross-sectional post-referendum survey conducted about 3 – 6 times per year with a sample size of approximately 1,000 respondents per wave. The only immigration-related question that is repeatedly asked in the VOX survey is identical to the one from the SHP but coded slightly differently insofar as answers were recorded on a six point scale from (1) Swiss citizens should be offered *better opportunities*, to (6) foreigners and Swiss citizens should be offered *equal opportunities*. We use the subset of, on average,  $N = 104$  respondents per year that reside in one of the 44 target municipalities. Figure S4 presents the VOX results. While there is some year-to-year variance due to the small sample size, the overall trends over the years 1996 – 2013 are remarkably stable.

<sup>3</sup>Unfortunately, the question about opportunities for Swiss natives and foreigners was discontinued in 2010.

<sup>4</sup>The data is hosted at <http://forsdata.unil.ch/projects/voxit/sondages.asp?>

Figure S4: Stability of attitudes towards immigrants over time; VOX survey

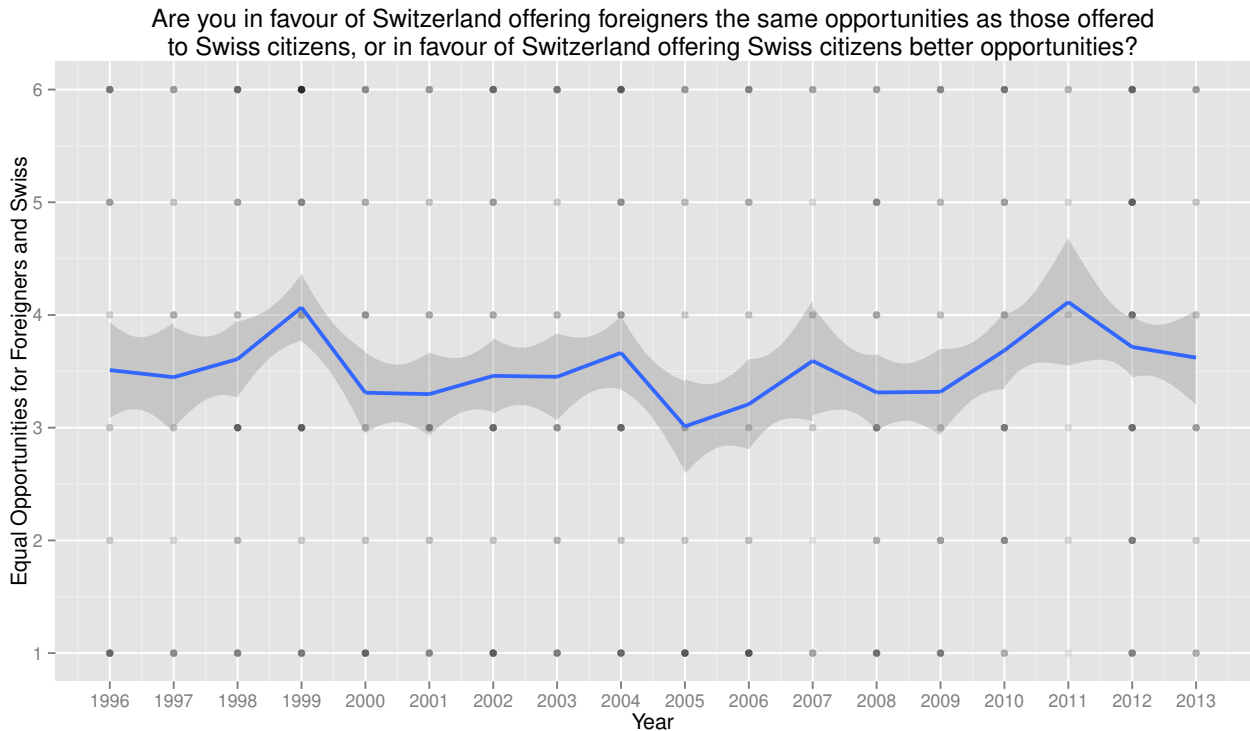


Figure shows year-to-year averages and corresponding 95 % confidence intervals. The variable measures attitudes towards immigrants over the years 1996–2013. Data: Swiss post-referendum survey VOX, focusing on respondents from target municipality. The sample size consists of, on average,  $N = 104$  respondents per year.

In summary, both tests provide some suggestive evidence that attitudes towards immigrants have remained fairly stable over the ten years separating the behavioral data from our survey experiment.

## S3 Survey Design and Sample

### Recruitment and Response Rate

Our main experiment was embedded in a survey that we conducted with the Swiss research firm gfs.bern. The field work took place between March 5 and July 25, 2014. The sampling design was a stratified random sampling. The recruitment was done by gfs.bern who contacted a stratified (by age and gender) random sample of 12,236 individuals in the target municipalities by telephone to invite them to participate in our online survey and collect baseline demographics and respondents' email addresses. Of these, 2,517 respondents agreed to participate in our online survey and were invited by email. Of those that expressed their willingness to participate,  $N = 1,979$  respondents completed the survey, yielding a retention rate of 78.6% from telephone interview to online survey.<sup>5</sup> Overall, this corresponds to a participation rate of 20.6 % and a cumulative response rate 3 (RR3) as defined by AAPOR of 12.8 %. Note that this RR3 is

<sup>5</sup>All respondents who initially agreed to participate in the online survey were reminded twice per email and a third time per telephone in the four weeks following the initial email invitation.

substantially higher than that of comparable online surveys. For example, a typical recent study conducted via Knowledge Networks, widely regarded as one of the best probability based online panels in the U.S., yields an RR3 of 2.8 % [2].

## Sample Descriptives

Table S1 shows the respondent characteristics for the unweighted survey sample, the Swiss post-referendum study VOX, and the reweighted survey sample. The VOX survey is the best available survey data on the Swiss voting population.

We see that the raw characteristics in our survey sample are quite close to the VOX survey. To address the small remaining differences we use entropy balancing [3] to reweight the survey sample based on the margins for age, gender, political interest, hypothetical participation in referendums, education, and employment to the margins computed from the VOX data. To create the margins for the reweighting procedure, we only focus on the VOX respondents that reside in the target municipalities between 2003–2013. After reweighting, the observable characteristics of the respondents in the two samples match very closely.

Table S1: Descriptive Statistics of Unweighted Survey, Target Sample Margins, and Weighted Survey

	Survey unweighted	2003–2013 VOX	Survey reweighted
Age	53.38	49.18	49.24
Female	0.50	0.53	0.53
Political Interest	3.31	2.87	2.88
Referendums	8.37	7.15	7.18
Education: 1	0.03	0.09	0.09
Education: 2	0.35	0.47	0.49
Education: 3	0.10	0.09	0.09
Education: 4	0.26	0.11	0.11
Education: 5	0.08	0.08	0.05
Education: 6	0.17	0.17	0.18
Employment	0.49	0.60	0.60

Table shows the descriptive statistics of the unweighted survey sample (Column 1), the VOX survey between 2003–2013 in the target municipalities that is our target sample (Column 2) and the reweighted survey sample (Column 3). Reweighting was performed using entropy balancing based on the following covariates: Age, Female (0/1), Political Interest (1–4), the number of referendums that respondents say they typically vote in assuming that there are 10 referendums per year (0–10), education (Education 1: compulsory schooling, Education 2: vocational training, Education 3: secondary schooling incl. *Matura*, Education 4: lower professional school, Education 5: higher professional school, Education 5: University degree) and Employment (0/1).

## Student Sample

In addition to the main survey, we also conducted a similar experiment on a sample of Swiss undergraduate and graduate students as well as administrative and faculty staff of the University of Zurich. The participants were recruited between July 11, 2014 and August 3, 2014 via an email invite sent out to all students and university employees. One-third of all respondents were randomly assigned to answer the paired profiles conjoint design with forced choice.  $N = 652$

respondents completed this survey and form the basis for the student sample. A primary purpose of this additional experiment was to examine whether the results in the main experiment could also be replicated on a separate sample representing a very different population.

## S4 Experimental Design

### Attributes and Attribute Levels

Table S2 details the attributes and attribute levels used to generate the profiles. The attribute levels are randomized under the following two constraints to rule out illogical combinations: age  $\geq$  years since arrival, and immigrants from Austria and Germany have a higher than “adequate” German language proficiency. The ordering of the attributes is fixed to match the typical leaflets as used in the actual naturalization referendums.

Table S2: Applicant Attributes and Attribute Levels

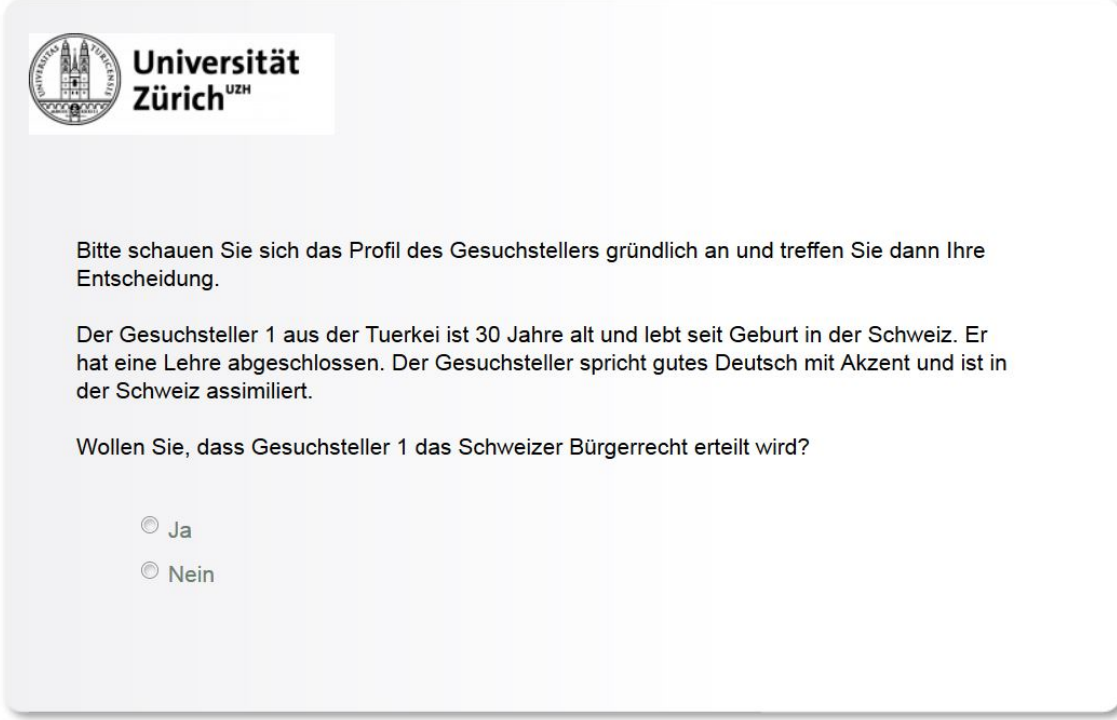
Attribute	Attribute Level
Gender $\in$	Male, Female
Origin $\in$	Netherlands, Germany, Austria, Italy, Turkey, Croatia, Former Yugoslavia, Bosnia and Herzegovina
Age $\in$	21 years, 30 years, 41 years, 55 years
Years since arrival $\in$	14 years, 20 years, 29 years, Born in CH
Education $\in$	Primary School, High School, University
German proficiency $\in$	“Adequate”, “Good with accent”, “Unaccented”, “Swiss German”
Integration status $\in$	“Assimilated”, “Integrated”, “Indistinguishable”, “Familiar with Swiss traditions”


### Treatment Conditions: Five Survey Designs

For the core of the experiment, we asked participants to decide on naturalization applicants of immigrants. We randomly allocated respondents to five groups of equal size and presented each group with one of five survey formats, namely the *single vignette*, *paired vignette*, *single conjoint*, *paired conjoint*, and the *paired conjoint with forced choice*. In the following we describe each design. Each respondent completed ten choice tasks.

Figure S5 shows a screenshot from the *single vignette*. The design presents a single immigrant profile 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.

Figure S5: Single Vignette



 **Universität  
Zürich** UZH

Bitte schauen Sie sich das Profil des Gesuchstellers gründlich an und treffen Sie dann Ihre Entscheidung.

Der Gesuchsteller 1 aus der Tuerkei ist 30 Jahre alt und lebt seit Geburt in der Schweiz. Er hat eine Lehre abgeschlossen. Der Gesuchsteller spricht gutes Deutsch mit Akzent und ist in der Schweiz assimiliert.

Wollen Sie, dass Gesuchsteller 1 das Schweizer Bürgerrecht erteilt wird?

Ja

Nein

Survey Powered By [Qualtrics](#) WEITER

Figure shows *single vignette* in German. Attributes levels for Gender, Origin, Age, Years since arrival, Education, German proficiency and integration status are randomized subject to logical constraints. Attribute order is fixed. Respondents are asked to vote “yes” or “no” on each applicant.

Figure S6 shows a screenshot from the *paired vignette*. This design is similar to the single vignette except that two immigrant vignettes are presented below each other and then respondents are asked to accept or reject each of the two applicants.

Figure S6: Paired Vignette

**Universität Zürich<sup>UZH</sup>**

Bitte schauen Sie sich die Profile der beiden Gesuchsteller gründlich an und treffen Sie dann Ihre Entscheidung.

Der Gesuchsteller 1 aus dem ehem. Jugoslawien ist 30 Jahre alt und lebt seit 29 Jahren in der Schweiz. Er hat die Universität abgeschlossen. Der Gesuchsteller spricht akzentfrei Schweizerdeutsch und ist vom Schweizer kaum zu unterscheiden.

Der Gesuchsteller 2 aus Oesterreich ist 30 Jahre alt und lebt seit 29 Jahren in der Schweiz. Er hat eine Lehre abgeschlossen. Der Gesuchsteller spricht akzentfrei Schweizerdeutsch und ist vom Schweizer kaum zu unterscheiden.

Wollen Sie, dass den Gesuchstellern das Schweizer Bürgerrecht erteilt wird?

	Ja	Nein
Gesuchsteller 1	<input type="radio"/>	<input type="radio"/>
Gesuchsteller 2	<input type="radio"/>	<input type="radio"/>

Survey Powered By [Qualtrics](#)

**WEITER**

Figure shows *paired vignette* in German. Attributes levels for Gender, Origin, Age, Years since arrival, Education, German proficiency and integration status are randomized subject to logical constraints. Attribute order is fixed. Respondents are asked to vote “yes” or “no” on each of the two applicants.

Figure S7 shows a screenshot from the *single conjoint*. This design presents a single immigrant profile in a conjoint table that resembles a CV with two columns. The first column lists the names of the attributes and the second column lists the attribute values. Again, respondents are asked to accept or reject the applicant.

Figure S7: Single conjoint

Universität Zürich

Bitte schauen Sie sich das Profil des Gesuchstellers gründlich an und treffen Sie dann Ihre Entscheidung.

	Gesuchsteller 1
<b>Geschlecht</b>	weiblich
<b>Herkunftsland</b>	Italien
<b>Alter</b>	55 Jahre
<b>In der Schweiz seit</b>	Geburt
<b>Bildungsstand</b>	obligatorische Schule
<b>Sprachkenntnisse</b>	spricht akzentfrei Schweizerdeutsch
<b>Integrationsstatus</b>	unterscheidet sich kaum vom Schweizer

Wollen Sie, dass Gesuchsteller 1 das Schweizer Bürgerrecht erteilt wird?

Ja  
 Nein

Survey Powered By [Qualtrics](#) WEITER

Figure shows *single conjoint* in German. Attributes levels for Gender, Origin, Age, Years since arrival, Education, German proficiency and integration status are randomized subject to logical constraints. Attribute order is fixed. Respondents are asked to vote “yes” or “no” on each applicant.



Figure S8 shows a screenshot from the *paired conjoint*. This design is similar to the single 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.

Figure S8: Paired conjoint

Universität Zürich UZH

Bitte schauen Sie sich die Profile der beiden Gesuchsteller gründlich an und treffen Sie dann Ihre Entscheidung.

	Gesuchsteller 1	Gesuchsteller 2
<b>Geschlecht</b>	weiblich	maennlich
<b>Herkunftsland</b>	Niederlande	Bosnien und Herzegowina
<b>Alter</b>	55 Jahre	55 Jahre
<b>In der Schweiz seit</b>	Geburt	14 Jahren
<b>Bildungsstand</b>	obligatorische Schule	obligatorische Schule
<b>Sprachkenntnisse</b>	spricht fließend Deutsch ohne Akzent	spricht akzentfrei Schweizerdeutsch
<b>Integrationsstatus</b>	unterscheidet sich kaum vom Schweizer	ist in der Schweiz integriert

Wollen Sie, dass den Gesuchstellern das Schweizer Bürgerrecht erteilt wird?

	Ja	Nein
Gesuchsteller 1	<input type="radio"/>	<input type="radio"/>
Gesuchsteller 2	<input type="radio"/>	<input type="radio"/>

Survey Powered By [Qualtrics](#) WEITER

Figure shows *paired conjoint* in German. Attributes levels for Gender, Origin, Age, Years since arrival, Education, German proficiency and integration status are randomized subject to logical constraints. Attribute order is fixed. Respondents are asked to vote “yes” or “no” on each of the two applicants.

Figure S9 shows a screenshot from the *paired conjoint with forced choice*. This design is identical to the paired conjoint except that respondents are asked to choose which of the two profiles they prefer for naturalization. In other words, respondents are forced to choose one of the two applicants and cannot accept or reject both.

Figure S9: Paired conjoint with forced choice

Bitte schauen Sie sich die Profile der beiden Gesuchsteller gründlich an und treffen Sie dann Ihre Entscheidung.

Welchen der beiden Gesuchsteller bevorzugen Sie für die Erteilung des Schweizer Bürgerrechts?

	<b>Gesuchsteller 1</b>	<b>Gesuchsteller 2</b>
<b>Geschlecht</b>	maennlich	maennlich
<b>Herkunftsland</b>	Niederlande	Italien
<b>Alter</b>	30 Jahre	41 Jahre
<b>In der Schweiz seit</b>	20 Jahren	Geburt
<b>Bildungsstand</b>	obligatorische Schule	obligatorische Schule
<b>Sprachkenntnisse</b>	spricht gutes Deutsch mit Akzent	kann sich auf Deutsch gut verstaendigen
<b>Integrationsstatus</b>	mit Schweizer Traditionen bestens vertraut	mit Schweizer Traditionen bestens vertraut
	<input type="radio"/>	<input type="radio"/>

Survey Powered By [Qualtrics](#) WEITER

Figure shows *paired conjoint with forced choice* in German. Attributes levels for Gender, Origin, Age, Years since arrival, Education, German proficiency and integration status are randomized subject to logical constraints. Attribute order is fixed. Respondents are forced to choose one of the two applicants.

## S5 Additional Results

This section reports additional analyses and robustness tests:

- Table S3 details the estimated effects of the applicant characteristics in actual and hypothetical naturalization referendums that are visualized in Figure 1 in the main text.
- Figure S10 shows the estimated differences in the effects of the applicant characteristics in the hypothetical and actual naturalization referendums. The estimates of the differences are generated based on a pooled dataset that combines the data from the hypothetical and actual naturalization referendums. In this pooled data we replicate the same model as in Table S3 and regress the rejection outcome on the attribute values, but also include a full set of indicator variables for the different survey experimental conditions plus the full set of interactions between these indicators and the attribute values. The coefficients on the interaction terms identify the differences in the estimates effects in hypothetical and actual naturalization referendums.
- Figure S11 and Table S4 replicate the main results based on the unweighted survey sample. The effects are very similar to the estimates based on the weighted sample displayed in Figure 1 and Table 1 in the main manuscript.
- Figure S12 and Table S5 replicate the main results but collapse the different country of origin indicators, following the coding of [1], into four roughly equal-sized categories: North West (Austria, Germany, Netherlands), South (Italy), Turkey, and Yugoslavia (Bosnia-Herzegovina, Croatia, and former Yugoslavia). Again, the results are very similar to the main results.
- Table S6 compares the estimated average rejection rate across the different survey designs to the behavioral benchmark. As discussed in the main text, most design underestimate the average rejection rate. The exception are the forced choice designs where the average rejection rate is by design fixed at 0.5.

Table S3: Attribute Effects in Actual and Hypothetical Naturalization Referendums

Condition	(1) Behavioral Benchmark	(2) Paired Conjoint	(3) Paired Conjoint Forced	(4) Paired Vignette	(5) Single Conjoint	(6) Single Vignette	(7) Paired Conjoint Forced Students
Gender:							
Male	0.0067 (0.0067)	0.013 (0.014)	0.067* (0.029)	0.00050 (0.015)	-0.0095 (0.019)	0.0088 (0.013)	0.027* (0.012)
Origin:							
Germany	0.028 (0.023)	0.077 (0.041)	0.11* (0.055)	0.077 (0.047)	0.00067 (0.059)	-0.0057 (0.031)	0.10** (0.031)
Austria	0.013 (0.038)	0.026 (0.035)	-0.020 (0.041)	-0.044 (0.036)	-0.021 (0.051)	-0.020 (0.044)	0.081** (0.031)
Italy	0.0030 (0.023)	0.0070 (0.022)	-0.011 (0.043)	-0.0085 (0.022)	-0.042 (0.048)	-0.037 (0.029)	0.0025 (0.024)
Turkey	0.17** (0.028)	0.16** (0.039)	0.19** (0.046)	0.087* (0.035)	0.077* (0.031)	0.036 (0.030)	0.044 (0.024)
Bosnia & Herzegovina	0.19** (0.033)	0.22** (0.052)	0.19** (0.046)	0.13** (0.042)	0.097* (0.042)	0.027 (0.038)	0.037 (0.030)
Croatia	0.15** (0.027)	0.12** (0.040)	0.11* (0.057)	0.10** (0.035)	0.0016 (0.043)	0.046 (0.039)	0.024 (0.029)
Yugoslavia	0.17** (0.026)	0.13** (0.039)	0.19** (0.053)	0.094** (0.033)	0.070 (0.037)	0.0015 (0.039)	0.015 (0.029)
Age:							
30 Years Old	0.012* (0.0057)	0.00017 (0.028)	0.0027 (0.039)	-0.040 (0.028)	0.0035 (0.024)	0.045 (0.030)	-0.012 (0.018)
41 Years Old	0.015* (0.0067)	0.013 (0.023)	-0.011 (0.040)	0.018 (0.037)	0.11 (0.073)	0.059** (0.019)	0.028 (0.019)
55 Years Old	0.0087 (0.0077)	0.0026 (0.025)	0.059 (0.045)	0.024 (0.031)	0.039 (0.035)	0.046* (0.020)	0.062** (0.019)
Years Since Arrival:							
20 Years	-0.0018 (0.0057)	-0.0034 (0.028)	-0.047 (0.029)	-0.028 (0.038)	-0.16** (0.037)	-0.061* (0.024)	-0.088** (0.016)
29 Years	0.0090 (0.012)	-0.071** (0.024)	-0.12** (0.029)	-0.089* (0.040)	-0.14* (0.065)	-0.11** (0.031)	-0.15** (0.018)
Born in Switzerland	-0.0074 (0.012)	-0.098** (0.026)	-0.22** (0.037)	-0.16** (0.034)	-0.19** (0.058)	-0.083** (0.028)	-0.29** (0.017)
Education:							
Middle	-0.0091 (0.0074)	-0.022 (0.028)	-0.095** (0.023)	-0.048* (0.024)	-0.039 (0.023)	-0.028 (0.021)	-0.12** (0.015)
High	-0.032** (0.012)	-0.071** (0.026)	-0.056 (0.032)	-0.023 (0.023)	-0.030 (0.032)	-0.015 (0.018)	-0.17** (0.016)
Integration Status:							
Assimilated	-0.035 (0.023)	0.036 (0.021)	0.073* (0.033)	0.042* (0.020)	0.043 (0.024)	0.037 (0.022)	0.020 (0.017)
Indistinguishable	-0.036* (0.014)	-0.035 (0.020)	-0.035 (0.031)	0.038 (0.021)	0.00032 (0.028)	-0.043** (0.016)	-0.092** (0.017)
Integrated	0.0016 (0.010)	0.027 (0.025)	0.0028 (0.029)	-0.00034 (0.016)	0.0090 (0.025)	0.0079 (0.021)	-0.048** (0.017)
German Proficiency:							
Good	0.0088 (0.025)	-0.042 (0.023)	-0.035 (0.031)	0.015 (0.023)	0.017 (0.028)	0.017 (0.033)	-0.087** (0.017)
Perfect	-0.015 (0.025)	-0.089** (0.022)	-0.13** (0.045)	-0.030 (0.020)	-0.029 (0.022)	-0.033 (0.022)	-0.20** (0.016)
Constant	0.37** (0.049)	0.24** (0.060)	0.57** (0.071)	0.21** (0.046)	0.23** (0.054)	0.15** (0.046)	0.82** (0.030)
Observations	1503	3910	3938	4274	2005	2173	6520

Ordinary least squares regression coefficients shown, with robust clustered standard errors in parentheses. Standard errors are clustered by the municipality (Model 1) or the respondents (Models 2-7) respectively. Model 1 is based on the actual naturalization referendums. Models 2-6 are based on our main survey and focus on the subsample of voters that is reweighted to match the margins of the Swiss post-referendum study VOX. Model 7 is based on the survey of the student sample. The reference categories for the various contrasts are: Gender: Female, Origin: Netherlands, Age: 21 Years, Years since Arrival: 14 Years, Education: Low, Integration Status: Traditions, German Proficiency: Adequate. Model 1 for the actual naturalization referenda also includes municipality and period fixed effects.

Figure S10: Differences in Effects of Applicant Attributes: Survey versus Behavioral Estimates

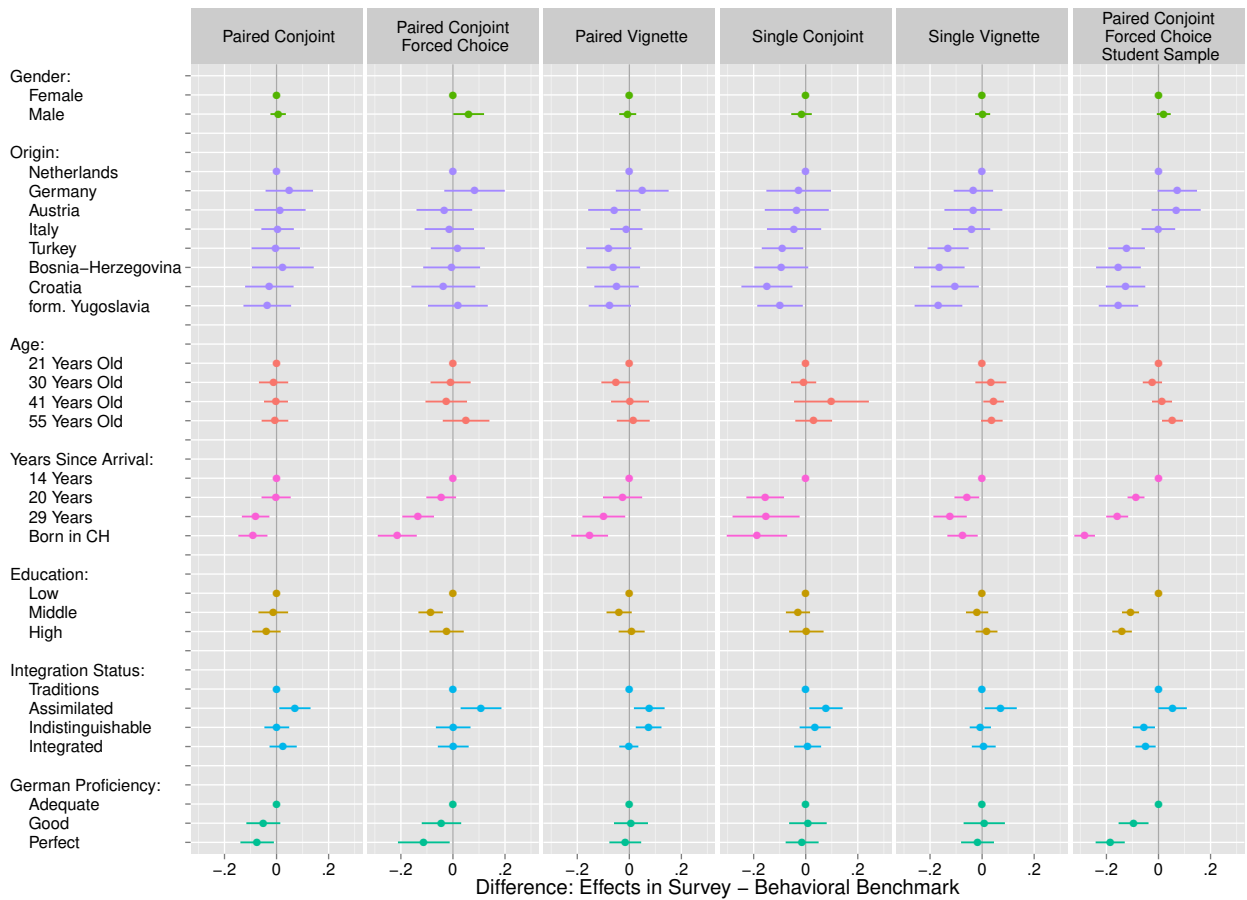


Figure shows point estimates (dots) and corresponding, cluster-robust 95 % confidence intervals (horizontal lines) from ordinary least squares regressions that identify the differences in the estimated effects in the survey conditions and the behavioral benchmark. The dots on the zero line without confidence intervals denote the reference category for each applicant attribute.

Figure S11: Effects of Applicant Attributes on Opposition to Naturalization Request (Un-weighted Survey Sample)

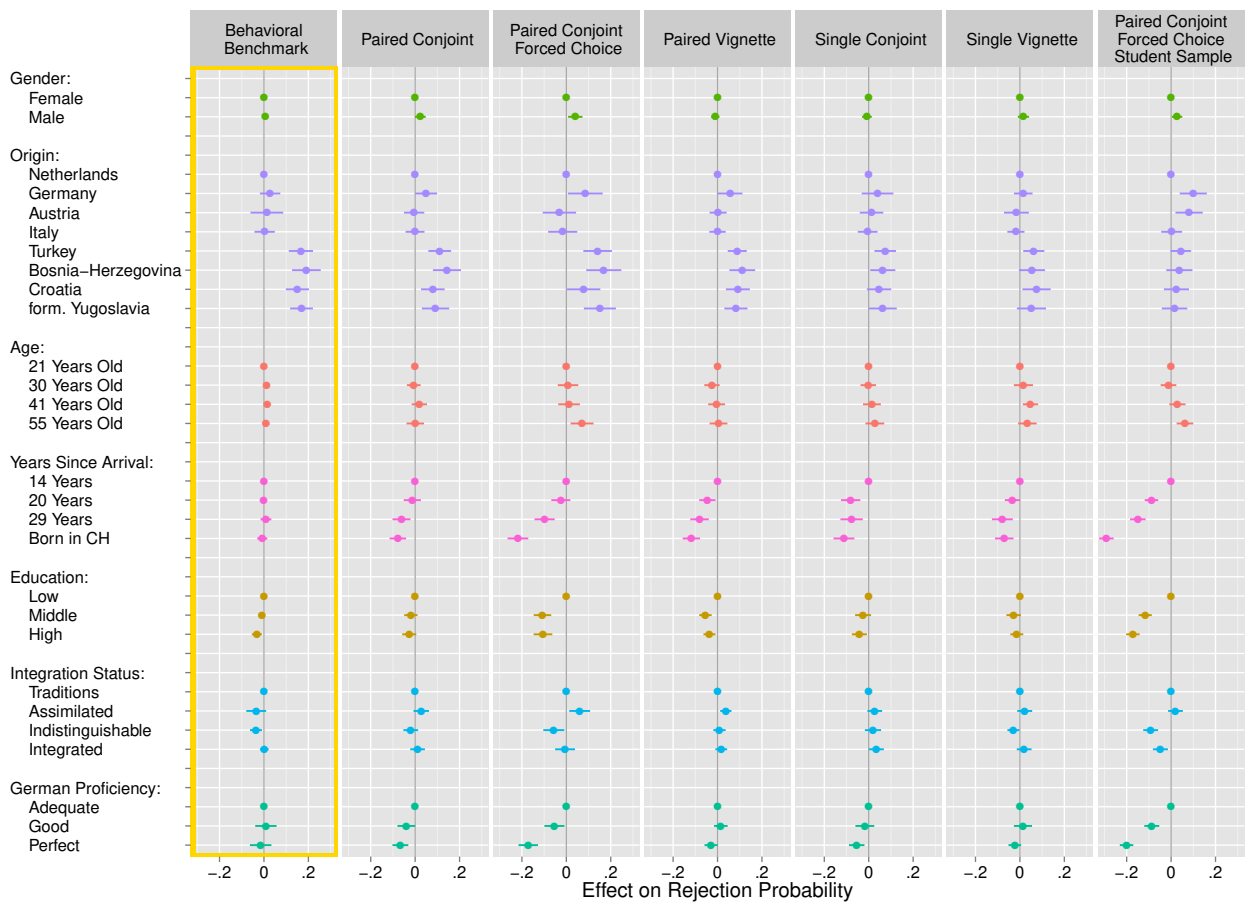


Figure shows point estimates (dots) and corresponding, cluster-robust 95 % confidence intervals (horizontal lines) from ordinary least squares regressions. The dots on the zero line without confidence intervals denote the reference category for each applicant attribute.

Table S4: Differences in Effects of Applicant Attributes: Survey versus Behavioral Estimates (Unweighted Survey Sample)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Design:	Absolute Differences			Sig. Diffs		Joint	Cor( $Y, \hat{Y}$ )	Cor( $\hat{Y}_b, \hat{Y}_s$ )
	mean	median	max	raw	adj	F-test		
Paired Conjoint	0.02	0.01	0.08	3/21	1/21	2.04	0.41	0.70
Paired Conjoint, FC	0.04	0.02	0.21	7/21	4/21	10.62	0.32	0.55
Paired Vignette	0.03	0.02	0.11	9/21	2/21	4.35	0.35	0.60
Single Conjoint	0.04	0.02	0.13	9/21	3/21	2.94	0.33	0.57
Single Vignette	0.03	0.02	0.14	6/21	2/21	2.82	0.35	0.60
Paired Conjoint, FC (Students)	0.07	0.06	0.28	14/21	11/21	26.69	0.13	0.23
Behavioral							0.58	

Table reports measures of performance for each survey design based on the unweighted sample of voters. Column 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 .05 significance level. Column 5 presents the same metric but with the Bonferroni correction. 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 bivariate correlation between observed shares of rejection votes and the predicted rejection probabilities. Column 8 presents the bivariate correlation between the predicted rejection probabilities based on the survey estimates and the fitted rejection rates in the behavioral regression. See main text for further details on the procedure used to generate columns 7 and 8.

Figure S12: Effects of Applicant Attributes on Opposition to Naturalization Request (Aggregated Origin Groups)

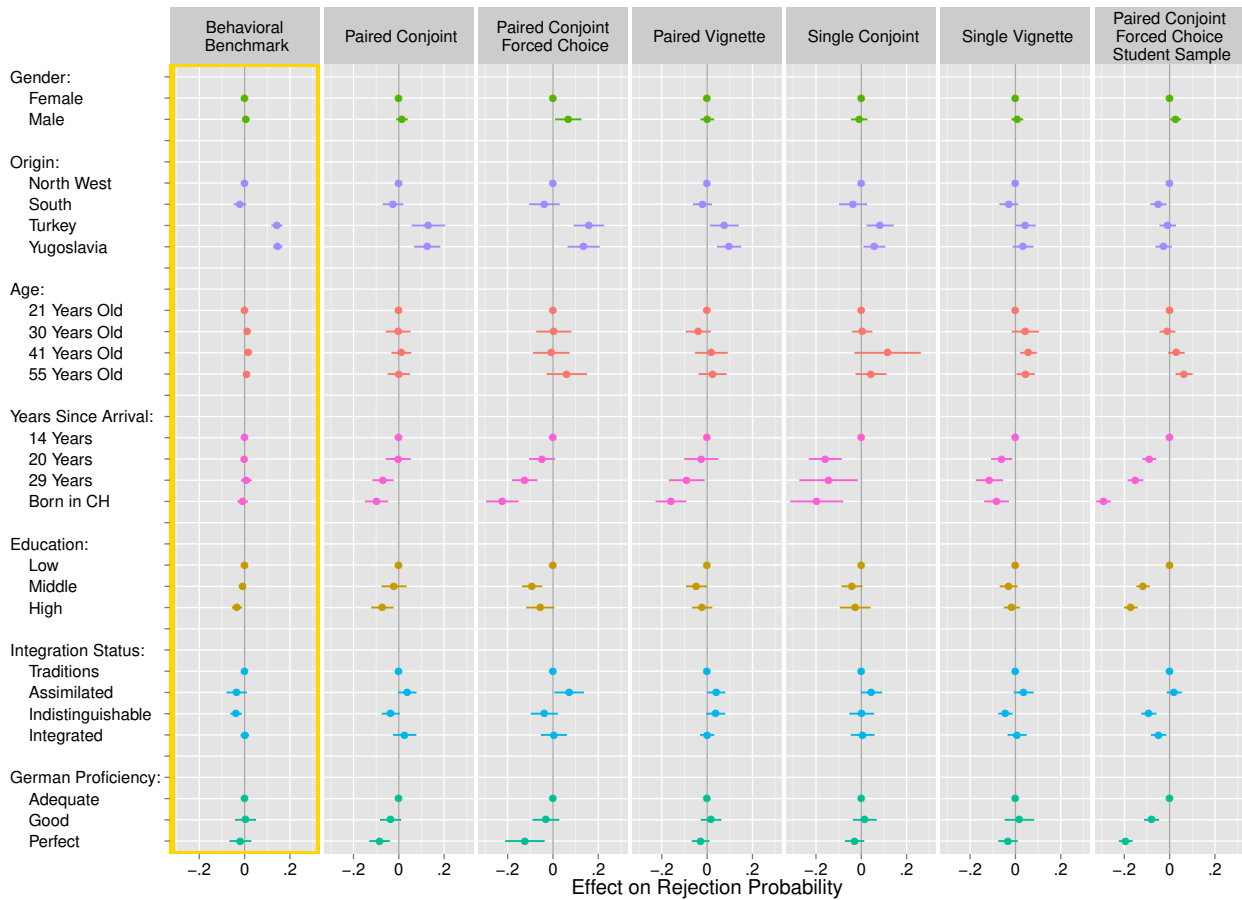


Figure shows point estimates (dots) and corresponding, cluster-robust 95 % confidence intervals (horizontal lines) from ordinary least squares regressions. The dots on the zero line without confidence intervals denote the reference category for each applicant attribute.

Table S5: Differences in Effects of Applicant Attributes: Survey versus Behavioral Estimates (Aggregated Origin Groups)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Design:	Absolute Differences			Sig. Diffs		Joint	Cor( $Y, \hat{Y}$ )	Cor( $\hat{Y}_b, \hat{Y}_s$ )
	mean	median	max	raw	adj	F-test		
Paired Conjoint	0.02	0.01	0.09	4/17	1/17	2.29	0.47	0.80
Paired Conjoint, FC	0.04	0.02	0.21	6/17	3/17	11.88	0.37	0.64
Paired Vignette	0.03	0.01	0.15	4/17	2/17	3.76	0.34	0.59
Single Conjoint	0.04	0.01	0.19	5/17	3/17	4.96	0.28	0.47
Single Vignette	0.03	0.01	0.12	7/17	3/17	3.94	0.28	0.49
Paired Conjoint, FC (Students)	0.07	0.05	0.28	12/17	8/17	31.15	0.17	0.30
Behavioral							0.58	

Table reports measures of performance for each survey design based on the weighted sample of voters and based on the aggregated origin groups. Column 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 .05 significance level. Column 5 presents the same metric but with the Bonferroni correction. 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 bivariate correlation between observed shares of rejection votes and the predicted rejection probabilities based on the survey estimates. Column 8 presents the bivariate correlation between the predicted rejection probabilities based on the survey estimates and the fitted rejection rates in the behavioral regression. See main text for further details on the procedure used to generate columns 7 and 8.

Table S6: Estimated Average Rejection Rate for the Applicants with Naturalization Referendums

	Estimated Average Rejection Rate
Behavioral Benchmark	.37
Paired Conjoint	.21
Paired Conjoint Forced	.49
Paired Vignette	.17
Single Conjoint	.12
Single Vignette	.10
Paired Conjoint Forced Students	.47

Table shows the estimated average rejection rate for the applicants with naturalization referendums. For the behavioral benchmark the rejection rate is simply the average proportion voting “no” in the referendum sample. For each survey condition we predict the rejection probability for the applicants in the referendum sample by taking their characteristics and multiplying them with the coefficients estimated from the survey respondents and then take the average of these predicted values. For observations with missing attribute information in the behavioral data, we impute missing values with their observed mean levels.



## S6 Survey Engagement

This section examines the differences in respondents' survey engagement across the different designs and thereby offers at least suggestive evidence for one particular causal pathway that runs through survey engagement, and explains why the paired designs produce better estimate of attribute effects than the single profile design.

Figure S13 shows that respondents in the paired and single profile conditions perceived no significant difference in the length of the survey, even though the actual response time was about 60% longer. Median response time used to complete the 10 decision tasks was 245 seconds for the *paired conjoint*, 291 seconds for the *paired conjoint with forced choice*, 253 seconds for the *paired vignette*, 166 seconds for the *single vignette*, and 153 seconds for the *single conjoint*.

Figure S14 shows that respondents perceived no significant difference in the difficulty of the survey, even though respondents in the paired profile conditions evaluated twice as many applicant profiles.

Figure S13: Perceived Survey Length Across Survey Designs

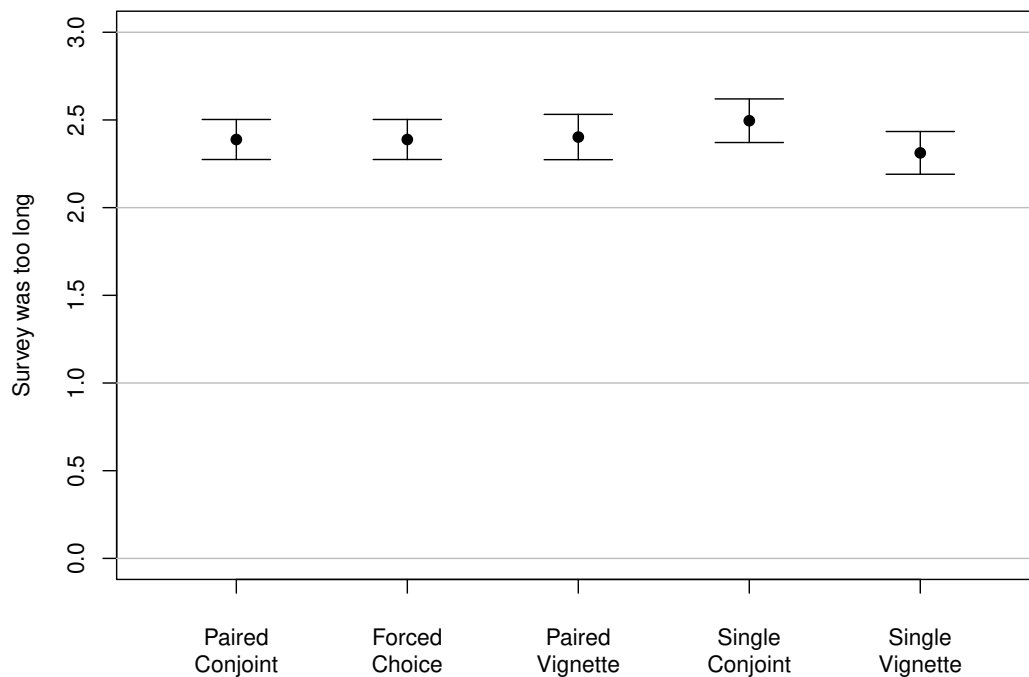


Figure shows estimated means and corresponding 95 % confidence intervals for perceived of survey length. At the end of the survey, respondents were asked if they agree that the survey was too long (4: completely agree, 3: agree, 2: neither, 1: disagree, 0: completely disagree)

Figure S14: Perceived Survey Difficulty Across Survey Designs

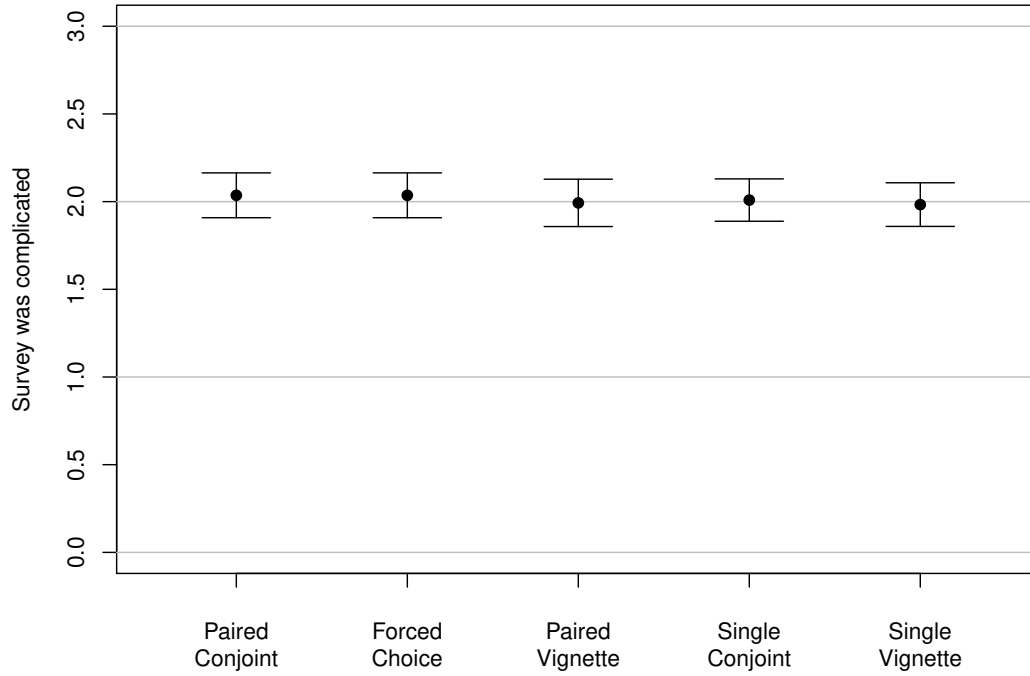


Figure shows estimated means and corresponding 95 % confidence intervals for perceived survey difficulty. At the end of the survey, respondents were asked if they agree that the survey was “complicated” (4: completely agree, 3: agree, 2: neither, 1: disagree, 0: completely disagree)

## References

- [1] Hainmueller J, Hangartner D (2013) Who gets a swiss passport? a natural experiment in immigrant discrimination. *American Political Science Review* 107:159–187.
- [2] Hainmueller J, Hopkins D (Forthcoming) The hidden american immigration consensus: A conjoint analysis of attitudes toward immigrants. *American Journal of Political Science* p DOI: 10.1111/ajps.12138.
- [3] Hainmueller J (2011) Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20:25–46.