Willingness to Say? Optimal Survey Design for Prediction

Charlotte Cavaillé, Karine van Der Straeten, Daniel L. Chen

WORKING PAPERS

N° 1424

March 2023

"Willingness to Say? Optimal Survey Design for Prediction"

Charlotte Cavaillé, Daniel L. Chen and Karine Van Der Straeten

Toulouse School of Economics

Willingness to Say? Optimal Survey Design for Prediction*

Charlotte Cavaille
Ford School of Public Policy
University of Michigan

Daniel L. Chen
Toulouse School of Economics
Institute for Advanced Study in Toulouse
CNRS

Ritesh Das
World Bank

Karine Van der Straeten
Toulouse School of Economics
Institute for Advanced Study in Toulouse
CNRS

March 31, 2023

Abstract

Survey design often approximates a prediction problem: the goal is to select instruments that best predict the value of an unobserved construct or a future outcome. We demonstrate how advances in machine learning techniques can help choose among competing instruments. First, we randomly assign respondents to one of four survey instruments to predict a behavior defined by our validation strategy. Next, we assess the optimal instrument in two stages. A machine learning model first predicts the behavior using individual covariates and survey responses. Then, using doubly robust welfare maximization and prediction error from the first stage, we learn the optimal survey method and examine how it varies across education levels.

*Chen and Van der Straeten acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

1 Introduction

Despite concerns with subjective measures (Bertrand and Mullainathan 2001), collecting attitudinal survey data is now part of researchers' and policymakers' basic tool kit.[1] As we demonstrate in this paper, recent machine learning advances in the estimation of individualized treatment assignment rules can be used at the piloting stage to inform survey design.

When deciding what type of survey instrument to use, researchers are in many cases solving a prediction problem. If the goal is to measure an unobserved construct, they will prefer instruments with higher validity, that is, instruments that best predict observed outcomes related to the unobserved construct being measured (e.g., a depression scale that predicts clinical depression symptoms). When the goal is to collect data to inform resource allocation (e.g., microtargeting of voter outreach efforts based on predicted turnout), the predictive goal is explicit: researchers will prefer survey instruments that maximize predictive accuracy. When choosing between competing survey instruments, a second concern is comparability across groups: a survey instrument can be optimal for one category of respondents but not for another.

We interpret the choice of a survey instrument among existing alternatives as a treatment choice problem, where the survey method is the treatment. Recent methodological advances combine experimental and observational data to optimize treatment choice. We apply these methods to learn which survey instrument to apply to different categories of survey respondents. The preferred instrument is the one that maximizes predictive accuracy. The recovered treatment assignment rules can highlight important heterogeneity in how people engage with a given instrument. This information, we argue, can inform survey design and help identify potential biases introduced by relying on one instrument over another. We illustrate this application using the measurement of preference intensity and the prediction of voter turnout as our running examples.

Because policy making involves trade-offs, knowing what voters want (i.e., preference orientation) is often not enough. Also important is knowing how much they want it (i.e., preference intensity). Preference intensity can be measured in at least one of two ways. One method asks respondents whether they support or oppose a given policy proposal and how important this policy is to them personally. Responses are collected using a categorical scale, the most common one being the Likert scale. We call these measurement instruments *Likert-type* instruments. Another — Quadratic Voting for Survey Research (QVSR) — gives respondents a fixed budget

[1] Attitudinal data is used to test whether a treatment has the intended effect or probe the mechanism underpinning a causal estimate (Kuziemko et al. 2015). A growing line of research also uses subjective survey data to understand immigration preferences, or examine how people reason about economic issues and form attitudes toward immigration or public policy (Alesina, Miano, and Stantcheva 2022, Stantcheva 2021). When building "black box" predictive models of behavioral outcomes of interest to policymakers (e.g., turnout or recidivism Wadsworth, Vera, and Piech 2018) data scientists routinely include attitudinal data.

to 'buy' votes in favor of, or against, each policy proposal, with the price for each vote increasing quadratically. By design, because of this fixed budget, they cannot report extreme opinions on all issues. In some versions, to buy votes, respondents use credits that have no pecuniary value. In other versions, the same credits can be converted into very small sums of money. We will call these measurement instruments *QVSR-type* instruments.

The underlying premise of QVSR-type instruments is that respondents who are passionate about certain policies will vote more frequently for them while compromising on policies they care less about, thus providing a better measure of preference intensity.[2] QVSR-type instruments also different from Likert-type ones in that they necessitate higher cognitive engagement from survey respondents. Whether this is detrimental to the quality of preference measurement is a priori unclear. On the one hand, higher cognitive engagement encourages respondents to give a more careful answer, thus providing a better measure of the target concept. On the other hand, some respondents might find the instruments too demanding and respond using bias-inducing heuristics instead (Krosnick 1991, Sauer et al. 2011). In the latter case, a simpler survey instrument such as Likert-type would do a better job.

To explore the performance of different survey instruments, we utilize data from a study where a sample of US citizens was randomly assigned to complete the same survey with the only variation being the technology used to measure policy preferences (e.g., support for increased gun control measured using a Likert-type scale versus QVSR).[3] Following the survey, respondents participated in a choice task associated with policy issues discussed in the survey (e.g., making a costly donation to a pro-gun control advocacy group). We also identified respondents who participated in the presidential election using official records.

In this study, we leverage the latest advancements in statistical learning methods (Athey and Wager 2021) to assess the performance of models utilizing QVSR-type instruments in predicting donations and turnout compared to those using Likert-type instruments. Additionally, we investigate if this performance varies across different education levels. Our focus on education stems from its role as an indicator of higher survey engagement costs.

We find that QVSR-type instruments best predict behavior in the donation task, with the best performance for intermediate education levels. Turnout, in contrast, is best predicted when using Likert-type instruments. In that case, we observe only limited heterogeneity across education levels.

The paper is organized as follows. First, we briefly present the methodology used to inform survey design. Second, we present the data used for our two running examples: the measurement of preference intensity and

[2] Cavaille, Chen, and Van der Straeten 2019 propose a theoretical model of answer choice in surveys, where respondents face a trade-off between reporting their "true policy preferences" and paying lip-service to their preferred party's norm. They provide some conditions under which QVSR-type instruments outperform Likert-type instruments in accurately gauging preference intensity.

[3] See Cavaille, Chen, and Van der Straeten 2022 for details on the data collection.

the prediction of voter turnout. Third, we present the results for each. We conclude with a short discussion of this methodology for survey design.

2 Survey Design: A Machine Learning Approach

When designing a survey, researchers often have to choose between several alternatives. It is common practice to run a pilot and see which instrument 'performs' the best, where 'performance' is an umbrella term covering concerns such as minimizing non-responses and survey attrition and maximizing information signal. In many cases, assessing performance amounts to a prediction exercise. For example, researchers seeking to measure an unobserved construct will pick the instrument that best predicts an outcome defined by their validation strategy. When researchers collect new data to better predict an observable outcome, the predictive goal is explicit.

Another concern when designing a survey is heterogeneity in how respondents engage with a given instrument. For example, researchers prefer when non-responses are uncorrelated with theoretically-relevant covariates, which can vary across instruments. If some items are more reliable for some respondents than others, this can bias estimates derived using the data generated by this instrument.

To investigate heterogeneous performance across survey instruments, we turn to statistical methods for policy learning, the subject of recent research (Athey and Wager 2021). Policy learning refers to the mapping of observed features (e.g., education level) on the one hand, to a treatment decision (e.g., which survey instrument to use) on the other. Simply put, the methodology uses a heterogeneous treatment effects estimator to derive a treatment assignment rule that is informative of heterogeneity in an instrument's impact on predictive accuracy.

We use attitudinal data to predict one of two behavioral outcomes. One behavioral outcome is defined by our validation exercise, the other is of interest per se. Our goal is to assess which survey instrument minimizes prediction error and whether this varies across different types of respondents. Being treated with one survey instrument over another has a causal effect on prediction error, which we capture using the Brier score. To estimate heterogeneous treatment effects on the Brier score, we use a doubly robust estimator that relies on propensity re-weighting to minimize bias in the treatment estimates. From this estimator, we can map education levels to the decision to assign a respondent to one type of survey instrument. If the treatment effect on prediction error of being assigned to a given instrument varies across education levels, then this will be reflected in the assignment rule.

3 Application: How to Best Measure Preference Intensity and Predict Voter Turnout?

The most standard instrument to measure policy preferences in surveys is the Likert item (Likert, for short). Respondents asked about their policy preferences using Likert items most often answer two consecutive questions: “Do you favor, oppose, or neither favor nor oppose: [Example] Giving same-sex couples the legal right to adopt a child?" followed by “Do you favor [oppose] that a great deal, moderately, or a little?” (Malhotra 2009). Respondents who select “neither" are not asked a follow-up question. Responses range from -3 (strongly oppose) to 3 (strongly favor) and are centered around 0 (neither/nor). Likert items’ wording suggests that answers collected using this survey tool capture a mix of preference alignment (whether people agree or disagreement with the specific details of a policy proposal) and preference intensity (how strongly they care about the issue at stake). Researchers interested in collecting additional information on preference intensity can combine the first question from the Likert item with an issue importance item that asks respondents whether or not an issue is personally important to them (Likert+, for short). Responses are collected using a categorical scale ranging from ‘not at all important’ (1) to ‘very important’ (5) (Miller and Peterson 2004; Howe and Krosnick 2017). The resulting variable is an ordinal scale ranging from -5 (oppose, extremely important) to $+5$ (favor, extremely important) and centered around 0 (neither/nor).

One potential problem with such Likert-type ordinal survey instruments is that survey respondents have only limited incentives to consider trade-offs across issues or how they prioritize these issues. They can report that they strongly support many proposals and that all are very important to them. In short, they place respondents in a world of abundance, and people’s answers may convey little information regarding which issues they prioritize and which issues they are willing to compromise on.

To address this limitation, researchers have designed survey instruments that force respondents to convey which issues they prioritize and which issues they don’t. One recent example is Quadratic Voting for Survey Research (QVSR). This method mimics real-world trade-offs by asking respondents to "vote" on a bundle of issues under conditions of scarcity: respondents are constrained by a fixed budget with which to ‘buy’ votes. Because on one given issue, the price for each vote is quadratic, it becomes increasingly costly to acquire additional votes to express more intense support for (or opposition to) the same issue. In one version of QVSR, respondents use credits that have no pecuniary value (QVSR for short). In another version, respondents have the option of converting unused credits into a very small amount of money. We call this technique QVSR with

Numeraire (QVSRN for short).[4] These QVSR-type instruments are a promising technology in that, by forcing respondents to consider trade-offs, they might theoretically elicit more information (Cavaille, Chen, and Van der Straeten 2019). Empirical studies indicate that, at least in the American context, QVSR-type instruments seem to generate more information than Likert-type instruments (see for example Quarfoot et al. 2017 and Cavaille, Chen, and Van der Straeten 2022).

Yet, a concern with QVSR-type instruments might be that, by requiring the allocation of a fixed budget across multiple items using quadratic pricing, they are arguably cognitively more costly than Likert-type instruments. In other words, decisions with QVSR require higher engagement with the survey, that is, more attention paid to the answer one is providing. On the one hand, this could be a desirable feature of QVSR-type technologies: inducing better survey engagement it might force respondents to think harder about their true policy priorities, improving the quality of survey answers. On the other hand, if the cognitive costs of using QVSR are too high for some individuals, they may drop out altogether or use easy ways out, such as allocating all their credits to one single issue. If this is the case, one might expect the relative performance of the different survey techniques to vary with some observable respondent characteristics, such as education level.

Another issue beyond cognitive costs is that of interpersonal comparisons: to what extent can we assume that 3 votes for one individual (9 credits) are similar to 3 votes for another individual? Note that this issue is a concern for most measurement tools. For example, in Likert+, not everyone imparts the same meaning to the "extremely important" response category. Still, in the case of QVSR, if the fixed budget is more constraining for some respondents than for others, then using votes to compare individuals will introduce noise. As a way to minimize this concern, we also assign respondents to QVSRN. With this instrument, respondents had the option to convert unused credits at a rate of 1 credit = $0.10.

When choosing a survey instrument to measure preference intensity, researchers face a possible trade-off between information gains and information loss due to the complexity of the tool. To assess this trade-off, we use the machine learning approach described earlier. To examine which instrument best measures preference intensity, we need a validation outcome, i.e., an outcome variable that plausibly proxies for preference intensity. For this, we use an incentivized behavior, namely donations to non-profit organizations advocating policy changes on various policy issues asked about in the survey.

We repeat this approach using voter turnout. In this case, our goal is not to choose the optimal measurement strategy but to optimize predictive accuracy as a goal in itself. Indeed, pollsters care about correctly predicting

[4] In the setup used in this paper, respondents were asked to vote on 10 issues using 100 credits. With QVSRN, respondents had the option to convert unused credits at a rate of 1 credit = $0.10.

levels of turnout. For example, in a country like the US, where electoral rules give a central role to geography, capturing the extent to which geographically concentrated sub-populations turn out at different rates is key. The 2016 election, the focus of our analysis, stood out for the salience of issues such as immigration and abortion, with Trump promising to build a border wall and pack the courts with anti-abortion judges. Such issues can motivate voters to register and turn out. We examine whether turnout models that include survey data on policy preferences perform better when these preferences are measured using Likert-type instruments or QVSR-type ones and whether this performance varies across education levels.

4 Intervention and Data Collection

We randomly assigned survey respondents to take the same survey varying only the survey instrument used to measure policy preferences. We test four survey instruments: Likert, Likert+, QVSR and QVSRN. The main component of the survey took place from October 5 to October 9, 2018 (for a detailed description of the experiment, see Cavaille, Chen, and Van der Straeten 2022 and Appendix A.1). The survey was administered to a general population of English-speaking US citizens over the age of 18 who reside in the US. The survey company, GfK-Ipsos, uses a probability-based web panel designed to be representative of the US population. Respondents were randomly assigned to one of the survey tools and asked whether they favor or oppose the following 10 policy proposals:

1. Giving same-sex couples the legal right to adopt a child
2. Laws making it more difficult for people to buy a gun
3. Building a wall on the US border with Mexico
4. Requiring employers to offer paid leave to parents of new children
5. Preferential hiring and promotion of blacks to address past discrimination
6. Requiring employers to pay women and men the same amount for the same work
7. Raising the minimum wage to $15 an hour over the next 6 years
8. A nationwide ban on abortion with only very limited exceptions
9. A spending cap that prevents the federal government from spending more than it takes
10. The government regulating businesses to protect the environment

Table 1 summarizes sample sizes. For each treatment: the first line gives the number of subjects who consented to take part in the study; the second line gives the number of respondents who answered the questions

about the 10 policy proposals, completed the donation task, and answered a number of questions we use as controls; the third line gives the number of respondents who answered the questions about the 10 policy proposals and answered a number of questions we use as controls (we got their actual turnout from GfK, which explains why the sample size is larger in the turnout prediction exercise than in the donation sample). If we define dropout rates based on respondents who consented but are not included in the donation task sample, they are 10%, 9%, 13%, and 14% for Likert, Likert+, QVSR and QVSRN, respectively. A large part of the differential attrition between the QVSR/QVSRN on the one hand and Likert/Likert + on the other hand occurred when respondents were asked to watch the video showing them how QVSR and QVSRN work. We found no evidence that dropping out was correlated with education level and other proxies of cognitive engagement (see Appendix A.2). The original study focused on comparing Likert, Likert+ and QVSR. QVSRN was included with the goal of collecting pilot data. Given smaller sample sizes, the QVSRN analysis should be interpreted with more caution.

Table 1: Sample Sizes

	Likert	Likert+	QVSR	QVSRN
Consented	1342	1337	1309	441
Donation	1206	1211	1134	380
Voter turnout	1299	1300	1218	412

Table 2: Outcome Variables.

Variable	Description	Mean
Voter Turnout	Turnout as recorded in official administrative records.	0.65
Donation	Equal to 1 if donated to any of the advocacy groups and 0 otherwise.	0.59

At the end of the survey, all respondents were given the opportunity to make donations to charities advocating on both sides of two highly-salient political causes: gun control and immigration. This donation task elicited the willingness to incur a monetary cost to promote a political cause one agrees with. At the beginning of the task, respondents were told that, as participants, they had been automatically entered into a lottery with "a prize of $100 for 40 randomly selected respondents (among 4000 or so)." They were then prompted to imagine that they were among the winners and asked whether they wanted to donate part of their lottery money to an advocacy group. They had a choice between four advocacy groups working in the areas of immigration or gun control. For each issue area, the chosen organizations fell on different sides of the political divide: for and against immigration, as well as for and against gun control. Respondents could choose not to donate or to donate

to one of the four advocacy groups. Whatever they did not donate they could keep. Table 2 provides an overview of the outcome variables. Socio-demographic variables (including education) were provided directly by GfK-Ipsos and were not asked in the survey. Turnout was also provided by GfK-Ipsos, and measured using official administrative records.

5 Analysis

The goal is to pick the survey instrument that best measures the concept of interest (e.g., preference intensity, as proxied by the decision to donate) or best predicts the outcome of interest (e.g., turnout). If we think of a survey instrument as the policy, then we can use policy learning algorithms to inform survey design.

Our data set includes attitudinal data, a treatment variable capturing which instrument was used to measure the latter (Likert, Likert+, QVSR or QVSRN), demographic covariates and outcome variables, as defined by the validation strategy (i.e., donation decision) or the prediction goal (i.e., turnout). We build a model using the survey responses and the demographic covariates (age, sex, education, ethnicity, income, employment status and party affiliation) to predict outcomes of interest. Prediction error is captured by the Brier score. By estimating the expected treatment effect on the Brier score, we estimate an assignment rule which assigns individuals to survey instruments solely based on their observed demographic features. With this analysis, we can assess whether being exposed to a specific survey instrument is, for individuals with a given set of covariates, particularly effective at accurately predicting the outcome of interest.

To identify the expected treatment effects conditional on observed covariates, we follow a two-step process. First, we build separate prediction models for each survey instrument using both survey responses and demographic covariates as predictors. To find which machine learning model is most suitable, we horse race several models and compare their predictive accuracy based on their F1 score. The predicted mean squared errors are summarized in Appendix A.3. Using predicted squared errors from the best machine learning models, we then calculate the Brier score to be used as the outcome variable in the next step. In total, with 4 instruments and 2 outcomes, we select 8 different models. Demographic and attitudinal predictors are the same across all models.[5] In contrast, the machine learning models retained to compute the Brier score differ across each of the 8 instrument/outcome combinations.

Second, we use a policy learning algorithm (Athey and Wager 2021), where the survey method is the policy. The algorithm maps the covariates to a particular treatment arm that results in the least Brier score. Since the

[5] With that caveat that the same 10 attitudes are measured using different survey instruments.

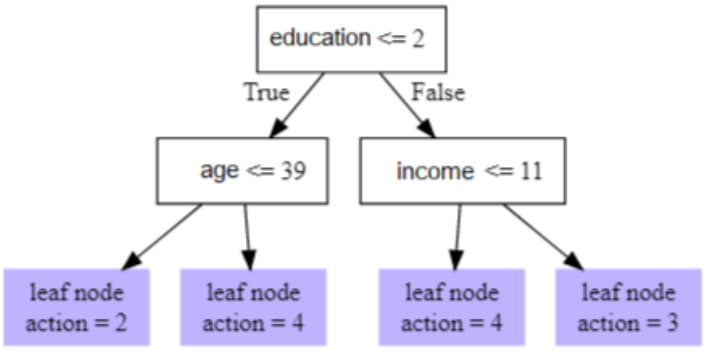

Figure 1: A sample iteration of the depth-2 policy tree.

algorithm assigns the treatment arm based on the maximum predicted treatment impact, the chosen treatment arm for the covariates is the one that yields the best prediction of behavior given the survey responses and the covariates. The algorithm computes the doubly robust scores for the treatment effect and learning policies by empirical maximization of the treatment effect. It uses the augmented inverse-propensity weighted scores (Robins, Rotnitzky, and Zhao 1994) with nuisance component estimates from generalized random forests (Athey, Tibshirani, and Wager 2019; Breiman 2001).

The policy assignment results in a sample iteration of the depth-2 policy tree as illustrated in Figure 1. We have four possible actions: 1 (assign to Likert), 2 (Likert+), 3 (QVSR), and 4 (QVSRN). In this figure, the policy tree algorithm hypothetically assigns QVSRN (action = 4) to individuals having lower educational qualifications and higher age and also to individuals having higher education and relatively lower income levels. It assigns Likert+ (action = 2) to individuals having lower education and lower age and QVSR (action = 3) to higher education and higher income individuals.

Based on the treatment assignment by the policy learning algorithm, we inspect for potential education-specific differences in treatment assignments. Specifically, we aggregate across all recommended actions, with aggregation taking place within a given education level. Put differently, while we are using all covariates when assigning the optimal policy, in our analysis, we aggregate across branches of the policy tree to capture how often the policy learning tree algorithm assigns those of a specific education level to a specific survey instrument.

Constructing confidence intervals is difficult to do with machine learning methods. As a second best, to estimate confidence intervals around the heterogeneous impacts of machine learning predictions, we use the bootstrap method. We bootstrap the policy assignment step and construct empirical confidence intervals that test the relative significance of the treatment heterogeneity conditional on specific demographic characteristics. Bootstrapping uses random sampling with replacement (e.g., mimicking the sampling process) in which we

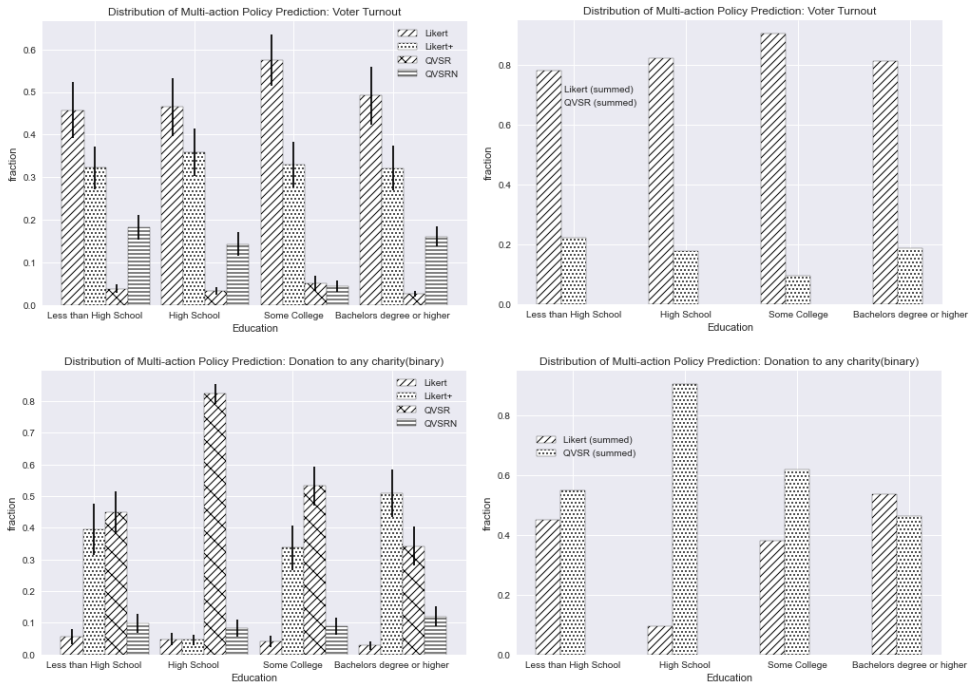

Figure 2: Optimal Assignment Based on Education Levels.

randomly select the training set for each iteration to train the policy tree that maps the covariates to the treatment arms. We then aggregate specific covariates for all the iterations. The aggregation yields the mean and 95% confidence intervals for each predicted treatment arm.

The final results are summarized in Fig 2. Each panel plots, by education level, the share of respondents assigned to a given instrument by the policy tree algorithm. The top row displays the results for voter turnout, and the bottom row for the donation variable. Each panel displays a bar plot that demonstrates the potential heterogeneous treatment effects based on education levels. The left column distinguishes the four different survey instruments, i.e., Likert, Likert+, QVSR, and QVSRN. The right column groups Likert and Likert+ into Likert-type instruments and QVSR and QVSRN into QVSR-type instruments. Any differences in assignment across education levels indicate that treatment effects vary by education.

As Figure 2 shows, if the goal is to predict donation behavior, then QVSR-type instruments are better suited than Likert-type ones: for a majority of respondents, QVSR is the optimal instrument, especially so for those with intermediate level of education. With regards to turnout, the results reverse: Likert-type instruments are better than QVSR-type instruments at predicting who cast a ballot in the 2016 elections. In that case, there is limited evidence of significant differences across education levels.

6 Discussion

As this analysis shows, survey instruments are not born equal: their performance varies with a researcher's goal. Under the assumption that donation behavior provides a reasonable proxy of preference intensity, our results indicate the QVSR-type instruments should be preferred when testing mechanisms hypothesized to affect preference intensity, or when probing causal relationships with effect sizes that vary with preference intensity.[6] We found some evidence of differences across education levels, with the performance of QVSR-type instruments the highest for intermediate levels of education. This suggests that this family of instruments, while perhaps encouraging some respondents to default to easy and noisy answers, encourages even more respondents to better engage with the survey and give answers that are more informative of preference intensity than standard survey instruments such as Likert scales.[7]

This conclusion does not extend to predicting turnout: when building a model of turnout that includes attitudinal data, researchers are better off relying on Likert-type instruments. What explains this difference? According to Cavaille, Chen, and Van der Straeten 2019, this could be due to Likert-type items' higher sensitivity to partisan signaling, thus providing an indirect measure of partisan strength. The intuition is straightforward: on highly-salient issues, when using Likert-type items, strong Republicans and strong Democrats are more likely to provide end-of-scale answers on a large subset of items. By design, QVSR-type instruments force such respondents to "de-bunch", i.e., prioritize voting on some issues at the expense of others. For strong partisans, this de-bunching introduces noise. If turnout is better predicted by partisanship strength than single-issue voting, then it becomes more optimal to rely on Likert-type items.

More generally, the analysis presented in this paper illustrates how advances in machine learning techniques can help inform survey design. We have discussed two main applications of these techniques. One is the development of a measurement instrument for an unobserved subjective quantity of interest. As we illustrate with the example of preference intensity, machine learning techniques can help navigate the trade-offs between conceptual fit on the one hand, and noise introduced because of differential engagement with the instrument on the other. This application requires identifying a second outcome (e.g., donation behavior in our running example) that will be used to assess predictive accuracy. In other words, it puts a strong emphasis on criterion validity, i.e., the extent to which a new measure predicts behavior in the future or alternative gold standard

[6] For example, the decision to switch party in response to a change in party platform will depend on preference intensity, see Carsey and Layman 2006

[7] As shown in Quarfoot et al. 2017, people interact with QVSR through an iterative process of vote adjustments. By forcing respondents to think in terms of trade-offs, QVSR-type instruments could make it easier for people to converge on an answer that better aligns with their underlying preferences.

measures made at the time of survey administration or shortly after (Boateng et al. 2018).

The second application discussed in this paper is black box prediction. Such exercises tend to rely on existing data to select the model with the best predictive accuracy. Using advances in machine learning described in this paper, an alternative is to improve model accuracy by collecting new subjective data. Given the black-box nature of the exercise, one could imagine using this methodology to fine-tune which survey instruments to use depending on a respondent's observable features. Take for example a researcher seeking to predict consumption behavior using both objective predictors like income and subjective predictors such as inflation expectations. Using this methodology, one might find that asking for specific inflation estimates is optimal for some respondents while asking for less precise estimation of overall inflation trends might be better for others.

While our running examples focus on subjective survey data, the methodology described in this paper extends to any situation in which collecting a given set of variables has a cost. This might be the case for example when building a dataset combining variables from administrative sources. Such endeavors can be costly, imposing a limit on the number of variables researchers can link. Relatedly, privacy concerns can favor collecting some variables over others. When deciding which variables are worth collecting and merging, researchers can use the optimal design strategy described in this paper to guide their decision-making process.

References

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva (2022). "Immigration and Redistribution". In: *The Review of Economic Studies*.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". In: *Annals of Statistics* 47.2, pp. 1148–1178.

Athey, Susan and Stefan Wager (2021). "Policy Learning With Observational Data". In: *Econometrica* 89.1, pp. 133–161.

Bertrand, Marianne and Sendhil Mullainathan (2001). "Do People Mean What They Say? Implications for Subjective Survey Data." In: *American Economic Review* 91.2, pp. 67–72.

Boateng, G.O. et al. (2018). "Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer." In: *Frontiers Public Health.*, pp. 6–149.

Breiman, L. (2001). "Random Forests." In: *Machine Learning* 45, pp. 5–32.

Carsey, Thomas and Geoffrey Layman (Jan. 2006). "Changing Sides or Changing Minds? Party Identification and Policy Preferences in the American Electorate". In: *American Journal of Political Science* 50, pp. 464–477. DOI: 10.2307/3694284.

Cavaille, Charlotte, Daniel L. Chen, and Karine Van der Straeten (2019). "Towards a General Theory of Survey Response: Likert Scales vs. Quadratic Voting for Attitudinal Research." In: *University of Chicago Law Review Online 22*.

— (2022). "Who Cares? Measuring Preference Intensity in a Polarized Environment." In: *TSE Working Papers 22-1297*.

Howe, L. C. and J. A. Krosnick (2017). "Attitude strength." In: *Annual Review of Psychology* 68, pp. 327–351.

Krosnick, J. A. (1991). "Response strategies for coping with the cognitive demands of attitude measures in surveys." In: *Applied Cognitive Psychology* 5.3, pp. 213–236.

Kuziemko, Ilyana et al. (2015). "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." In: *American Economic Review* 150.4, pp. 1478–1508.

Malhotra, N.K. (2009). "Review of Marketing Research". In: *Emerald Group Publishing Limited* 5.

Miller, J. M. and D. A. M. Peterson (2004). "Theoretical and Empirical Implications of Attitude Strength." In: *The Journal of Politics* 66.3, pp. 847–867.

Quarfoot, David et al. (2017). "Quadratic voting in the wild: real people, real votes." In: *Public Choice* 172, pp. 283–303.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of Regression Coefficients When Some Regressors are not Always Observed." In: *Journal of the American Statistical Association* 89.427, pp. 846–866.

Sauer, Carsten et al. (2011). "The Application of Factorial Surveys in General Population Samples: The Effects of Respondent Age and Education on Response Times and Response Consistency." In: *Survey Research Methods* 5, pp. 89–102.

Stantcheva, Stefanie (2021). "Understanding Tax Policy: How do People Reason?" In: *The Quarterly Journal of Economics* 136.4, pp. 2309–2369.

Wadsworth, Christina, Francesca Vera, and Chris Piech (2018). "Achieving Fairness through Adversarial Learning." In: *arxiv*.

A Appendix

A.1 Overview of Survey Design

To recruit participants, we relied on the GFK/Ipsos KnowledgePanel. It is the oldest and largest probability-based online panel in the U.S.—with about 60,000 members. Panelists take on average two to three KnowledgePanel surveys a month, minimizing respondent fatigue and attrition. Panel participants are rewarded through the provision of free internet and a tablet to access it. Participants' consent was obtained on the first page of the survey. On this page, we provided information on the topic of the survey, the length and potential benefits from participating (entering a $100 lottery). We clearly stated that the survey was anonymous.

To allocate respondents across the four survey tools, we used a randomized-block design (see details in Cavaille, Chen, and Van der Straeten 2022). We first formed 27 blocks on the basis of partisan identity (Republican, Independent, Democrat), subjective ideology (liberal, middle of the road, conservative), and vote in 2016 (Clinton-other, Trump, did not vote/too young to vote). These variables are important predictors of individuals' policy positions on politicized issues such as immigration, gay rights or budget deficits, as well as predictors of partisan identity and partisan strength. Within each block, we implemented a complete randomization.

Balance tables are included below (Tables A1-A4). They reveal only small differences, whether in terms of socio-demographic covariates or donation behavior. For example, more respondents donated to a gun advocacy group in the Likert condition but this difference is substantively small and non-significant when using donation amounts. We also find no evidence that people who own a gun, have a gun in the household, or have no gun (see "gun ownership score") are distributed differently across the treatment conditions. The same apply to people with an immigrant background (see "immigration score").

Table A1: Balance Table — Likert Treatment —

	Likert +/QVSR-QVSRN (pooled)	Likert	Δ
% donated (gun)	0.372	0.340	-0.032*
	(0.483)	(0.474)	(0.017)
% donated (wall)	0.223	0.233	0.010
	(0.416)	(0.423)	(0.014)
Amount donated† (gun)	10.275	8.649	-1.626
	(34.004)	(32.871)	(1.164)
Amount donated† (wall)	1.347	1.989	0.642
	(28.484)	(27.629)	(0.976)
Turnout	0.647	0.668	0.021
	(0.478)	(0.471)	(0.016)
Party identity (1-7)	4.182	4.111	-0.071
	(2.215)	(2.203)	(0.076)
Ideology (1-7)	4.099	4.113	0.013
	(1.619)	(1.618)	(0.056)
% women	0.510	0.491	-0.020
	(0.500)	(0.500)	(0.017)
Age	52.230	52.420	0.191
	(16.102)	(16.606)	(0.562)
White	0.754	0.758	0.003
	(0.430)	(0.429)	(0.015)
Black	0.090	0.079	-0.011
	(0.286)	(0.269)	(0.010)
Other	0.069	0.070	0.002
	(0.253)	(0.256)	(0.009)
Hispanic	0.087	0.093	0.006
	(0.282)	(0.290)	(0.010)
HS or less	0.280	0.310	0.030*
	(0.449)	(0.463)	(0.016)
Some college	0.317	0.288	-0.029*
	(0.465)	(0.453)	(0.016)
BA or more	0.403	0.402	-0.001
	(0.491)	(0.491)	(0.017)
Income	13.424	13.364	-0.060
	(4.523)	(4.562)	(0.157)
Gun ownership score (0-2)	0.640	0.676	0.036
	(0.873)	(0.886)	(0.030)
Proximity to immigration score (1-3)	1.240	1.246	0.006
	(0.598)	(0.603)	(0.021)
Observations	2,725	1,206	3,931

† Respondents could donate anywhere between 0 and 100 dollars. For these variables, this amount is multiplies by -1 if respondents donate to the pro-gun or anti-immigration non-profits. As a result, these variables captures both how much people donated and in which "direction."

Table A2: Balance Table — Likert + Treatment —

	Likert/QVSR-QVSRN (pooled)	Likert +	Δ
% donated (gun)	0.358	0.372	0.014
	(0.480)	(0.483)	(0.017)
% donated (wall)	0.228	0.222	-0.006
	(0.420)	(0.416)	(0.014)
Amount donated† (gun)	9.496	10.407	0.912
	(33.012)	(35.094)	(1.163)
Amount donated† (wall)	1.415	1.834	0.419
	(27.520)	(29.750)	(0.975)
Turnout	0.649	0.665	0.016
	(0.478)	(0.472)	(0.016)
Party identity (1-7)	4.151	4.182	0.032
	(2.207)	(2.222)	(0.076)
Ideology (1-7)	4.110	4.088	-0.022
	(1.602)	(1.655)	(0.056)
% women	0.500	0.515	0.016
	(0.500)	(0.500)	(0.017)
Age	52.387	52.066	-0.321
	(16.182)	(16.428)	(0.562)
White	0.753	0.761	0.008
	(0.431)	(0.426)	(0.015)
Black	0.085	0.089	0.004
	(0.279)	(0.285)	(0.010)
Other	0.069	0.069	-0.001
	(0.254)	(0.253)	(0.009)
Hispanic	0.092	0.081	-0.011
	(0.289)	(0.273)	(0.010)
HS or less	0.292	0.283	-0.009
	(0.455)	(0.451)	(0.016)
Some college	0.304	0.316	0.012
	(0.460)	(0.465)	(0.016)
BA or more	0.404	0.400	-0.004
	(0.491)	(0.490)	(0.017)
Income	13.404	13.409	0.005
	(4.574)	(4.446)	(0.157)
Gun ownership score (0-2)	0.664	0.620	-0.044
	(0.883)	(0.863)	(0.030)
Immigration score (1-3)	1.242	1.240	-0.002
	(0.600)	(0.599)	(0.021)
Observations	2,720	1,211	3,931

† Respondents could donate anywhere between 0 and 100 dollars. For these variables, this amount is multiplies by -1 if respondents donate to the pro-gun or anti-immigration non-profits. As a result, these variables captures both how much people donated and in which "direction."

Table A3: Balance Table — QVSR Treatment —

	Likert-Likert+/QVSRN (pooled)	QVSR	Δ
% donated (gun)	0.360	0.369	0.009
	(0.480)	(0.483)	(0.017)
% donated (wall)	0.223	0.234	0.011
	(0.416)	(0.423)	(0.015)
Amount donated† (gun)	9.574	10.275	0.701
	(34.103)	(32.570)	(1.185)
Amount donated† (wall)	1.904	0.655	-1.249
	(28.141)	(28.414)	(0.993)
Turnout	0.660	0.638	-0.022
	(0.474)	(0.481)	(0.017)
Party identity (1-7)	4.157	4.170	0.014
	(2.210)	(2.216)	(0.078)
Ideology (1-7)	4.097	4.120	0.023
	(1.631)	(1.587)	(0.057)
% women	0.507	0.498	-0.009
	(0.500)	(0.500)	(0.018)
Age	52.214	52.471	0.257
	(16.489)	(15.673)	(0.572)
White	0.758	0.749	-0.010
	(0.428)	(0.434)	(0.015)
Black	0.084	0.093	0.009
	(0.277)	(0.290)	(0.010)
Other	0.070	0.066	-0.004
	(0.256)	(0.249)	(0.009)
Hispanic	0.087	0.093	0.005
	(0.282)	(0.290)	(0.010)
HS or less	0.292	0.281	-0.011
	(0.455)	(0.450)	(0.016)
Some college	0.304	0.317	0.012
	(0.460)	(0.465)	(0.016)
BA or more	0.403	0.402	-0.001
	(0.491)	(0.491)	(0.017)
Income	13.389	13.445	0.056
	(4.520)	(4.571)	(0.160)
Gun ownership score (0-2)	0.649	0.656	0.008
	(0.876)	(0.878)	(0.031)
Immigration score (1-3)	1.245	1.233	-0.012
	(0.602)	(0.594)	(0.021)
Observations	2,797	1,134	3,931

† Respondents could donate anywhere between 0 and 100 dollars. For these variables, this amount is multiplies by -1 if respondents donate to the pro-gun or anti-immigration non-profits. As a result, these variables captures both how much people donated and in which "direction."

Table A4: Balance Table — QVSRN Treatment —

	Likert-Likert+/QVSR (pooled)	QVSRN	Δ
% donated (gun)	0.360	0.384	0.024
	(0.480)	(0.487)	(0.026)
% donated (wall)	0.230	0.195	-0.035
	(0.421)	(0.397)	(0.023)
Amount donated† (gun)	9.768	9.855	0.087
	(33.552)	(34.744)	(1.817)
Amount donated† (wall)	1.510	1.858	0.348
	(28.615)	(24.282)	(1.523)
Turnout	0.657	0.618	-0.039
	(0.475)	(0.486)	(0.026)
Party identity (1-7)	4.154	4.218	0.064
	(2.213)	(2.197)	(0.119)
Ideology (1-7)	4.107	4.075	-0.032
	(1.620)	(1.602)	(0.088)
% women	0.502	0.532	0.030
	(0.500)	(0.500)	(0.027)
Age	52.316	52.032	-0.284
	(16.249)	(16.349)	(0.878)
White	0.756	0.750	-0.006
	(0.429)	(0.434)	(0.023)
Black	0.087	0.084	-0.003
	(0.281)	(0.278)	(0.015)
Other	0.068	0.076	0.008
	(0.253)	(0.266)	(0.014)
Hispanic	0.089	0.089	0.001
	(0.284)	(0.286)	(0.015)
HS or less	0.292	0.266	-0.026
	(0.455)	(0.442)	(0.024)
Some college	0.307	0.318	0.012
	(0.461)	(0.466)	(0.025)
BA or more	0.402	0.416	0.014
	(0.490)	(0.494)	(0.026)
Income	13.405	13.408	0.003
	(4.524)	(4.631)	(0.245)
Gun ownership score (0-2)	0.651	0.653	0.002
	(0.876)	(0.887)	(0.047)
Immigration score (1-3)	1.240	1.259	0.019
	(0.599)	(0.612)	(0.033)
Observations	3,551	380	3,931

† Respondents could donate anywhere between 0 and 100 dollars. For these variables, this amount is multiplies by -1 if respondents donate to the pro-gun or anti-immigration non-profits. As a result, these variables captures both how much people donated and in which "direction."

A.2 Survey attrition

Respondents assigned to Likert or Likert + had a 10% probability of dropping out after having consented to taking the survey. For QVSR and QVSRN, the probability is 14%, with the additional attrition originating from having to watch a video.

Table A.2 examines whether observable covariates help predict who, in the QVSR and QVSRN treatment groups, is most likely to drop out once asked to watch the explanatory video. People who did not vote in 2016 are slightly more likely to drop out in QVSR-type instruments. Effect sizes are substantively small and similar to those found for Likert +. Older people are also more likely to drop out, something true across survey instruments.

Table A5: Predicting Dropout

(OLS)	QVSR-QVSRN b/se	Likert b/se	Likert+
Voted in 2016	-0.048**	-0.015	-0.042**
	(0.017)	(0.016)	(0.016)
Partisanship	-0.004	0.003	-0.006
	(0.005)	(0.005)	(0.005)
Ideology	-0.002	0.007	-0.004
	(0.007)	(0.006)	(0.006)
Gender	0.007	0.025	0.006
	(0.017)	(0.015)	(0.015)
Age	0.002***	0.002***	0.001*
	(0.001)	(0.000)	(0.000)
Black [ref: White]	0.002	-0.013	0.009
	(0.031)	(0.029)	(0.027)
Other	0.006	0.000	0.031
	(0.034)	(0.031)	(0.030)
Hispanic	-0.007	0.044	0.034
	(0.031)	(0.028)	(0.029)
Some college [ref: HS or less]	-0.001	-0.024	0.008
	(0.022)	(0.019)	(0.019)
BA or more	-0.009	-0.028	0.010
	(0.022)	(0.020)	(0.020)
Income	-0.001	-0.001	-0.005*
	(0.002)	(0.002)	(0.002)
Gun ownership score	-0.008	-0.015	0.010
	(0.010)	(0.009)	(0.009)
Prox. to immigration score	0.017	-0.012	-0.000
	(0.015)	(0.014)	(0.013)
_cons	0.053	-0.039	0.115
	(0.070)	(0.060)	(0.059)
N	1719	1303	1298

A.3 Machine Learning: Mean Squared Errors

Table A6: Mean squared errors for all the ML models for each of the survey methodologies.

ML Method	Mean Squared Error (voter turnout)	Mean Squared Error (donation)
	Likert	
Logistic Regression	0.3406	0.3785
Decision Tree	0.4227	0.4069
Random Forest	0.3375	0.3596
XGBoost	0.3659	0.3880
K Nearest Neighbor	0.3343	0.3880
ADA Boost	0.4195	0.3848
Elastic Net	0.2340	0.2490
	Likert+	
Logistic Regression	0.3501	0.4542
Decision Tree	0.4290	0.4889
Random Forest	0.3501	0.4353
XGBoost	0.3343	0.4952
K Nearest Neighbor	0.3722	0.4195
ADA Boost	0.4195	0.5047
Elastic Net	0.2312	0.2483
	QVSR	
Logistic Regression	0.3402	0.4604
Decision Tree	0.3814	0.4432
Random Forest	0.3642	0.3951
XGBoost	0.3608	0.4432
K Nearest Neighbor	0.3333	0.4226
ADA Boost	0.3573	0.4776
Elastic Net	0.2223	0.2484
	QVSRN	
Logistic Regression	0.3762	0.5049
Decision Tree	0.4356	0.4356
Random Forest	0.3465	0.4059
XGBoost	0.4455	0.3366
K Nearest Neighbor	0.4257	0.4257
ADA Boost	0.4455	0.3861
Elastic Net	0.2379	0.2517