Who Cares?

Measuring Differences in Preference Intensity

Charlotte Cavaillé*
Ford School of Public Policy, University of Michigan
Daniel L. Chen
Toulouse School of Economics - IAST - CNRS
Karine Van der Straeten
Toulouse School of Economics - IAST - CNRS

How well do existing survey instruments differentiate between opinions that affect individual behavior and opinions that don't? To answer this question, we randomly assigned U.S. respondents to one of three survey instruments: Likert items (Likert), Likert items followed by personal importance items (Likert+) and Quadratic Voting for Survey Research (QVSR), which gives respondents a fixed budget to buy 'favor' or 'oppose' votes, with "vote" price increasing quadratically. We find that, relative to Likert, both Likert+ and QVSR better identify people who care enough about an issue to act in opinion-congruent ways, with QVSR offering the most consistent improvement overall. Building on these results, we show how conclusions regarding the relationship between policy opinions and self-interest can differ across measurement strategies.

Word Count: 9,658 with references; 8,958 without references.

Keywords: Survey research, Public opinion, Preference intensity, Attitude extremity, Attitude importance, Likert item, Quadratic voting for survey research, Self-interest, Determinants of policy preferences

* cavaille@umich.edu

People have opinions on many issues but do not care about all of them. If people care deeply about an issue, they are more likely to act in opinion-congruent ways, e.g., object (through voting or protest) when a proposed policy does not align with what they prefer. If people are mostly indifferent, then they are more likely to compromise, e.g., accept an outcome that does not align with what they prefer (by going to the beach instead of protesting). Conceptually, these differences across opinions —which we will later call differences in preference intensity—are central to theories of democratic accountability (Hill 2022), issue voting (Rabinowitz and Macdonald 1989) and party-switching (Carsey and Layman 2006).

How do researchers empirically differentiate between opinions that affect political behavior and opinions that don't? One type of measurement strategy sequentially asks people how they feel about a set of issues, with responses recorded using a uni or bidirectional scale. This *simply-ask* approach assumes that respondents, when prompted with words such as "strongly" or "important," report to the best of their ability how much they care about an issue. An alternative approach seeks to elicit the same information by forcing respondents to trade-off across all issues considered jointly. In this case, expressing their opinion on some issues comes at the expense of doing so on others: respondent have to choose. Such *forced-choice* approach assumes that respondents arbitrate in ways that are informative of how people behave when confronted with the costs of real world opinion-congruent action.

From the existing research, we already know that the *simply-ask* approach helps distinguish people who care about an issue from those who don't: survey respondents who indicate feeling "strongly" about a policy they find "very important," are more likely to behave in opinion-congruent ways than people who pick the "weakly" and "not important" response categories (e.g., Krosnick and Petty (1995); Carsey and Layman (2006)). For people with well-formed opinions on issues asked about in the survey and no obvious reasons to misrepresent their "true" opinion, this approach should suffice. Yet, concerns that few respondents match this profile have lead many scholars to avoid subjective survey data altogether (Bertrand and Mullainathan 2003). *Forced-*

choice elicitation strategies could offer a compromise solution (Cavaillé, Chen and Van Der Straeten 2019; Hanretty, Lauderdale and Vivyan 2020). As we discuss later in the paper, they can help people better realize "in the moment" how important an issue is to them or mitigate the measurement bias introduced by competing motives such as partisanship. Alternatively, if concerns regarding the *simply-ask* approach are overblown, then applied researchers need not rush to look for better ways of identifying "who cares."

To investigate these issues, we asked respondents from a representative sample of U.S. citizens their opinion on 10 policy issues, randomly varying the method used to measure their opinion. One method is the Likert item (Likert for short), which asks people to report (on a 3-point scale) the "strength" of their support/opposition. The other method (Likert+) combines the Likert item with a personal importance item that asks respondents to further specify if the issue previously mentioned is "personally important" to them (on a 5-point scale). The third method —Quadratic Voting for Survey Research (QVSR)—uses a variant of the *forced-choice* approach. It gives respondents a fixed budget to 'buy' votes in favor or against the 10 policy proposals, with the price for each vote increasing quadratically. Because of this price schedule, it becomes increasingly costly to acquire additional votes to express more intense support or opposition to a given policy (Lalley and Weyl 2018). After expressing their opinion using one of these three measurement tools, respondents performed a number of choice tasks commonly associated with issue-specific political action (e.g., a donation to a non-profit advocating for gun control or letter writing to a senator about a minimum wage bill). We compare each tool's ability to distinguish between respondents whose opinion-congruent behavior suggests they care intensely about a given issue and those whose non-congruent behavior suggests they do not care as much.

First, we document Likert's reasonably good performance. We find that the addition of a personal importance item offers some improvement, with QVSR offering the most consistent improvement overall. One important difference appears to be QVSR's ability to de-bunch in informative ways, that is, generate meaningful differences in votes

cast among people who, under alternative measures, would end up picking the same response category.[1]

Because of these differences, the measurement strategy used to measure people's opinions has implications for applied research. We demonstrate this point by revisiting the claim, common among public opinion scholars, that people's policy opinions do not reflect their material self-interest (e.g., Sears and Funk 1990). In line with previous studies, we find that support for a policy measured using a Likert scale conveys only limited information about a respondent's position as a potential beneficiary of this policy. In contrast, QVSR votes help distinguish between respondents who would directly benefit from a policy and respondents who would not be affected. This suggests that conclusions regarding the importance of material self-interest can vary with the measurement strategy used to measure individual support for a given policy.

1 Measuring Who Cares? Conceptual and Theoretical Considerations

Consider a status quo changing policy (e.g., Brexit). Of all people who favor this policy, when given the opportunity (e.g., a referendum), only a subset will translate this support into opinion-congruent action (e.g., turn out to vote in favor). Formally, we capture these individual differences with a real number u_{ik} in the interval $[-1,1]$, where the likelihood of taking costly action in favor of the reform k and against the status quo increases as u_{ik} gets closer to 1. Conversely, the likelihood of taking costly action against the reform and in favor of the status quo increases as u_{ik} gets closer to -1. The preference ranking captured by u_{ik} can be further decomposed into two terms. One, preference orientation, is an indicator variable which captures whether the respondent prefers the reform over the status quo ($u_{ik} > 0$) or the status quo over the reform ($u_{ik} < 0$). The other, preference intensity is the *extent* to which the respondent prefers

[1] For readers familiar with theses concepts: QVSR improves over Likert on both discrimination and calibration (e.g., Tetlock (2017)). For some outcomes, QVSR improves over Likert+ on calibration.

one over the other and is captured by the absolute value of u_{ik} ($|u_{ik}|$).

Both preference orientation and intensity vary with the specific content of the proposed policy and how it compares to the status quo. Preference intensity, in addition, varies with the relative importance of the policy domain. Take, for example, two individuals with moderate preferences in favor of a given reform. One individual might feel strongly about the policy domain yet be quite satisfied with the status quo, resulting in moderately intense preferences. Another might not care much about a policy domain yet be dissatisfied enough with the status quo to be moderately supportive of the reform. People with the most intense preferences will include those who care strongly about a given policy domain and are highly dissatisfied with the status quo.

As a summary concept aimed at describing a complex subjective mental state, u_{ik} cannot be observed, it can only be imperfectly measured. To recover meaningful information about u_{ik} in general, and $|u_{ik}|$ in particular, researchers rely on two broad families of measurement strategies. Next, we discuss the pros and cons of each. For expository purposes, we build our discussion around the specific measurement strategies used in the empirical section of this study.

1.1 The Simply-Ask Approach: Likert and Personal Important Items

When measuring u_{ik}, researchers who favor the *simply-ask* approach described in the introduction most often rely on the two-step version of the Likert item (Malhotra, Krosnick and Thomas 2009). First, respondents are asked if they "favor, oppose, or neither favor nor oppose" a status quo changing policy k. Respondents who pick the favor or oppose option then see the following prompt: "Do you favor [oppose] that a great deal, moderately, or a little?" Respondents who initially select 'neither nor' are not asked a follow-up question. Recorded responses range from -3 (strongly oppose) to $+3$ (strongly favor) and are centered around 0 (neither-nor). Once normalized, the resulting response variable $\widehat{u}_{ik}^{L}$ ranges from -1 to 1.

A common practice is to supplement information provided by Likert items using a

follow-up personal importance item. This item asks respondents "how important" a given issue is to them "personally." Respondents answer using a categorical scale ranging from 'not at all important' (1) to 'extremely important' (5) (Miller and Peterson (2004); Howe and Krosnick (2017)).[2] A recurrent finding is that opinion-congruent behavior is higher among people who "strongly favor" a policy and among those who report finding the issue personally important to them (Krosnick and Petty 1995: e.g.). In other words, both Likert and personal importance items recover meaningful information about u_{ik} in general and $|u_{ik}|$ in particular. Combined with a Likert item, this suggest a second straightforward way of measuring u_{ik}, namely:

$$\widehat{u}_{ik}^{L+} = F\left(\widehat{u}_{ik}^{L}, \widehat{Imp}_{ik}\right)$$

with the answers to the personal importance item denoted by $\widehat{Imp}_{ik}$.

Before discussing the pros and cons of this *simply-ask* approach, a quick note on how preference intensity, per our definition, relates to similar concepts in public opinion research. We have defined our main quantity of interest in reference to a spatial model of politics most commonly found in political economy (see Appendix B.1 for more details). Likert and personal importance items were developed by social psychologists to measure what is called attitude extremity and attitude importance. While attitude extremity captures "the degree to which the person likes or dislikes the object," attitude importance captures "an individual's subjective judgment of the significance he or she attaches to his or her attitude" (Howe and Krosnick 2017: 329).[3] What is the relation between u_{ik} (the combination of preference orientation and intensity) on the one hand, and attitude extremity and importance on the other? There are significant epistemological differences underpinning spatial models' emphasis on preference ranking and

[2] Per this definition and measurement strategy, personal importance is distinct from "national importance" which captures respondents' subjective evaluation of how important a policy is "for the country as a whole" (Miller, Krosnick and Fabrigar 2017: 157).

[3] Krosnick and Abelson (1992) identify at least five attitude features. Given our interest in explaining *"attitude-congruent" behavior*, we leave aside attributes (e.g., attitude accessibility) most relevant for cognition and attitude formation only (Howe and Krosnick 2017).

social psychology's emphasis on attitudes, which we discuss in Appendix A. Still, for our purpose, we can put these differences aside: u_{ik} is, by construct, the *total sum effect* of attitude extremity, attitude importance and any other attitude features that affect the decision to act in an opinion-congruent way or compromise instead.[4] Our goal is to *measure* differences in u_{ik} to the best of our abilities, not to explain these differences. As a result, attitude extremity and importance are absent of our conceptualization or analysis: preference intensity supersedes these concepts. Note that a concept such as attitude strength, which Krosnick and Abelson (1992) define as the extent to which a given attitude "affects one's cognition or behavior," does not provide an adequate substitute for preference intensity as defined here. One important reason is that social psychologists relate strong attitudes to stable attitudes that are hard to change. In contrast, based on our definition of $|u_{ik}|$, preference intensity can vary over time depending, for example, on changes in the status quo.[5]

The main advantage of the *simply-ask* approach is its simplicity. One major disadvantage for researchers interested in measuring preference intensity $|u_{ik}|$ is that it puts respondents in a world where talk is cheap.[6] First, there are no consequences for misrepresenting one's true opinion or reporting an opinion even if one has none. A second concern is that respondents are asked about policy issues sequentially, with no incentives to arbitrate between intense preferences for two mutually exclusive policies.

If people have some prior sense of how their opinion on one issue compares to their opinion on another and report these truthfully, these concerns might be relatively minor. But scholars have reasons to worry. Partisan motives have been shown to system-

[4] Given the different epistemological starting, building a 1-to-1 conceptual match between extremity and importance on the one hand, and preference and preference intensity on the other, is far from straightforward. We discuss this in more detail in Appendix B.2.

[5] The concept of preference intensity, per our definition, is also different from that of "attitude intensity," which social psychologists define as "the strength of the emotional reaction provoked by the attitude object" (Krosnick et al. 1993: 1132). In contrast, preference intensity is defined only in reference to the *net* utility of the proposed change relative to the status quo, which might or might not be driven by an emotional response to an object.

[6] Note that this is not a concern for people interested in measuring attitudes, which are conceptually different from the preference ranking defined as u_{ik}.

atically bias survey responses (Bullock and Lenz 2019). In the U.S. context, polarized ideological messaging and affective partisanship can generate bi-modal response distributions. In this case, the same response category (e.g., 'favor a great deal' or 'very important') might include respondents who care about the issue and respondents who do not care as intensely and are merely "paying lip service to the party norm" (Zaller 2012). Not only do researchers have limited variation to build on, whatever variation they have, it is difficult to interpret. Furthermore, with only two parties to choose from, many U.S. voters hold a combination of mutually exclusive policy preferences: behaving in an opinion-congruent way on one policy often means having to compromise on another (e.g. support for Republicans' strong stance on balanced budgets means compromising on support for abortion rights). With the *simply-ask* approach, respondents are in a world of abundance, where compromise is not needed, meaning that the information recovered might carry too little information about opinion-congruent behavior in the real world.

1.2 Forced-Choice Approach: Quadratic Voting for Survey Research

These concerns have lead some scholars to turn away from subjective survey data and stated preferences and rely instead on in-survey behavioral outcomes in the form, for example, of a donation or a real effort task. While ideal for studies limited to one or two issues, in-survey behavioral proxies are difficult and/or costly to scale up to include a larger number of issues. An intermediate solution is to rely on stated preferences but use a measurement strategy that leverages a force-choice design that makes talk a little less cheap by confronting people with trade-offs.

QVSR, developed by Posner and Weyl (2018), is one such measurement strategy.[7] Like Likert items, it asks respondents the extent to which they favor a given set of policies,

[7] Another type of *forced-choice* method asks respondents to choose between bundles of policies (Hainmueller, Hopkins and Yamamoto 2014; Hanretty, Lauderdale and Vivyan 2020; Sides, Tausanovitch and Vavreck 2022). Because these methods return *group-level* estimates of preferences (instead of measuring individual-level differences), they are outside our scope of inquiry (Abramson, Koçak and Magazinnik 2022; Ganter 2023).

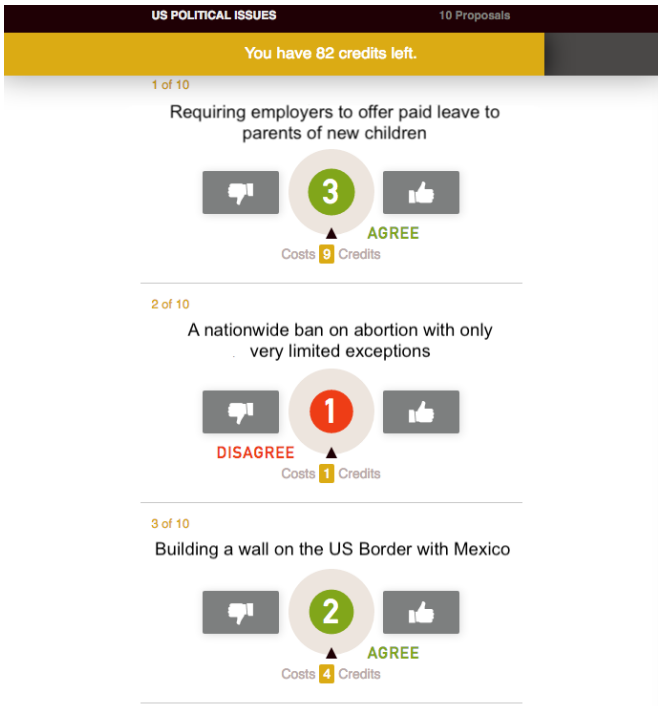
but the technology used to measure people's answers is very different. Respondents express their preferences on a bundle of policies under the constraint of a fixed budget of credits with which to buy units of support (votes in favor) and units of opposition (votes against).[8] A distinctive feature of QVSR is that the price schedule is quadratic: buying one vote for one proposal costs one credit; buying two units for the same proposal costs four credits; buying three units costs nine credits; and so on. In our own survey, respondents assigned to QVSR were given a budget of 100 credits to spend across ten different survey questions. Figure 1 shows what such survey looks like to respondents. Respondents can scroll down to report their preferences on all the issues examined in the survey. Remaining credits are displayed at the top of the screen. Respondents can go back to revise their answers until they are satisfied with how they have allocated their credits. The maximum that respondents can spend in favor or against any question is 10 units of support/opposition (which costs 100 credits) though doing so would mean not being able to express (however mild) support for or opposition to any of the other 9 issues. Respondents do not have to spend all of their 100 credits.[9] Recorded responses range in theory from -10 to +10. Once normalized, the resulting response variable $\widehat{u}_{ik}^{QVSR}$ ranges from -1 to 1.

QVSR's *forced-choice* design compels individuals to compare across issues. This can improve the quality of responses in three ways. First, QVSR better approximates the real-world opportunity costs of opinion-congruent behavior. Second, it does not require people to have well-formed opinions: by forcing people to compare across issues, QVSR can induce people to themselves realize what it is they care the most about. Third, as discussed in Appendix B, when partisan concerns generate misreporting and end-

[8] A related method asks respondents to rank policies by order of importance. When choosing a *forced-choice* method to evaluate, we opted for QVSR for three reasons. First, it jointly measures preference orientation and preference intensity. In contrast, a ranking exercise first requires a battery of Likert items to capture preference orientation. Second, ranking exercises tend to be limited to five items while piloting with QVSR shows that respondents are comfortable with a larger set of items (Quarfoot et al. 2017). Third, QVSR has the potential to generates more information in the form of a cardinal (instead of an ordinal) scale.

[9] Description of how respondents interact with this interface is not the focus of our study, see Quarfoot et al. (2017) for more information.

Figure 1: Screenshot of the QVSR Version of the Survey



of-scale bunching, QVSR forces people to de-bunch in ways that are informative of preference intensity. In the abundance world of Likert and personal importance items, people can inflate their reported preference intensity at no cost. The combination of a fixed budget and quadratic pricing makes this type of misreporting costly: expressing a strong preference (through multiple votes) for a policy one does not care about comes at the cost of doing so for a policy one truly cares about. Take, for example, a set of respondents who all report strongly supporting unrestricted abortion and finding this issue personally important to them. Some might provide these answers because they are sincerely reporting their true u_{ik}. Others might have a lower u_{ik} yet choose end of scale responses out of partisan concerns (e.g., strong support for abortion rights is what defines a strong Democrat). Assuming respondents compromise (in terms of the number of votes cast) on policies they do not sincerely care about, then we can plausibly expect QVSR to be more informative of differences in preference intensity.

Still, QVSR has several important drawbacks. One is that it requires higher cognitive

engagement from survey respondent, something that might improve the quality of responses for some but decrease it for others. For example, some respondents might find the instrument too demanding and respond using bias-inducing heuristics (Krosnick (1991), Sauer et al. (2011)). For these respondents, a simpler survey instrument such as a combination of a Likert and personal importance items would do a better job. A second drawback is that, while plausibly approximating the type of arbitrage most relevant to the measurement of preference intensity, QVSR's budget constraint might also introduce measurement error. For example, if the budget constraint is too constraining, then respondents can end up randomly picking which issue to give fewer votes to in order to free enough credits for other issues. A related concern is that of interpersonal comparisons. Take, for example, two respondents who both used 9 credits (3 votes) to express support for a given proposal: can we reasonably assume that they care about this proposal to the same extent? Note that this issue is a concern for most subjective measurement tools. For example, with personal importance items, not everyone imparts the same meaning to the 'extremely important' response category.

1.3 Comparing Methodologies

How much is gained by measuring preference intensity using QVSR instead of Likert items? Assuming QVSR offer an improvement over Likert items, how does this improvement compare to merely adding a follow-up personal importance item? How much more informative is the personal importance item relative to using a Likert item alone? We conclude this overview by providing speculative, if informed, answers to these questions. Likert provides our benchmark. A measurement strategy, to be of any value, should perform better than simply (and sequentially) asking people how strongly they favor a given set of status quo changing policies.

To compute Likert+, we multiply answers from the Likert and personal importance items. The resulting scale ranges from -15 to $+15$ ('strongly oppose/favor' and 'extremely important') and is centered around 0 (neither-nor). Once normalized, the

response variable $\widehat{u}_{ik}^{L+}$ ranges from -1 to 1. In Appendix E, we discuss alternative ways of combining the information captured by these two survey items. Results remain unchanged.

Mechanically, given the addition of an item, researchers have more variation to work with when using Likert+ than when using Likert. Because Likert+ generates novel information in the form of a new prompt about a different facet of preference intensity, we expect Likert+ to outperform Likert. Compared with Likert, QVSR should also provide a better measure of preference intensity. In contrast to Likert, QVSR is less prone to end-of-scale bunching and forces respondents to engage in between-issue comparison.

How do Likert+ and QVSR compare? As we discuss in Appendix B, the answer partly depends on the strength of the partisan motive and its impact on the prevalence of uninformative end-of-scale bunching. It also depends on the amount of error introduced by QVSR's previously discussed disadvantages. When it comes to comparing QVSR and Likert+, we remain agnostic on which methodology will outperform the other.

In Appendix C, we also offer a systematic comparison of the three methods focusing on costs (software, survey time and drop out rates). The creation of several QVSR web applications[10] have brought software costs down to zero. While median time spent answering preference-related questions is shorter for respondents assigned to Likert, it is roughly the same for respondents assigned to Likert+ and QVSR. The main difference time-wise for QVSR is a 90 second video explaining how the tool works.[11] QVSR, in our study, has one additional extra cost, namely a higher drop out rate (though see Quarfoot et al. (2017) who find no such difference).

As the above discussion suggest, each methodology comes with advantages and disadvantages. Which method outperforms the others is an empirical question. Next, we explain how we propose to answer this question.

[10] See footnote 22 on page 26.

[11] The video can be found at: `https://www.youtube.com/watch?v=GrY_RzDsqLY`.

2 Empirical Design

A measurement tool can be thought of as a classification instrument that distributes the surveyed population across a fixed number of response categories. Each tool differs in terms of the number of available response categories and the technology used to distribute people across categories. The tool that best measures preference intensity is the one that best classifies respondents from the most to least likely to to behave in an opinion-congruent way. To compare each survey tool's classification abilities, we use an experimental design. In this section, we first describe this design. Next, we describe how we use the data collected to compare Likert, Likert+ and QVSR.

2.1 Survey Design

We asked people to take the same survey, randomly varying the measurement tool used to measure policy opinions. The survey was administered to a general population of U.S. citizens over the age of 18 (N = 3,551). The survey company, GfK-Ipsos, uses a probability-based web panel designed to be representative of the U.S. population. The main data collection effort took place from October 5 to October 9, 2018. For an overview of the survey design, see Appendix C and H.

Respondents were randomly assigned to one of the three survey tools and asked to provide their opinion on the following 10 policy issues:[12]

Do you Favor or Oppose:

- **[sameS]** Giving same sex couples the legal right to adopt a child
- **[gunC]** Laws making it more difficult for people to buy a gun
- **[wall]** Building a wall on the U.S. Border with Mexico
- **[paidL]** Requiring employers to offer paid leave to parents of new children
- **[affA]** Preferential hiring and promotion of blacks to address past discrimination
- **[equalP]** Requiring employers to pay women and men the same amount for

[12] In each treatment, the order in which the 10 proposals are presented is fully randomized.

the same work

- **[minW]** Raising the minimum wage to $15 an hour over the next 6 years
- **[abort]** A nationwide ban on abortion with only very limited exceptions
- **[cap]** A spending cap that prevents the federal government from spending more than it takes
- **[env]** The government regulating business to protect the environment

After expressing their opinion, respondents were given the opportunity to take action by donating lottery money to single-issue advocacy groups. First, respondents were told that, as participants to the survey, 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 two issue areas: immigration and gun control. For each issue area, we chose organizations that fall 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, and one only, of the four advocacy groups. Whatever they did not donate, they could keep. Two weeks after the end of the survey, 40 randomly selected respondents received their prize money, which was disbursed by GfK-Ipsos.[13]

Four months later (between January 31 and February 18, 2019), we recontacted a random subset of respondents and asked them to answer the same 10 survey questions using the survey tool they were assigned to in the first wave (number of responses, N= 1569).[14] We then collected information on two additional behavioral tasks.

First, we asked each respondents how they would behave in three dictator games: one involving a Republican, another a Democrat and a third an Independent (the order was

[13] Because of regulations preventing tax-exempt research funds from being used for political purposes —something we failed to realized at the design stage of the study—we ultimately made no donations and, several weeks after the end of the study, the lottery winners received the full $100 amount, alongside an email explaining the reason why.

[14] Participation in wave 2 is not predicted by treatment condition and policy preferences in wave 1, nor by partisanship. See Appendix C for more on balance across treatment conditions in wave 1 and on wave 2 participation.

randomized). Respondents had the option to donate anywhere between $0 and $100 of some lottery money (the set up was similar to the one in wave 1). After they made their decisions, respondents were asked again about their donation to the Independent. We explained that, in wave 1, this Independent had donated to the pro-immigration organization and to the anti-gun control organization.[15] We asked respondents if they wanted to change the amount they had previously decided to donate to this individual. In other words, they had to choose between doing nothing, "punishing" the Independent (by decreasing the amount originally donated) or "rewarding" them (by increasing the amount originally donated). Because few people in our survey (based on wave 1 results) are both pro-immigration *and* anti-gun control, most respondents faced a trade-off: rewarding this fellow survey participant meant condoning a position one is in agreement with while also condoning a position one is in disagreement with.

Second, respondents were also given the opportunity to write to their Senators about real bills that were moving through Congress at the time of the survey. One bill was about abortion and the other was about raising the minimum wage. We did not mention who the bill sponsors were, only the content of the bills. The texts provided by the respondents were then integrated into a letter, which was ultimately sent to the Senate committees in charge of reviewing the policy proposals (Adida, Lo and Platas 2018). Comments were anonymous. This task was designed to capture respondents' willingness to spend time and effort promoting a political cause they agree with.

As we discuss in Appendix C, in the QVSR treatment condition, dropout rates are higher by 13 percentage points. We found no evidence that dropping out was predicted by observable covariates including partisanship and ideology. Table 1 provides an overview of the outcome variables derived from the three behavioral tasks and used in the remainder of the analysis. Throughout the paper, when we examine the relationship between survey answers and behavior, we only use answers collected in wave 1.[16] Using

[15] In practice, this was impossible as respondents could only donate once. When disbursing the funds, we consequently used survey answers, not donation decisions, to identify Independents to disburse the funds to.

[16] The reader should keep in mind however that, in at least two cases (dictator game and letter writing

data collected in wave 2 does not change the results (See Appendix D).

Table 1: Behavioral Outcomes and Relevant Survey Question

Variable	Description	Mean (Stand. dev.)	Survey question expected to correlate with behavior
Donation to gun-related advocacy group	Equal to the $ amount donated multiplied by 1 if donated to pro gun control and -1 if donated to anti gun control advocacy group.	9.8 (33.5)	Laws making it more difficult for people to buy a gun
Donation to immigration-related advocacy group	Equal to the $ amount donated multiplied by 1 if donated to pro immigration and -1 if donated to anti immigration advocacy group.	1.6 (28.7)	Wall on the border with Mexico
Punishment of Independent respondent (1)	Equal to the $ amount *taken off* the amount previously donated to the Independent. If respondent *gave* additional $ then amount multiplied by -1.	4.9 (13.5)	Laws making it more difficult for people to buy a gun/ Wall on the U.S. border with Mexico
Punishment of Independent respondent (2)	Equal to the $ amount *taken off* the amount previously donated to the Independent as a *proportion* of the amount originally donated. If respondent *gave* additional $ then amount multiplied by -1.	0.17 (0.38)	Laws making it more difficult for people to buy a gun/ Wall on the U.S. border with Mexico
Letter writing on the minimum wage bill	Equal to the length of text written (number of characters).	76 (139)	Raising the minimum wage to $15/h over the next 3 years (absolute values)
Letter writing on the abortion bill	Equal to the length of text written (number of characters).	59 (90)	A nationwide ban on abortion with only very limited exceptions (absolute values)

2.2 Estimation Strategy

Each survey tool generates a response variable ($\widehat{u}_{ik}^{L}$, $\widehat{u}_{ik}^{L+}$ or $\widehat{u}_{ik}^{QVSR}$) that differs from the other two in terms of 1) the total number of ordinal categories and 2) the distribution of observations across these categories. Likert has 7 response categories ranging from -3 to $+3$ and Likert+ has 23 response categories ranging from -15 to $+15$. While QVSR has 21 response categories in theory (from -10 to $+10$), in practice, few people

tasks), the attitudinal data is analyzed alongside behavioral data collected four months apart.

put more than 7 votes on the same issue, resulting in 15 response categories (from -7 to 7).[17] To insure comparability, we normalize $\widehat{u}_{ik}^{L}$, $\widehat{u}_{ik}^{L+}$ and $\widehat{u}_{ik}^{QVSR}$ such that the lowest possible answer corresponds to zero ($-3/-15/-7$ for Likert, Likert+ and QVSR respectively) and the highest possible answer to 1 (3/15/7).

As shown in the bottom panel of Figure 2, when preferences are measured using a Likert item, the distribution of answers to the gun control item is uni-modal: answers bunch on one extreme of the scale (i.e., strong support for gun control). This pattern is much less pronounced with Likert+, implying that, while most respondents strongly support gun control, not everyone believes this issue to be personally important to them. Partly by design, responses in QVSR exhibit no such bunching patterns.[18]

More response categories and less bunching imply more information (i.e., higher entropy) for QVSR and Likert+ on the one hand than for Likert on the other.[19] If Likert+ and QVSR's higher entropy is more than just noise then, when comparing individuals with a higher score to individuals with a lower score, the former's behavior should signal more intense preferences than the latter's. Put differently, if a response category is a bin, people in a bin with a higher value should be, on average, more likely to take action then people in a bin with a lower value. Quantitatively, this implies a positive and monotonic relationship between ordinal response categories on the one hand, and the mean of the outcome of interest—conditional on the response category—on the other. We examine this expectation by regressing each of the behavioral outcomes described in Table 1 over the corresponding normalized survey response variable (X) interacted with a categorical variable identifying the method used:

$$Y_i = \sigma_0 + \mu_1 D_{i,Likert+} + \mu_2 D_{i,QVSR} + \sigma_1 X_i + \sigma_2 X_i D_{i,Likert+} + \sigma_3 X_i D_{i,QVSR} + \sigma_4 J_4 + ... + \sigma_j J_j + \varepsilon_i \quad (1)$$

[17] A few respondents chose to vote 8 times or more for the same issue. To avoid presenting results from bins that only include very small numbers of observations, we re-coded the 8 votes or more answers into a 7 votes answer.

[18] See Quarfoot et al. (2017) for within-individual evidence on this de-bunching process.

[19] Shannon entropy scores capturing this difference are provided in Appendix F

where $J_4, ...J_j$ are dummy variables that indicate membership in a block used for block randomization (see Appendix C for more details). Regression coefficients σ_1, $\sigma_1 + \sigma_2$ and $\sigma_1 + \sigma_3$ can be interpreted as the difference between $E(Y/X = 1)$ and $E(Y/X = 0)$ for Likert, Likert+ and QVSR respectively. The better tool is the one with not only more variation (or higher entropy) but also more informative variation in the form of a larger difference between the two quantities of interest, that is, the one with a larger regression coefficient. Monotonicity is also key: in the next section, we assess it visually by plotting the average value of Y_i for all respondents with the same value for X.[20]

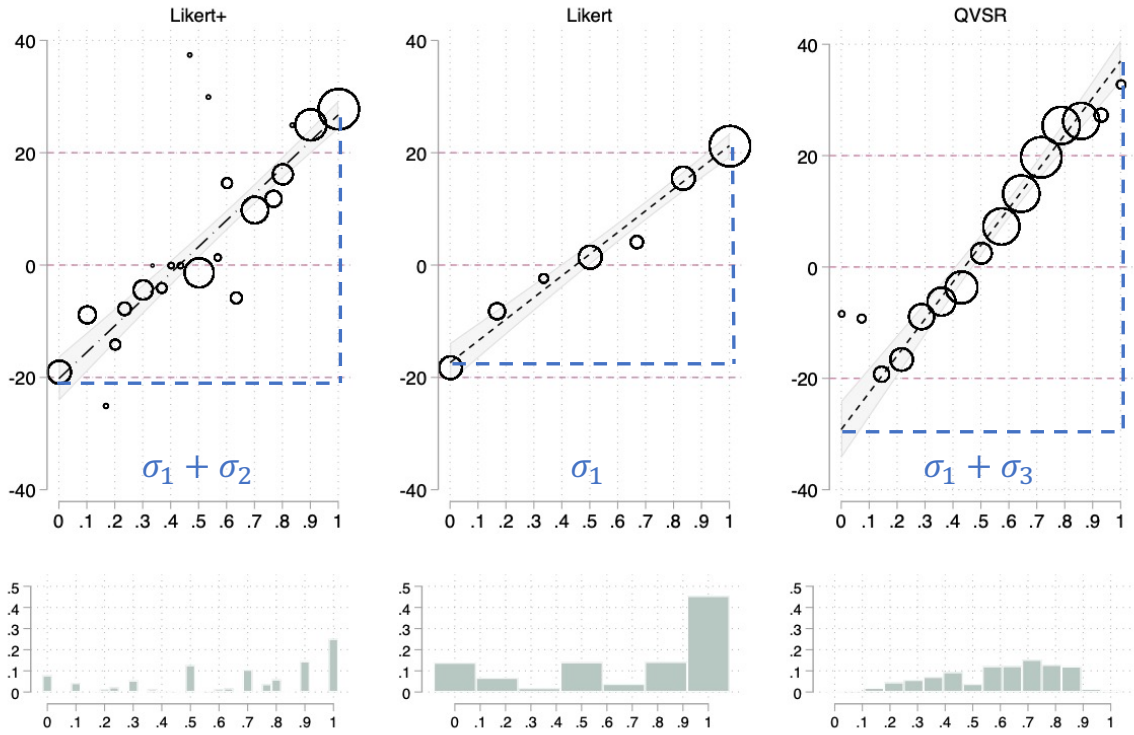
3 Results

Figure 2 plots average donations to the gun control charities by response to the gun control question, further broken down by survey instrument. The lines capture the three regression coefficients mentioned in the previous section (see Figure 3 for the actual estimates). As shown on this figure, the regression slope is larger for QVSR than for Likert. This mean that individuals who choose the end-of-scale response categories in Likert end up de-bunching under QVSR in ways that align with their behavior on the donation task. Specifically, people who donate less choose, on average, smaller values in QVSR than people who donate more. This is captured by the magnitude of the regression slope: individuals who do not donate are no longer pulling the regression slope down by 'sharing' the extreme response categories with people who care enough to donate. Comparing the regression coefficients, we can also see that, in this case, the discrimination achieved with QVSR better aligns with preference intensity than that achieved with Likert+. For all three survey tools, the relationship between response category and average behavior is monotonic. Exceptions are due to sparsely populated bins.

As Figure 2 (center panel) shows, Likert does recover some information about prefer-

[20] We also checked for a non-parametric relationship and find that, for all methods, the standard errors do not allows us to rule out a monotonic relationship.

Figure 2: Donation to Gun-Related Advocacy Group and Responses to Gun Control Item



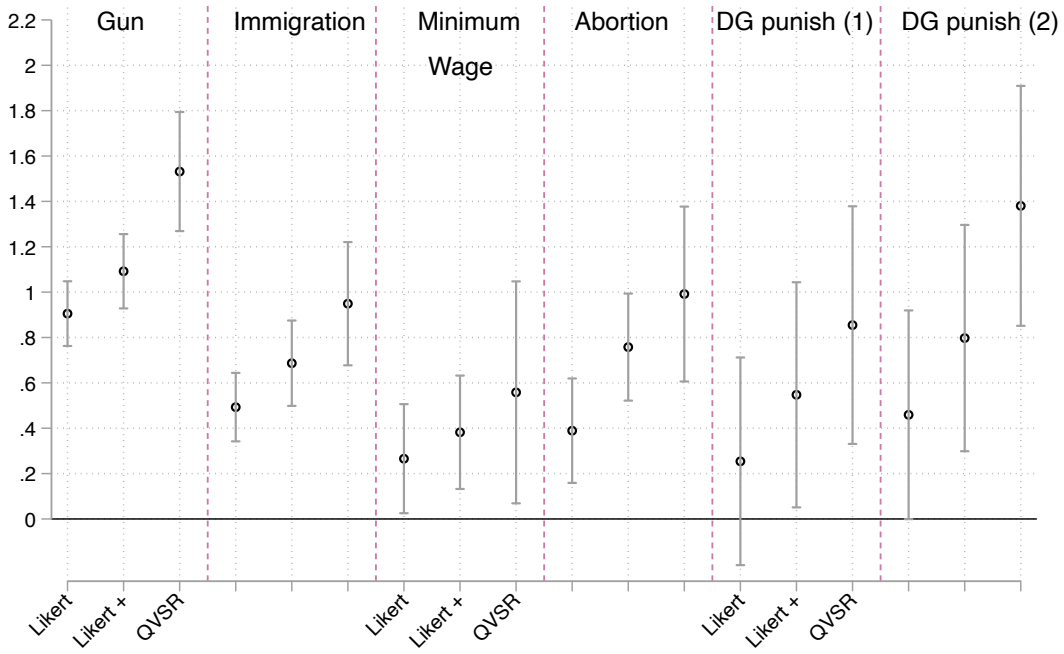
Y-axis: Donation amount. *X-axis:* survey answers by survey method, normalized to vary from 0 to 1. Survey item used: **[gunC]**. *Interpretation*: Scatter plot represents the average donation for respondents with the same X value, i.e., E(Y/X=x). Dots are proportional to the number of observations. Likert, in the center of the figure provides the benchmark. A visual comparison indicates that the coefficient for Likert+ is only marginally larger than that for Likert. Notice the difference in slope between Likert and QVSR. The full estimates are available in Figure 3. Compare also the bunching in Likert and the variation recovered under Likert+ and QVSR. Sparsely populated bins in Likert+ means less than 23 dots are visible to the naked eye.

ence intensity, as proxied by donation behavior. People who 'strongly oppose' (1) or 'strongly favor' (0) gun control donate more dollars to an organization that advocates for their preferred policy outcome than people who only 'oppose' or 'favor' gun control. The benefits of Likert+ and QVSR is that they distribute people across more response categories in ways that are informative of average donation behavior (that is, more variation/less bunching, a larger coefficient and monotonicity).

Figure 3 presents the same analysis for all tasks. Specifically, it plots regression coefficients obtained using equation 1 for all Ys and corresponding Xs described in Table 1. Note a few important differences in how Xs were computed. For the donation outcomes (first two columns in Figure 3), we use the normalized values of the response variables

(X). When predicting the number of characters written, we use the normalized *absolute* values of the response variables (i.e., $0-3$ for Likert, $0-15$ for Likert+ and $0-7$ for QVSR). Indeed, our outcome variable does not capture *what* was written about the bill (i.e., for or against), only the overall effort spent writing about it. When predicting punishment in the dictator games, we use the normalized *difference* between responses on gun control and responses on the border wall (See Table 1). Higher positive values indicate that one favors gun control more intensely than one opposes the wall. Higher negative values indicated that one favors the wall more intensely than one opposes gun control. Given that the Independent recipient in the dictator game was opposed to gun control and opposed to the wall, we examine whether larger differences predict a higher likelihood of punishing the Independent.

Figure 3: Regression Coefficients for all Behavioral Outcomes



Interpretation: a switch from the smallest response category (0) to the largest (1) is associated with a σ increase in Y. The increase is equal to σ times the standard deviation of Y. For the letter writing tasks (Minimum wage and Abortion), the predictor is the normalized absolute value of the response variable. For the punishment task, the predictor is the normalized difference between the gun control and the border wall response variables. For details on each, see text.
* Sample sizes for the Gun and Immigration donation tasks (wave 1) are double the size of the samples sizes for the other tasks (wave 2). As a result, effect sizes are more precisely estimated for these two tasks. For details on each task, see text.

The higher the regression coefficient in Figure 3, the better a given tool is at distin-

guishing between respondents with high and low preference intensity (as proxied by task-specific behavior). Again, Likert's performance is noticeable: in line with the claim that Likert items capture a mix of preference orientation and preference intensity, people with end-of-scale answers behave differently from others (in all cases, the coefficient is positive and substantively large). Overall, the main issue with this measurement tool is whether, on hyper-partisan issues, such as gun control or abortion, there are enough people who do *not* choose end-of-scale answers to identify who truly cares and who doesn't (see Appendix F for response histograms).

While Likert+ appears to carry more information on preference intensity than Likert, its discriminatory power (as captured by $\sigma_1 + \sigma_2$) is statistically indistinguishable from Likert's on all 6 outcomes. Overall Likert+ relative performance is far less consistent than QVSR's. For wave 1 outcomes (donation to an advocacy group task), QVSR outperforms Likert both substantively and statistically. Due to smaller sample sizes, results for wave 2 tasks exhibit larger standard errors. Still, a comparison of regression coefficients suggests that QVSR is more informative of preference intensity than Likert: on all 4 outcomes, QVSR coefficients are at least twice the size of those found with Likert. In contrast, the coefficients for Likert+ represent, relative to Likert, a 50% increase at best.

Because of QVSR's budget constraint, for individuals who use all their credits, votes on one issue is a linear combination of votes on other issues. As a result, the error terms across outcome-specific (or covariate-specific) equations are likely correlated. As a robustness check, we consequently re-run the analyses underpinning Figure 3 and estimate seemingly unrelated regressions models that account for this correlation (Zellner 1962). Table 2 reports differences in coefficient size between methods. The results remain unchanged.

The bottom row of Table 2 reports the F-statistics under the null-hypothesis that, within a data collection wave, the sum of all between-method differences is equal to 0. This allows us to compare the performance of methods within a wave. For both waves,

Table 2: Differences in Coefficient Size (Seemingly Unrelated Models)

	QVSR vs. Likert b/se	Likert + vs. Likert b/se	QVSR vs. Likert + b/se	QVSR vs. Likert b/se	Likert + vs. Likert b/se	QVSR vs. Likert + b/se
Gun	0.61***	0.19	0.41**			
	(0.14)	(0.10)	(0.15)			
Immigration	0.48***	0.21*	0.27			
	(0.14)	(0.10)	(0.15)			
Minimum wage				0.06	0.08	0.11
				(0.25)	(0.19)	(0.29)
Abortion				0.63**	0.40*	0.30
				(0.22)	(0.17)	(0.23)
DG punish (1)				0.59	0.44	0.04
				(0.35)	(0.32)	(0.35)
DG punish (2)				1.00**	0.41	0.50
				(0.37)	(0.32)	(0.37)
N	2336	2471	2343	964	1029	979
F_test	33.9	8.3	11.4	9.1	3.9	1.5
Prob not rej. the null	< 0.000	< 0.004	< 0.000	< 0.003	< 0.05	< 0.22

$^{*}p<.05$, $^{**}p<.01$ $^{***}p<.001$. We replicate Figure 3 analysis using seemingly unrelated models. This table reports the interaction between the preference variable and a dummy variable identifying the survey methods used. For example, for the gun donation outcome, the difference between the coefficient for Likert and that for QVSR is equal to 0.61. Bottom row: F-test of the null-hypothesis that the sum of the coefficients is equal to 0.

the F-statistic is 3 to 4 times larger when comparing QVSR and Likert then it is when comparing Likert+ to Likert, further indicating that, relative to Likert, the information gained with QVSR is substantively larger than that gained with Likert+. Still, when comparing QVSR and Likert+, fewer observations in wave 2 mean we cannot reject the null of no differences between QVSR and Likert+ at conventional levels.

If QVSR, or even Likert+, convey information on preference intensity that is not captured by Likert, then a test of a theory where preference intensity is a theoretically relevant concept could be affected by the measurement tool used. Next, we examine this conjecture focusing on a longstanding debate in political science on the relationship between policy preferences and material self-interest.

4 Where Theory and Measurement Meet

A common starting point when studying preference formation is to expect people to support policies that positively affect their economic conditions and oppose policies that negatively affect them. According to public opinion scholars, this expectation finds limited empirical support. Instead, to explain preference formation, researchers have emphasized non-economic modes of reasoning such as value-based or partisan-

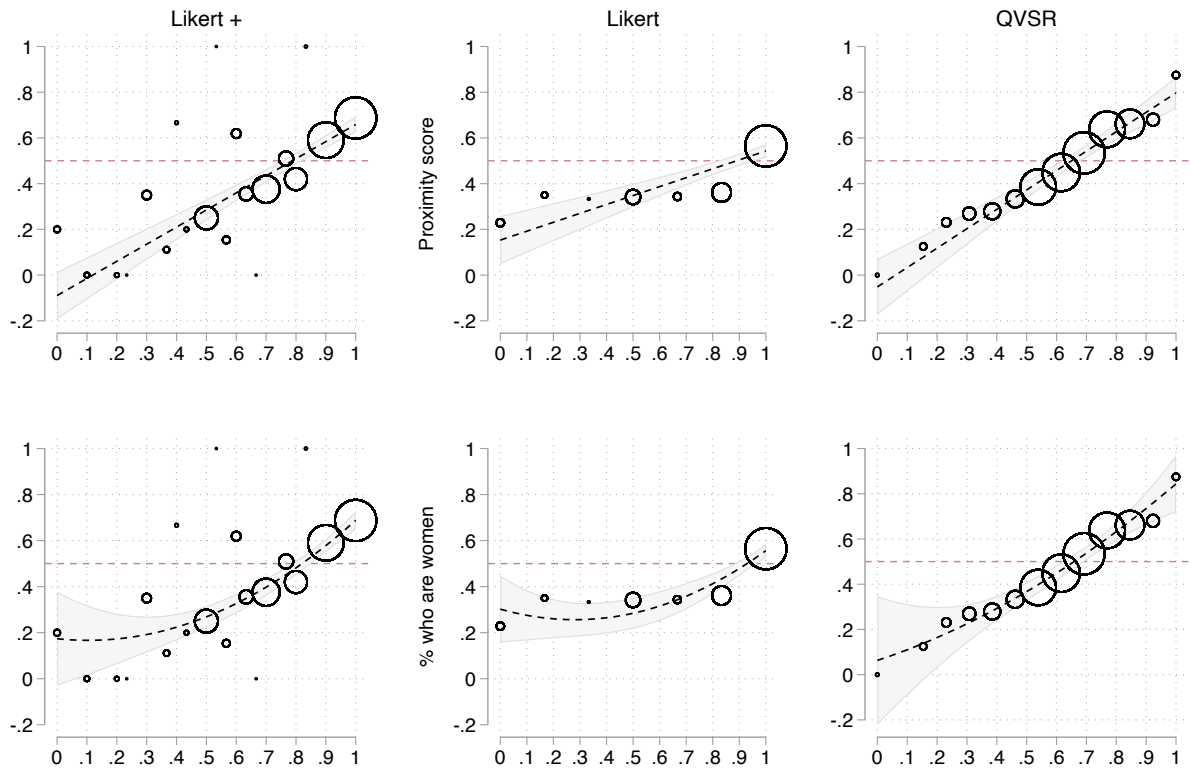
motivated reasoning (Sears and Funk 1990; Margalit 2013; Cavaille 2023). Still, when it comes to preference intensity and the likelihood of behaving in opinion-congruent ways, material self-interest likely plays a key role. For example, while both men and women might support equal pay for equal work out of fairness concerns, when it comes to taking action, women will be more likely to do so than men, meaning that women have stronger preferences on this issue than men. This point has been made repeatedly by John Krosnick when discussing the related concepts of attitude extremity and importance (Howe and Krosnick 2017: 328).

Somewhat surprisingly, empirical analyses of preference formation rarely emphasize the distinction between preference orientation and preference intensity (or related concepts). Yet, this distinction has implications for measurement strategy. If material self-interest is hypothesized to affect preference orientation, then a binary variable measuring support for a given policy should, a priori, be enough to test this argument. If material self-interest is hypothesized to affect preference intensity, then QVSR might be a better measurement strategy.

Figure 4 examines the implication of overlooking the importance of measurement when examining the role of material self-interest. It plots the relationship between gender on the one hand and support for gender equality in the workplace on the other, measured using Likert, Likert+ and QVSR. Notice how, in Likert (middle panel), there is very little variation in survey answers: most people appear to *strongly* support workplace gender equality. The additional information gained by switching from Likert to Likert+ is informative of respondents' gender: women are more likely than men to be in the highest response category. In QVSR, the de-bunching is more consequential and, unlike Likert+, there is a clear linear and monotonic relationship between the number of votes in QVSR and the percentage of women as a share of individuals who cast the same number of votes.

As Figure 5 shows, the same pattern emerges when comparing parental leave preferences and a measure of one's proximity to childbirth. In Appendix G, we show similar

Figure 4: Respondent's Gender and Response to Pay Equity Item



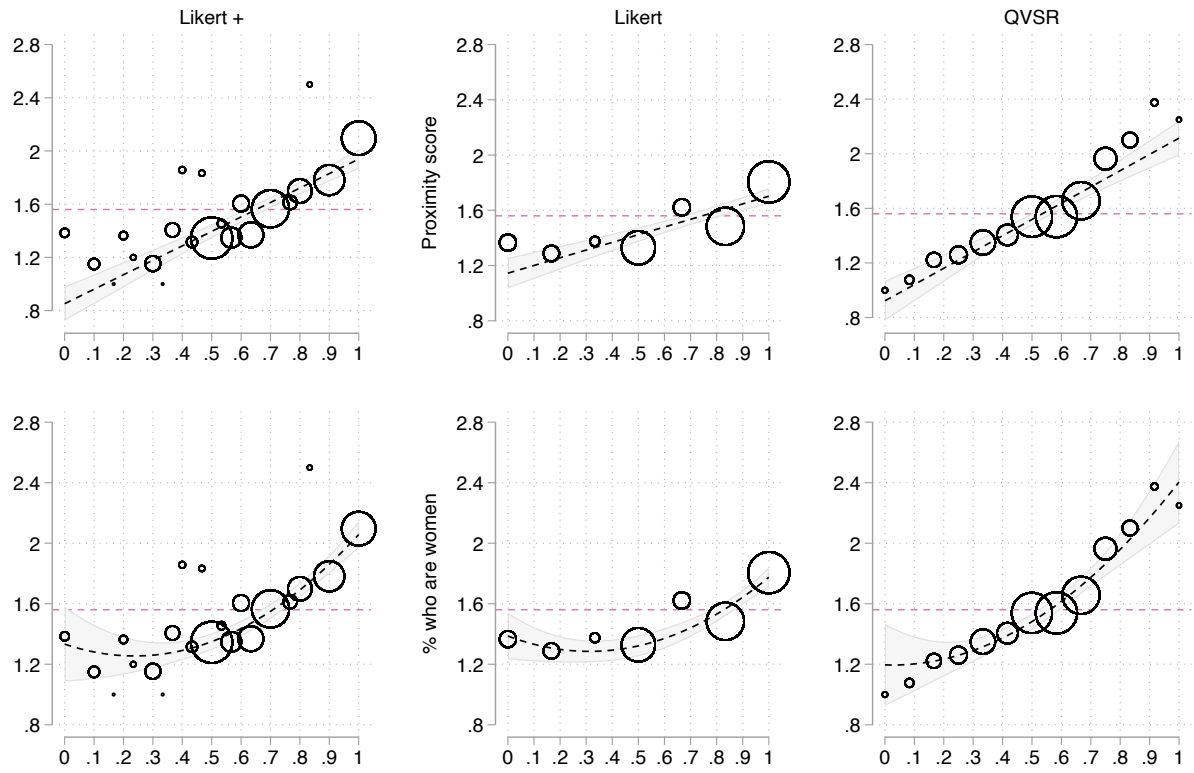
Y-axis: gender of respondent (female = 1, 0 otherwise). *X-axis:* survey answers by survey method, normalized to vary from 0 to 1. Survey item used: **[sameS]**. *Interpretation*: Scatter plot represents the share of women among respondents with the same X value, i.e., E(Y/X=x). Dots are proportional to the number of observations. In the top panel, scatter plots are overlayed with a linear fit line. In the bottom panel, scatter plots are overlayed with a quadratic fit line. Sparsely populated bins in Likert+ means less than 23 dots are visible to the naked eye.

results for affirmative action and race, gun control and gun ownership, as well as increasing the minimum wage and the likelihood of benefiting from such increase. Because differences in preference intensity are imperfectly captured by Likert, using this item alone can produce the type of empirical patterns that have lead researchers to dismiss the theoretical relevance of material self-interest.

5 Conclusion

What do our argument and results imply for scholars interested in measuring preference intensity? Choosing a measurement strategy involves trade-offs between 1) maximizing interpretable variation, 2) minimizing survey costs, and 3) minimizing noise.

Figure 5: Respondent's Proximity to Childbirth and Response to Parental Leave Item



Y-axis: proximity to childbirth score (= 1 if no young child and no plans to have any in future, = 2 young children but no plans to have more, = 3 if children planned or just had a child). *X-axis:* survey answers by survey method, normalized to vary from 0 to 1. Survey item used: **[paidL]**. *Interpretation*: see note Figure 4.

QVSR performs well on 1) in the form of a larger number of better discriminating response categories. When it comes to documenting the importance of material self-interest, this can have substantive implications. QVSR does marginally worse on 2) in the form of longer survey time and more respondents dropping out. On 3), the improvement is minimal: standard errors remain similarly sized across methods, meaning that, as the quality of the signal increases, so does the noise, thus keeping the signal-to-noise ratio somewhat stable. In QVSR's case, this could be due to the type of measurement error induced by a budget constraint that is too tight for some or too loose for others.

Additional work is thus needed to better understand where *forced-choice* methods like QVSR succeed and where they can be improved. For example, is the bulk of the work done by forcing respondents to consider issues jointly or does the quadratic pricing also play a key role? How might design-tweaks —e.g., using a different cost function—help

improve the noise-to-signal ratio? To answer these questions, future work could use a linear rather than quadratic pricing, and compare QVSR to ranking methods.[21] Relatedly, we have yet to examine the impact of changing the menu of options: would results differ had we included an item on the introduction of a wealth tax, or one on reparations? Within individuals, we would expect differences in the number of option-specific votes cast in one menu versus another. Still, how this will affect the cardinal information conveyed by the votes remains to be investigated. To facilitate such follow-up studies, we have made available a web application enabling researchers to vary QVSR's key features including pricing (e.g., linear versus quadratic) and the number of credits relative to the number of options.[22] We hope this will help spur future innovations in the measurement of policy opinions.

Ultimately, which measurement strategy to choose will depend on the type of financial constraints a researcher faces (e.g., survey time) as well as the type of policy issue being measured. When it comes to highly politicized issues, individual-level variance is much lower in Likert than in Likert+ and QVSR, which speaks in favor of QVSR. For less politicized issues, Likert+ might be enough. Possible menu effects are both a weakness and a strength of QVSR in particular and *forced-choice* methods in general (e.g., conjoint analysis). On the one hand, they raise concerns about cross-study comparisons. On the other, they compel researchers to pick a menu of options that reflect theoretically relevant real-world constraints. This emphasis on theoretically-grounded design, while a weakness for exploratory research might be a strength in the deductive stage of a research project.

If there is one main take-away from our inquiry is that, faced with the expansion of survey-based research beyond descriptive public opinion polls, researchers need to take measurement seriously. Disciplinary boundaries have made it difficult: to the best of our knowledge, this is the first study to systematically compare and contrast measure-

[21] See footnote 8.

[22] The resulting survey can be embedded into other online platforms, such as Qualtrics. This web application can be found at `https://qvsr.io`.

ment strategies derived from two distinct conceptualizations of human cognition and behavior, social psychology' (in the case of Likert/Likert+) and economics (in the case of QVSR). We hope our conceptual and theoretical framework (see Appendix B for the more extensive discussion) will help future scholarship more clearly specify their quantities of theoretical interest and identify the tools and strategies best adapted to measuring them.

References

Abramson, Scott F, Korhan Koçak and Asya Magazinnik. 2022. "What do we learn about voter preferences from conjoint experiments?" American Journal of Political Science 66(4):1008–1020.

Adida, Claire L, Adeline Lo and Melina R Platas. 2018. "Perspective taking can promote short-term inclusionary behavior toward Syrian refugees." Proceedings of the National Academy of Sciences 115(38):9521–9526.

Bertrand, Marianne and Sendhil Mullainathan. 2003. "Enjoying the quiet life? Corporate governance and managerial preferences." Journal of political Economy 111(5):1043–1075.

Bullock, John and Gabriel Lenz. 2019. "Partisan Bias in Surveys." Annual Review of Political Science 22:325–342.

Carsey, Thomas M and Geoffrey C Layman. 2006. "Changing sides or changing minds? Party identification and policy preferences in the American electorate." American Journal of Political Science 50(2):464–477.

Cavaille, Charlotte. 2023. Fair enough? Support for redistribution in the age of inequality. Cambridge University Press.

Cavaillé, Charlotte, Daniel L. Chen and Karine Van Der Straeten. 2019. "A Decision-Theoretic Approach to Understanding Survey Response: Likert vs. Quadratic Voting for Attitudinal Research." University of Chicago Law Review .

Ganter, Flavien. 2023. "Identification of preferences in forced-choice conjoint experiments: Reassessing the quantity of interest." Political Analysis 31(1):98–112.

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2014. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." Political analysis 22(1):1–30.

Hanretty, Chris, Benjamin Lauderdale and Nick Vivyan. 2020. "A Choice-Based Measure of Issue Importance in the Electorate." American Journal of Political Science 64(3):519–535.

Hill, Seth J. 2022. Frustrated Majorities: How Issue Intensity Enables Smaller Groups of Voters to Get What They Want. Cambridge University Press.

Howe, Lauren and Jon Krosnick. 2017. "Attitude Strength." Annual Review of Psychology 68:327–351.

Krosnick, J. A. 1991. "Response strategies for coping with the cognitive demands of attitude measures in surveys." Applied Cognitive Psychology 5(3):213–236.

Krosnick, Jon A, David S Boninger, Yao C Chuang, Matthew K Berent and Catherine G Carnot. 1993. "Attitude strength: One construct or many related constructs?" Journal of personality and social psychology 65(6):1132.

Krosnick, Jon A and Richard E Petty. 1995. "Attitude strength: An overview." Attitude strength: Antecedents and consequences 1:1–24.

Krosnick, Jon and Robert Abelson. 1992. The Case for Measuring Attitude Strength in Surveys. In Questions about Questions: Inquiries into the Cognitive Bases of Surveys, ed. J.M. Tamur. Russell Sage Foundation pp. 177–203.

Lalley, Steven P and E Glen Weyl. 2018. Quadratic voting: How mechanism design can radicalize democracy. In AEA Papers and Proceedings. Vol. 108 pp. 33–37.

Malhotra, Neil, Jon A Krosnick and Randall K Thomas. 2009. "Optimal Design of Branching Questions to Measure Bipolar Constructs." Public Opinion Quarterly 73(2):304–324.

Margalit, Yotam. 2013. "Explaining social policy preferences: Evidence from the Great Recession." American Political Science Review 107(01):80–103.

Miller, Joanne and David Peterson. 2004. "Theoretical and empirical implications of attitude strength." The Journal of Politics 66(3):847–867.

Miller, Joanne M, Jon Krosnick and Leandre A Fabrigar. 2017. The origins of Policy Issue Salience. In Political psychology: New explorations, ed. Jon A Krosnick, I-Chant A Chiang and Tobias H Stark. Routledge pp. 125–171.

Posner, Eric and E Glen Weyl. 2018. Radical Markets: Uprooting Capitalism and Democracy for a Just Society. Princeton: Princeton University Press.

Quarfoot, David, Douglas von Kohorn, Kevin Slavin, Rory Sutherland, David Goldstein and Ellen Konar. 2017. "Quadratic Voting in the Wild: Real People, Real Votes." Public Choice 172(1-2):283–303.

Rabinowitz, George and Stuart Elaine Macdonald. 1989. "A directional theory of issue voting." American political science review 83(1):93–121.

Sauer, Carsten, Katrin Auspurg, Thomas Hinz and Stefan Liebig. 2011. "The Application of Factorial Surveys in General Population Samples: The Effects of Respondent Age and Education on Response Times and Response Consistency." Survey Research Methods 5:89–102.

Sears, D.O. and C.L. Funk. 1990. "The limited effect of economic self-interest on the political attitudes of the mass public." Journal of Behavioral Economics 19(3):247–271.

Sides, John, Chris Tausanovitch and Lynn Vavreck. 2022. The bitter end: The 2020 presidential campaign and the challenge to American democracy. Princeton University Press.

Tetlock, Philip E. 2017. Expert political judgment. Princeton University Press.

Zaller, John. 2012. "What Nature and Origins Leaves Out." Critical Review 24(4):569–642.

Zellner, Arnold. 1962. "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias." Journal of the American statistical Association 57(298):348–368.