"I can show what I really like.": Eliciting Preferences via Quadratic Voting

TI-CHUNG CHENG*, University of Illinois at Urbana-Champaign, USA
TIFFANY WENTING LI*, University of Illinois at Urbana-Champaign, USA
YI-HONG CHOU, The Chinese University of Hong Kong, Hong Kong SAR
KARRIE KARAHALIOS, University of Illinois at Urbana-Champaign, USA
HARI SUNDARAM, University of Illinois at Urbana-Champaign, USA

Surveys are a common instrument to gauge self-reported opinions from the crowd for scholars in the CSCW community, the social sciences, and many other research areas. Researchers often use surveys to prioritize a subset of given options when there are resource constraints. Over the past century, researchers have developed a wide range of surveying techniques, including one of the most popular instruments, the Likert ordinal scale [49], to elicit individual preferences. However, the challenge to elicit accurate and rich self-reported responses with surveys in a resource-constrained context still persists today. In this study, we examine Quadratic Voting (QV), a voting mechanism powered by the affordances of a modern computer and straddles ratings and rankings approaches [64], as an alternative online survey technique. We argue that QV could elicit more accurate self-reported responses compared to the Likert scale when the goal is to understand relative preferences under resource constraints. We conducted two randomized controlled experiments on Amazon Mechanical Turk, one in the context of public opinion polling and the other in a human-computer interaction user study. Based on our Bayesian analysis results, a QV survey with a sufficient amount of voice credits, aligned significantly closer to participants' incentive-compatible behaviors than a Likert scale survey, with a medium to high effect size. In addition, we extended QV's application scenario from typical public policy and education research to a problem setting familiar to the CSCW community: a prototypical HCI user study. Our experiment results, QV survey design, and QV interface serve as a stepping stone for CSCW researchers to further explore this surveying methodology in their studies and encourage decision-makers from other communities to consider QV as a promising alternative.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; **Collaborative and social computing design and evaluation methods**; *HCI design and evaluation methods.*

Additional Key Words and Phrases: Quadratic Voting; Likert scale; Empirical studies; Collective decision-making

*Both authors contributed equally to this research.

Authors' addresses: Ti-Chung Cheng, tcheng10@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; Tiffany Wenting Li, wenting7@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; Yi-Hong Chou, hank0982@link.cuhk.edu.hk, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR; Karrie Karahalios, kkarahal@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; Hari Sundaram, hs1@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

2573-0142/2021/4-ART182 $15.00
https://doi.org/10.1145/3449281

ACM Reference Format:
Ti-Chung Cheng, Tiffany Wenting Li, Yi-Hong Chou, Karrie Karahalios, and Hari Sundaram. 2021. "I can show what I really like.": Eliciting Preferences via Quadratic Voting. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 182 (April 2021), 43 pages. https://doi.org/10.1145/3449281

1 INTRODUCTION

Surveys are a common instrument to gauge opinion for scholars in the CSCW community, the social sciences, and many other research areas. The Likert ordinal scale [49] is a familiar instrument in surveys that elicit ratings, where survey takers indicate their level of agreement, satisfaction, or perceived importance along an intensity scale [58]. Nearly a hundred years after Likert [49], we find the ordinal scale in widespread use in online surveys, including those on Qualtrics, SurveyMonkey, and Google Forms platforms. Prior work indicates that self-reports by survey-takers, including responses recorded via an ordinal scale, can be inaccurate [4, 75]. Today, scholars utilize online surveys to reach a larger audience than paper-based surveys. However, the issue of inaccurate self-reports persists since many online incarnations of these ordinal scales are virtually indistinguishable from Likert scale surveys administered on paper-based forms nearly a century ago. Therefore, in this paper, we ask: *Do the affordances of modern computer interfaces, coupled with a modern computer's ability to make real-time calculations, allow us to develop novel, screen-based techniques to elicit more truthful responses from survey-takers?* By "truthful-responses," we mean that the survey instrument is incentive-compatible—truth-telling is a dominant strategy for the respondent. Throughout this paper, we shall use truthful-response, an informal phrase, instead of the more formal "incentive-compatible," in use in Game theory, to intuitively capture our intent.

We examine our motivating question within the context of surveys where the survey creator makes decisions by aggregating respondents' elicited opinions. At the same time, these decisions could potentially impact these survey respondents. Examples of such surveys include those that ask respondents about which policies the government ought to pursue [51], ask users for their opinions on different interface designs [47], and ask customers what products to stock at grocery stores [72]. The decision-maker (i.e., the group or the organization that uses the survey results) then uses the survey aggregated responses to implement one of the several options presented to the survey-takers. These decision-makers often cannot implement all the choices they gave to the survey-takers due to limited resources including money, time, space, and human resources. In this paper, we term such surveys, where respondents help the decision-maker with limited resources choose among a set of options, *resource-constrained surveys*. More specifically, we focus on resource-constrained surveys where respondents receive some utility from the survey outcome (e.g., the specific policy that the government adopts), even though the exact utility value or the time to receive that utility may remain uncertain. In the CSCW community, researchers often used Likert scale surveys to compare across multiple tools [78] and experimental conditions [12], and sometimes across different attributes of the same object [14, 52]. In such studies, researchers often aimed to choose the best or most critical options based on participants' preference; hence the surveys in these studies fit our definition of resource-constrained surveys. In contrast, surveys that, for example, aim to develop rankings by asking respondents to identify their favorite hobby, place to visit, or their favorite celebrity, are not survey types that we consider in this paper.

In resource-constrained surveys, decision-makers ask the survey respondents two questions: *how much* do you prefer each option; how would you *prioritize* (i.e., rank) the options? While in the past century, decision-makers have used ratings and rank-based survey instruments to answer such questions, the instruments answered *one* of the two questions, but not both [58]. Next, we examine two common survey techniques: one based on ratings and the other based on rankings.

In a rating-based survey, survey takers indicate their level of agreement, satisfaction, or perceived importance along an intensity scale [58]. The Likert scale is the most commonly used rating-based technique that elicits the *strength* of a survey taker's preferences [49]. Rating-based surveys are easy to understand and administer; this may explain why the Likert scale is the current norm in many survey settings. The Likert ordinal scale is problematic in the resource-constrained survey types that are of interest in this paper. Consider, for example, the Pew Research Center annual surveys with a 3-point Likert-like scale that ask how participants would budget federal government resources [51]. Respondents in such a survey can freely choose "increase spending" on the scale for *all* government program areas. However, a government with limited resources cannot operationalize this option. This behavior is known as "extreme responding" bias [5, 23, 56]. A Likert scale survey does not safeguard against exaggerated or unrealistic preferences [4, 75]. Furthermore, it may inaccurately elicit how participants *prioritize* the options since each option is independently assessed [2].

Ranking-based surveys focus on understanding how participants *prioritize* among the possible options [58]. For example, rank voting Benade et al. [6] is a popular technique that asks voters to prioritize a set of given options. Rank-based surveys force people to make trade-offs between options, which aligns with the goal of a resource-constrained survey. However, ranking-based surveys cannot elicit the strength of a participant's preferences, i.e., how much more a participant values option A over B. Besides, even though the rankings better demonstrated individuals' relative preferences, they are harder to be statistically analyzed and more cognitive demanding to participants [58].

This paper examines Quadratic Voting (QV) as a possible preference elicitation mechanism in online surveys. Posner and Weyl [64] recently advocated for the use of QV as an alternative for the traditional one-person-one-vote system. In the Posner and Weyl [64] formulation, each person has access to a fixed number of voice credits B (i.e., similar to the budget in participatory budgeting) to allocate across options. Then, each person can cast more than one vote (for, or against) for each option, with the proviso that the votes have a quadratic cost. Thus, casting n_k votes on option k would cost $c(n_k) \propto n_k^2$ in voice credits. Furthermore, the total cost in voice credits across all options cannot exceed the budget B. Therefore, a person casting positive or negative votes across k options has to satisfy $\sum_k n_k^2 \leqslant B$, where n_k is the number of votes cast on option k.

What makes QV of interest to the CSCW community that uses online surveying tools is that it appears to straddle the two familiar elicitation mechanisms of rating and ranking. While respondents can cast any number of voice credits modulo budgeting constraints on an option and thus indicating the strength of a preference, unlike Likert, they *cannot* allocate a high number of votes to *all* options. Notice that the quadratic vote costs imply that every additional vote on an option, say a change from n to $n+1$ votes, increases the marginal cost by $(n+1)^2 - n^2 = 2n+1$ voice credits. Due to the quadratic budget constraint, QV gently nudges the respondents to prioritize among the options. Prior work further showed that aggregated QV results of an option in a survey sample approximate a normal distribution in contrast to the "extreme responding" behavior found in Likert surveys [65]. Lalley and Weyl [45] also proved *robust optimality* properties of QV, suggesting that QV may be incentive-compatible (i.e., truth-telling is the dominant strategy for the respondent). Yet, to our best knowledge, this optimality assertion lacks empirical validation in the survey contexts of interest to this paper. Since QV requires real-time calculations to let the respondent know if they are allocating voice credits across options within the overall budget, QV requires a computational engine. The ubiquity of smartphones and tablets, the desire for the CSCW community to reach a wide audience through online surveys, and QV's robust optimality all suggest that QV is worth investigating as an alternative online survey instrument. Though there were many unanswered questions—from designing QV interfaces to communicating what QV is to users—we believe examining whether QV elicits more accurate responses from participants than Likert scale surveys is an important first step.

In our study, we examined QV's ability to elicit true (i.e., incentive-compatible) preferences in resource-constrained surveys. Since we were not aware of any prior work that empirically explored a way to determine the voice credit budget of QV, we assessed three different budgets in our QV survey experiment. Given that Likert scale surveys are the most widely adopted to date [58], we compared a QV survey to a Likert scale survey [49] in a resource-constrained setting. We identified two distinct types of survey questions in this setting: (1) choosing among K independent options of one topic, and (2) choosing among K dependent options that jointly contribute to the same topic. While the former is typical in public policy surveys, the latter is common to interface design surveys. Therefore, we studied the two types in two separate research questions:

RQ 1a How well do QV responses and Likert-scale responses align with the respondent's true preferences[1] in a survey where a survey respondent chooses among K independent options of one topic?

RQ 1b How do variations in the number of voice credits available to QV survey respondents—a small ($O(K)$), medium ($O(K^{1.5})$), and large ($O(K^2)$) budget—impact this outcome?

RQ 2 How well do QV responses and Likert-scale responses align with the respondent's true preferences in a survey where the survey respondent chooses among K dependent options that jointly contribute to the same topic?

To answer RQ1, we designed a public polling survey, similar to the annual Pew Research Center surveys, to assess which social causes need more support, a typical scenario for choosing among K independent options. To measure participants' true preferences on the same topic, we created an incentive-compatible donation task, in which participants should only donate more for cause A than cause B if they truly care more about cause A [11]. In a randomized controlled experiment, participants completed either a 5-point Likert scale version or a QV version of the survey and the donation task on the Amazon Mechanical Turk (MTurk) platform. We measured the similarity between each individual's survey result and their true preferences as reflected in the donation task and applied Bayesian analysis to compare the degree of similarity to true preferences in QV and Likert scale surveys.

In the case of choosing among K dependent options that jointly contributed to the same topic in RQ2, we used an interface design user study scenario to understand how users made trade-offs across visual and audio elements in a video streaming experience given limited internet bandwidth. We used QV and 5-point Likert scale surveys to obtain participants' self-reported responses and created an incentive-compatible product design and pricing task to elicit their willingness-to-pay for each video element [68]. Similar to the first experiment, we conducted the randomized controlled experiment on MTurk and analyzed the results using Bayesian analysis.

Contributions This work contributes to the extensive body of work on survey techniques that elicit self-reported preferences in resource-constrained decision-making in two ways: we find an improved accuracy of responses via QV compared to the Likert scale norm, and we extend the use of QV to a user study typical in the CSCW and HCI communities.

QV more accurately elicited true preferences: We empirically showed that QV with a medium (i.e., $O(K^{1.5})$) to large (i.e., $O(K^2)$) voice credit budget elicited true preferences more accurately than Likert scale responses when the survey respondents were asked to either (1) choose among K independent options of the same topic, or (2) choose among K dependent options jointly contributing to the same topic, under resource constraints. Unlike prior empirical work that compared the characteristics of QV and Likert scale responses [65, 59], we focused on accurate preference elicitation. Based on two carefully-designed randomized controlled

[1]That is, the phrase "align with true preferences" is equivalent to determining if the survey instrument is incentive-compatible.

experiments, our Bayesian analysis showed that QV responses aligned significantly closer to participants' responses in an incentive-compatible task than the 5-point Likert scale responses with a medium to high effect size (0.56 for RQ1 and 0.51 for RQ2). This finding is important because it shows that QV, a computational-powered alternative survey approach that combines ratings and rankings, can assist resource-constrained decision-making with more accurate self-reported responses.

Application of QV to HCI: We extend QV's application scenario from typical public policy and education research to a problem setting familiar to the CSCW community: a prototypical HCI user study. The limited prior empirical studies and applications of QV focused mostly on public policy [65, 1] and education research [59]. To the best of our knowledge, no prior HCI study has applied QV in an online survey. We designed an HCI user study with a research question that involved resource-constrained decisions. Compared to a Likert scale survey, a norm in the CSCW and HCI communities [47], we show that a QV survey more accurately elicited users' preferences on an HCI-related research question. Our experiment results, QV survey design, and QV interface serve as a stepping stone for HCI researchers in the CSCW community to further explore the use of this surveying method in their studies and encourage decision-makers from outside of CSCW to consider QV as a promising alternative.

While our study demonstrates the potential of QV as a computational tool to facilitate truthful preference elicitation, the goal of this research is *not* to convince decision-makers to replace their Likert scale surveys with QV surveys entirely. Indeed, in cases when the survey requires paper-based responses (e.g., when technological aids are unavailable), when the survey involves a list of unrelated options, or when the survey outcome is not resource-constrained, QV may not be an appropriate option. Instead, our goal with this paper is to show that QV may be a compelling alternative to the Likert scale when conducting online surveys under resource-constrained scenarios where it's beneficial to elicit both the respondents' ratings and rankings preferences.

2 RELATED WORK

2.1 Limitations in Likert scale surveys

The Likert scale is an intensity scale used to elicit participants' level of agreement, satisfaction, and perceived importance [49]. Invented by psychologist Rensis Likert in 1932, the initial design of the Likert scale aimed to identify clusters of opinions within a crowd [38]. The initial Likert scale survey was a 5-point scale[2], and subsequent researchers 3, 7, 11, or 12-point Likert scale variations [24, 22]. Some researchers developed alternative forms of Likert scale interfaces such as slider scales [67] or phrase completions [33] to improve usability of the traditional Likert scale survey.

Likert scale surveys have grown in popularity since they are easy to understand and administer across domains. However, as a ratings approach, Likert scale has its limitations. The primary limitation is not able to accurately understand how participants' prioritize a set of options due to several response biases, a phenomenon where survey respondents' stated opinions do not align with their true opinions.

Likert scale surveys suffer from "acquiescence bias," a type of response bias where participants tend to select the same level across the entire survey [2, 58]. To minimize this bias, researchers typically design the same ratio of positively and negatively framed options [44].

Another type of bias, "extreme responding," occurs when participants only answer on the extreme ends of the Likert scale [5, 23, 56], misaligned with their true preferences. An empirical study by

[2]The original design used: Strongly Approve (1), Approve (2), Undecided (3), Disapprove (4), Strongly Disapprove (5) as the five scales.

Quarfoot et al. [65] observed this phenomenon — participants either expressed polarized opinions or did not express an opinion at all, making it hard for survey creators to form an optimal strategy based on the survey responses [64]. Cavaillé et al. [10] provided a theoretical explanation for this bias; respondents exaggerate their opinions on purpose to influence the outcome of the survey. Researchers have developed statistical methods and suggested best practices for better question designs to mitigate such biases [27].

Although researchers have designed various solutions to alleviate these response biases stemmed from the *mechanism* of Likert scale, in this work, we examine QV as a potential alternative that is less prone to these response biases.

2.2 Quadratic Voting

Posner and Weyl [64] developed Quadratic Voting (QV), a collective decision-making mechanism [46] to circumvent the tyranny of the majority in traditional one-person-one-vote mechanisms, where the majority favors one option over the rest, always limiting the voice of the minority. Since QV participants are not bound to a single vote, QV does not have such a concern. Inspired by the Vickrey-Clarke-Groves mechanism [69], in QV, the marginal cost to cast an additional vote grows proportionally to the votes already cast on that option, inducing rational participants to vote proportionally to how much they care about an issue [64]. This design is why, unlike many traditional voting methods, each QV vote comes with a quadratic cost.

Here we formally define QV. Consider collecting responses from S participants, where each person has access to a fixed number of voice credits B to allocate across options. Then, each person can cast more than one vote (for, or against) for each option, with the proviso that the votes have a quadratic cost. Thus, casting n_k votes on option k would cost $c(n_k) \propto n_k^2$ in voice credits. Furthermore, the total cost in voice credits across all options cannot exceed the budget B. Therefore, a person casting positive or negative votes across k options has to satisfy $\sum_k n_k^2 \leqslant B$, where n_k is the number of votes cast on option k. At last, survey creators analyze the aggregated results by comparing the total number of votes from all participants across each option.

Several works explored the theoretical properties of QV. Lalley and Weyl [46] proved theoretically that QV's total welfare loss converges as the number of respondents increases. Similarly, a recent work by Eguia et al. [19] theoretically proved that the probability of QV aggregates reaching a socially efficient outcome converged to one as the number of participants increased in a resource-constrained survey. Lalley and Weyl [45] also proved *robust optimality* properties of QV, suggesting that QV may be incentive-compatible (i.e., truth-telling is the dominant strategy for the respondent). Our study examines if QV can elicit incentive-compatible results from an empirical lens instead of a theoretical lens. We next discuss empirical studies that compared QV with Likert scale.

2.3 Comparing QV with Likert Scale

Since QV is a relatively new voting mechanism, we are only aware of two studies that empirically compared QV with Likert scale. Both of these studies focused on comparing the characteristics of responses from QV and Likert scale surveys.

Quarfoot et al. [65] surveyed 4500 participants on their opinions for ten public policies; each participant completed either a Likert scale survey, a QV survey, or both. The study found that, for the same group of participants, responses on any option from the QV survey followed a normal distribution while those from the Likert scale survey were heavily skewed or polarized into "W-shaped" distributions. Researchers also noticed that individuals deliberated their responses more in QV and revealed more fine-grained attitudes. Thus, the study concluded that QV provided a clearer picture of the crowd's opinions to policy-makers on polarized issues [65].

Even though the study showed that the Likert scale survey and the QV survey produced *different* results, it did not compare the survey results to the participants' true preferences. Our study takes one step further and compares their degree of alignment with participants' true preferences. Besides, the study focused on controversial policies, such as "same-sex marriage", on which voters had a strong tendency to agree or disagree at the extremes. It's unclear how QV and Likert perform when survey options are less polarized, e.g., choosing one's favorite ice cream flavor. Our study analyzes this latter scenario.

In another empirical study, Naylor et al. [59] utilized QV for an educational research to understand students' opinions towards a list of factors that impacted their success at universities. Results showed that QV provided more insights than the Likert survey, such as distinguishing good-to-have factors from must-have ones. Again, our study differs from this work as we focus on comparing the accuracy of survey responses.

Overall, to the best of our knowledge, prior work has neither studied the degree of alignment between QV responses and participants' true preferences nor compared it with the degree of alignment between Likert responses and true preferences. Therefore, our work is the first to examine QV's ability to elicit true preferences.

3 METHODS – EXPERIMENT ONE: CHOOSING AMONG INDEPENDENT OPTIONS

We designed a between-subjects randomized controlled experiment to answer our first research question:

> RQ1a: How well do QV responses and Likert-scale responses align with the respondent's true preferences[3] in a survey where a survey respondent chooses among K independent options of one topic?
>
> RQ1b: How do variations in the number of voice credits available to QV survey respondents—a small ($O(K)$), medium ($O(K^{1.5})$), and large ($O(K^2)$) budget—impact this outcome?

The study was in the context of a public opinion polling to understand participants' preferences towards various societal causes, such as the environment, education, veterans. We focused on the topic of societal causes because public goods and resource allocation across causes is a problem relevant to every citizen. Since resources are limited in public sectors, this problem is a typical example of choosing among K independent options. Each participant completed one of the two kinds of surveys on the importance of the nine societal causes and a donation task. We detail the flow of our experiment in this section.

3.1 Participants Recruitment

We recruited participants located in the US from Amazon Mechanical Turk (MTurk) through the CloudResearch platform [50] in the 1st quarter of 2020. The MTurk population skews towards the younger and higher educated population [8], but an ideal surveying tool should be inclusive and serve people of all ages and education levels. Therefore, we did our best to align the participants' age and education level distribution to the latest available United States census estimates in 2018 [18]. We randomly assigned them into two groups: the Likert group and the QV group. Participants in the Likert group had a median completion time of 11.5 minutes and received $0.75, and those in the QV group received $2.5 due to a longer study length, with a median completion time of 20 minutes 56 seconds.

[3]That is, the phrase "align with true preferences" is equivalent to determining if the survey instrument is incentive-compatible.

3.2 Experimental Flow

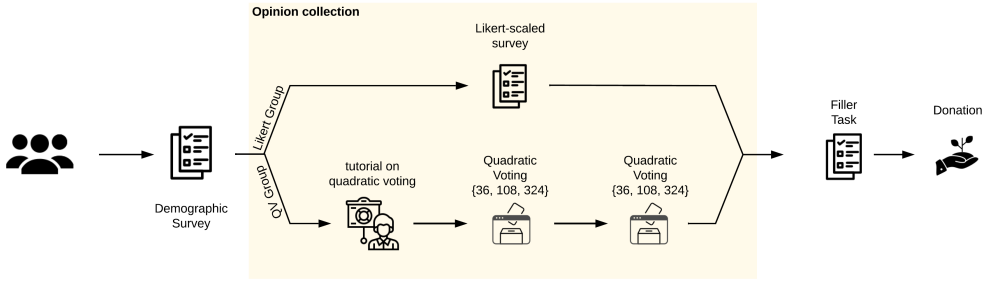


Fig. 1. Experiment one was a between-subjects experiment. We randomly assigned participants into two groups. Participants that took the upper path were in the Likert Group, and they expressed their attitudes toward various social causes through a five-point Likert scale. QV group participants reported their attitudes through two of the three variations of QV survey, with 36, 108, and 324 voice credits respectively.

Figure 1 summarizes the experimental procedure. The experiment consisted of four steps: 1) demographic survey, 2) Likert or QV survey, 3) a filler task, and 4) a donation task. We provide the complete experimental protocol in the supplementary materials.

Step 1: Demographic Survey. Participants joined the study under the impression that the goal of the study was to understand their opinions towards *the importance of various societal causes*. After signing the consent form, participants completed a demographic survey that asked for their gender, ethnicity, age, household income level, education level, and current occupation.

Step 2.1: Group 1 – Likert Scale Survey. The experiment randomly assigned some of the participants to the Likert group, shown in the upper path of Figure 1. In the prompt, we explicitly told the participants that there are limited resources in society, and that people have different preferences at allocating resources to various societal causes. The survey focused on nine societal issues, including 1) pets and animals, 2) arts, culture and humanities, 3) education, 4) environment, 5) health, 6) human services, 7) international causes, 8) faith and spiritual causes, and 9) veterans[4]. We derived the nine societal causes from the categorization of charity groups on Amazon Smile[5], a popular donation website that has accumulated over 100 million dollars of donations.

We asked the participants to rate each of the nine societal issues on a 5-point Likert scale: "For each of the issues listed below, how important do you think the issue is to you and that more effort and resources should be contributed towards the issue?". The 5-point Likert options ranged from "Very important" to "Very Unimportant." While there are a variety of Likert scales (3-point, 5-point, 7-point, and even 11-point), we used a 5-point Likert scale since it is one of the most commonly used scale [16].

Step 2.2: Group 2 – QV Survey The QV group took the lower path in Figure 1. We first asked participants to watch a pre-recorded tutorial video that introduced how QV works and how to use our QV interface since we did not expect participants to know about QV before taking part in the study, as opposed to Likert scale. Participants had unlimited time to interact with a demo QV interface to familiarize themselves with QV. To ensure that the participants paid attention

[4]For detailed definitions of each cause, please refer to Appendix A.1
[5]https://smile.amazon.com/

and understood QV, they needed to correctly answer at least three of the five multiple-choice quiz questions related to QV to continue with the study.

Once they passed the quiz, participants encountered two of the three versions of the QV surveys at random. The three versions of QV had 36, 108, and 324 voice credits, respectively. We showed them the same prompt as in the Likert group and instructed them to answer the same question for the nine identical causes, but with QV instead. Participants cast positive votes for causes they considered important and vice versa.

To our knowledge, no prior work discussed about how to decide on the voice credit budget in QV empirically. Therefore, we designed three versions of the QV survey to answer the second question in RQ1: how does the amount of voice credits in QV impact QV's ability to elicit true preferences? To examine how larger voice credit budgets impact people's choices, we set an exponential increase based on the number of options (K) on the survey ($O(K)$, $O(K^{1.5})$ to $O(K^2)$). We investigated three levels of voice credits: $K \times O$, $K^{1.5} \times O$, and $K^2 \times O$, where K is the number of options in the survey and O is the number of credits required to express an attitude in QV that is equivalent to the strongest attitude in a 5-point Likert scale survey, where "Neutral" in Likert = 0 vote in QV and one level in Likert = one vote in QV. In this experiment, $K = 9$ corresponded to the 9 societal causes. We used a 5-point Likert scale survey with extreme levels at ±2; hence each participant needed four voice credits ($2^2 = 4$) to express the extreme Likert levels in QV, which translated to $O = 4$. Thus, the three levels of voice credits in the experiment were $9 \times 4 = 36$ (QV36), $9^1.5 \times 4 = 108$ (QV108), and $9^2 \times 4 = 324$ (QV324). In all three cases, participants could afford to express any results from Likert in the form of QV. In addition, we asked participants to complete two of the three QV surveys to help understand how an increase or decrease in voice credits impacted their response behavior. To minimize the learning effects between the two versions, we provided participants a playground to get familiar with the QV interface and mechanism before taking the surveys. We randomized the order of the two versions.

Step 3: Filler Task After both groups of participants completed their surveys on the nine societal causes, as a filler task, they answered a free-form text question about their thoughts on another set of societal issues unrelated to those presented in the previous stage, such as increasing funding for Medicaid, strengthening gun control, and tighten social media regulation. For a complete list of causes, please refer to the supplementary material. Using a filler task in an experiment is common in psychology [37] and HCI experiments [31, 60] to prevent participants from inferring the experiment's purpose and form strategies when completing successive tasks. We designed this survey to prevent participants from directly connecting their survey responses with the upcoming donation tasks.

Step 4: Donation Task In the last step of the experiment, we need to design an incentive-compatible mechanism to elicit participants' true preferences towards the societal causes in the QV and Likert scale surveys. We designed a binding voluntary donation task with lottery-incentives, where their donation amount should reflect how much they truly care about the causes. In this task, to the best of their interest, they should donate more to organization A than organization B only if they care more about the cause of organization A.

We believe a binding out-of-pocket voluntary donation was a suitable task to estimate the participants' true preferences for two reasons. First, prior works frequently estimated true preferences with voluntary donations using actual binding financial consequences in either a lab or field setting [77, 26, 66, 7, 25]. They used participants' estimated true preferences based on actual voluntary donations as a reference to test the validity of contingent valuation methods on public or quasi-public goods. Xiao et al. [76] operationalized a voluntary donation with lottery-incentives in an MTurk setting. Second, voluntary donation, either online or offline, is an ecologically valid task in real-life settings. In most donation and crowdfunding websites, participants can navigate

across a wide range of options to donate. It does not require specialized knowledge and is simple to conduct online and at scale. We now describe how our donation mechanism worked.

We showed participants a list of nine charities in a randomized order with an introduction and an official website. We selected one charity for each of the nine societal causes in the QV and Likert scale surveys via Amazon Smile. But we chose not to show their mapping to the causes explicitly to the participants to maintain as much independence between the survey responses and the donation decisions as possible.

Participants had the chance to win $35 as a bonus through a lottery with an odds of 1 in 70. We asked them whether they would like to donate part of this bonus to any of the nine charity groups if they were the winner. They would keep the remaining amount after the donation to themselves. While participants could donate any amount to any organization as long as the total donation value did not exceed $35, they would want to maximize their bonus, which encouraged them to be truthful about the amount to donate. To increase participants' chance of donating a non-zero total amount [55], the research team promised to match $1 to each dollar they donated to an organization and execute the donation on their behalf if they won the lottery.

3.3 System Design

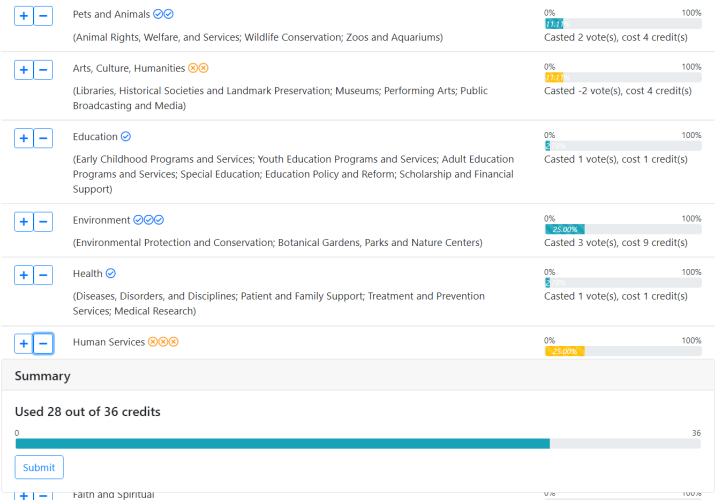


Fig. 2. Our QV interface design in the experiments. We omitted the prompt in this figure. After multiple design iterations from pretest and early pilots, the final interface allows participants to vote, with real-time feedback of how the credits are allocated. The progress bar design was inspired by the knapsack voting interface by [30].

We designed the QV interface through iterative-design (process detailed in Appendix D.2) with the goal to reduce participants' cognitive load with visual information [63]. Figure 2 shows the body section of the voting panel that contained a list of options to vote on. To the left of each option, participants voted using the plus and minus buttons. Buttons for an item were automatically disabled if the number of voice credits remaining did not permit an additional vote for that item. The number of blue check or yellow cross icons next to the item represented the number of "for" and "against" votes. We provided participants a bar with percentages to the right of each option, which showed the proportion of voice credits used for that option. In the summary panel, a progress

bar showed the number of voice credits the participants have and have not used out of the total budget. We floated the summary panel at the bottom of the page at anytime to ensure visibility.

For our experimental system, we used Angular.js for the front-end, Python Flask for the back-end implementation, and MongoDB Atlas for the database. The experimental system source code is publicly available[6], and so is the code for the standalone QV interface[7].

3.4 Analysis Method – Opinions Alignment Metric

We break our research question on "alignment" into two parts. First, we ask how similar these individual survey responses are to a participant's incentive-compatible behavior. Then we ask, how does QV and Likert compare overall in terms of the degree of alignment? To answer the first question, we need a metric for alignment.

We first clarify the definition of "alignment" in our analysis. A perfect alignment between a survey response and the same participant's donation choices requires an individual to express their preferences in survey with the same relative strength as their donation amount. More formally, it is defined as the following:

> A set of survey response $\vec{v} = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$, and a set of donation amount $\vec{d} = (d_1, d_2, \ldots, d_n) \in \mathbb{R}_+^n$, where n is the number of topics or options involved in the decision, are perfectly aligned if there exists a positive constant $k > 0$ that satisfies $k\vec{v} = \vec{d}$.

Notice now we represent the a participant's response in Likert scale survey, QV, and donation each with a vector of the same length. In addition, we focus the definition of alignment on the *relative* strength across opinions for two reasons. First, the results from our four types of surveys (Likert, QV36, QV108, QV324) and the donation task were not on the same scale. For example, the maximum possible number of votes on a topic in QV36 was 6, while the maximum donation amount on a topic was \$35. A relative scale maps each response onto the same space. The other reason is when any two participants donated different absolute amounts, possibly due to other factors such as income level or level of education, we want to capture how they *distributed* their total donation amounts. In other words, participants might donate with the *same* set of preferences across topics, but with *different* levels of total amount they were willing to donate. Hence, we decided to care only about the relative strength in opinions across topics.

Next, we need a metric that measures the degree of alignment. This metric needs to be monotonic with respect to the amount of discrepancy between two preference vectors in terms of the relative strength across preferences. In addition, this metric needs to be easily interpretable. Therefore, we decided to make use of the cosine similarity metric as our alignment metric and represent the difference between the survey results and the donation amount with an angle θ. It is formally defined as the following:

> The cosine similarity angle θ between a set of survey response $\vec{v} = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$, and a set of donation amount $\vec{d} = (d_1, d_2, \ldots, d_n) \in \mathbb{R}_+^n$, where n is the number of topics or options involved in the decision, is calculated via $\theta = \arccos\left(\frac{\langle \vec{v}, \vec{d} \rangle}{\|\vec{v}\| \|\vec{d}\|}\right)$, $\theta \in (0, \pi)$.

Cosine similarity is a commonly used similarity metric that measures the cosine of the angle between two non-zero vectors [73]. Instead of reporting a value between 0 and 2π radians, we report the angle in degrees, allowing for a more intuitive interpretation. Cosine similarity is monotonic with respect to the relative orientations of the two vectors, i.e., the relative strength in opinions,

[6] https://github.com/a2975667/QV-app
[7] https://github.com/hank0982/SimpleQV

and does not take into account the magnitude of the vectors, i.e., absolute vote or donation amount. Two sets of perfectly align opinions will yield a cosine similarity angle of zero, while two sets of completely opposite opinions will result in an angle of 180 degree.

For the Likert group, we map the ordinal responses into a vector where the result for each topic ranges from −2 to 2. For each of the three QV conditions, the vector contains the number of votes of the topics as is. Then, for each individual, we computed the cosine similarity angle between the Likert or QV vector and the absolute donation amount of the same individual.

Once we gathered these data, we moved on to the next step where we uncovered how the cosine similarity angles compared across the Likert group and the three QV variances (QV36, QV108 and QV324). We set up a Bayesian Model with these four sets of cosine similarity angle for each condition, as described in the next subsection.

3.5 Analysis Method – A Bayesian Approach

We used Bayesian analysis to compare if the distribution of cosine similarity angle in the QV group significantly differs from that in the Likert group. While Non-Bayesian inference techniques are widely available, and in hands of an experienced statistician, can dramatically reduce time to make an inference, we use Bayesian inference techniques for the following four reasons. First, as Kay et al. [39] point out Bayesian inference allows for accumulation of knowledge within the HCI community, where subsequent researchers can use prior outcomes as informative priors. Second, Bayesian models are transparent—researchers foreground all the assumptions in the model. There are no assumptions (e.g., Normality assumptions for the t-test) that need to be cross-validated with the data. The Bayesian model transparency avoids the intentionality pitfall in null hypothesis significance testing (NHST) [35, 43]—the choice of critical p-value is dependent on researchers' intentions, including sample size adjustments and pre-selection of hypotheses. Third, Kay et al. [39] suggest a that Bayesian formulation shifts the conversation focus from "did it work" to "how strong is the effect." A posterior probability distribution of estimated effect size in a Bayesian analysis plays a critical role when stakeholders perform a cost-benefit analysis on deciding which surveys to use [39]. While NHST can estimate the effect size (a point estimate) and a confidence interval, the NHST logic relegates this information as secondary to the p-value and under-emphasizes it. Finally, as McElreath [54] points out, since Bayesian priors use maximum entropy distributions (e.g., Normal, Gamma distributions), the inference is the *most conservative* given the evidence.

Now, we discuss our Bayesian formulation. There is one outcome variable: $\theta_{i|j}$, the cosine similarity angle between a survey response vector and a donation amount vector of each participant i under each of the four experimental conditions j: Likert, and three QV conditions (with 36, 108 and 324 credits). In summary, we aim to fit a distribution for the mean (i.e. the expected) angle between responses from each survey method and the donation amount. Then we compare how different the four distributions are.

In a Bayesian formulation, we need to define a likelihood function to model the cosine similarity under each condition. In general, this is a parametric formulation, and consistent with McElreath [54]. The likelihood function represents the modeler's view of the data, and not a claim about the world. The likelihood function is often parametric and we treat each model parameter as a random variable, drawn from a distribution (its prior). Typically, these priors are "weakly informative"—conservative priors which allow for all possible values of the parameter but chosen in a manner that promotes fast convergence.

We use a Student-t distribution to characterize the mean (i.e. the expected) angle in all four survey conditions (Likert and the three QV conditions). A Student-t, unlike a Normal distribution, is heavy-tailed, in the sense that the Student-t distribution doesn't fall off as quickly as does a Normal distribution and will thus be able to better account for outliers in the data [39]. The Student-t

distribution has three parameters: the degrees of freedom (ν), the experimental condition dependent mean (μ_j) and scale (σ_j). These parameters are random variables and we need to define their priors. Since our goal is to model the *average* angle, the fact that the Student-t is unbounded while the angle $\theta \in [0, \pi]$ is bounded is unimportant.

$$\theta_{i|j} \sim \text{Student} - \text{t}(\nu, \mu_j, \sigma_j), \quad \text{likelihood function to model donation} \tag{1}$$
$$\nu \sim 1 + exp(\lambda), \quad \text{degrees of freedom} \tag{2}$$
$$\mu_j \sim N(M_0, \sigma_0), \quad \text{modal angle in condition } j \tag{3}$$
$$\sigma_j \sim \Gamma(\alpha, \beta), \quad \text{scale parameter for condition } j \tag{4}$$

Equation (1) describes that the response $\theta_{i|j}$ of each group j is modeled as a Student − t distribution with mode μ_j, scale σ_j and with ν degrees of freedom. Next, we explain the model parameters.

Degrees of Freedom: We draw the degrees of freedom ν from a shifted exponential distribution, to ensure $\nu \geq 1$; $\nu = \infty$, corresponds to a Normal distribution assumption.

Modal contribution μ_j in each condition j: The mode μ_j corresponding to each group is drawn from a Normally distributed random variables with constant mean M_0 and variance σ_0.

Scale σ_j of each condition j: The scale σ_j of the likelihood function is drawn from a Gamma distribution $\Gamma(\alpha, \beta)$, with mode α and scale β; this prior on σ_j ensures that $\sigma_j > 0$.

Constants: The constants $M_0, \sigma_0, \alpha, \beta$ are set so that the priors are generous but weakly informative so that despite exploring all possible values, we ensure rapid MCMC convergence.

We performed the Bayesian analysis using PyMC3 [70], a popular Bayesian inference framework. We used one of the common computational techniques for Bayesian inference, Markov Chain Monte Carlo (MCMC), a stochastic sampling technique. It samples the posterior distribution $P(\theta|D)$, the distribution functions of the parameters in the likelihood function given the data observations D.

4 EXPERIMENT ONE RESULTS

In this section, we first describe the demographic distribution of the participants. Then, we present the descriptive statistics of the data in our first experiment. Lastly, we discuss results from our Bayesian analysis approach.

4.1 Participant Demographics

We collected 223 complete responses in the first experiment[8]. We removed four poor-quality responses where participants responded the qualitative question facetiously or they misunderstood the prompt. Among the 219 remaining responses, 56 completed the Likert path. Since the remaining participants each completed two of the three QV surveys with 36 credits (QV36), 108 credits (QV108) and 324 credits (QV324) in random, we collected 107 responses for QV36, 108 for QV108 and 111 for QV324. Since we have two experiment conditions for each QV version (choosing two of the three versions with randomized order), we collected more QV participants than the Likert group. As Bayesian analysis is less sensitive to the sample size, we didn't see an issue having differences in sample sizes.

We recruited the participants to match the demographic distribution in the US 2018 census estimates in terms of age and education level, as shown in Table 1. This reduced bias from the sample and allowed more balanced voices from different subgroups of the population, which is

[8]The experiment data that support the findings of this study are openly available in the Illinois Data Bank at: https://doi.org/10.13012/B2IDB-1928463_V1 [13]. The associated computation notebook are openly avaliable at: https://github.com/CrowdDynamicsLab/QV_True_Preference_Analysis.

Table 1. Experiment one sample demographics statistics align closely with 2018 US census across all groups and subgroup. Of a total of 219 experimental subjects, 56 subjects took the Likert scale survey, 107 subjects took the QV36 survey, 108 the QV108 survey and 111 subjects took the QV324 survey.

Demographics	Likert (%)	QV36 (%)	QV108 (%)	QV324 (%)	**All (%)**	Census (%)
Education						
No High School	16.1	14.0	13.9	14.4	**14.6**	10.2
High School	25.0	27.1	25.9	26.1	**26.0**	27.7
College Associate	28.6	26.2	34.3	34.2	**32.9**	33.1
Bachelor's Degree and above	30.4	32.7	34.3	34.2	**32.9**	33.1
Age						
18–24	21.4	15.9	14.8	15.3	**16.9**	13.7
25–39	26.8	29.9	30.6	29.7	**29.2**	30.7
40–54	25.0	29.9	28.7	29.7	**28.3**	28.3
55–69	26.8	24.3	25.9	25.2	**25.6**	27.3

generally hard to achieve in MTurk studies without specific control [17]. Having a representative sample is critical to ensure the generalizability of voting and survey tools.

Overall, 48.18% of the participants identified with male, 50.45% with female, 0.9% with non-binary, and 1 preferred not to disclose. As for the distribution of races, 77.27% of the participants were White, 11.82% were Asian, 9.09% were Black or African American, 1.36% were other races and one of them did not disclose the information. These distributions among each group closely resembled one another.

4.2 Descriptive Statistics

Since cosine similarity angle, our metric that measured the degree of alignment between survey responses and donation amount, could not take on all-zero vectors, we could only analyze participants who donated a non-zero amount. Therefore, we further filtered the dataset to keep only the responses that had a non-zero total donation amount. Across all the conditions, the average non-zero donation rate was 73.3%, consistent with the results provided by Fehr and Gintis [21] in 2007, which suggested that about 30% of the population would always free-ride in public goods provision regardless of what others do. The number of valid responses after dropping zero-donation participants for Likert, QV36, QV108, and QV324 were 44, 76, 76 and 84 respectively.

Overall, we found that the total amount of donation per participant was large enough to be distributed across charities in a way that could represent the full picture of their true underlying preferences for nine topics in most cases. Among those who donated a non-zero amount, about 60% of them donated part of the lottery winning amount ($5 - $20) but still kept a significant portion for themselves. Another 25% of the participants contributed a majority of the lottery reward ($\geq$ $33), if not the full amount. The total donation amount distribution across four surveying methods were relatively consistent, except that almost twice the proportion of participants in the Likert condition donated almost the full amount compared to the other QV conditions. We also confirmed that most participants donated to more than two charities and are consistent within all four experimental groups (Likert, QV36, QV108, QV324). For more details, please refer to Appendix B.1 and Appendix B.2.

Figure 3 shows the sample distributions of Likert responses across the nine societal causes. Distributions of most topics skewed towards positive opinions, with a median of either "Important"

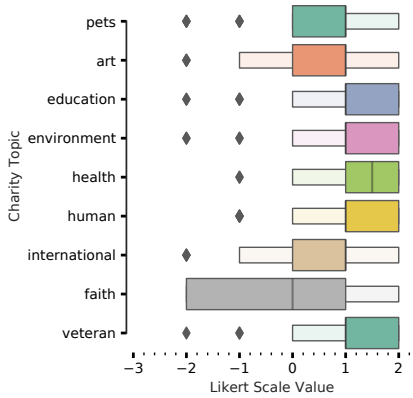


Fig. 3. Distribution of Likert scale responses per societal causes in Boxen plot. Each level from -2 to 2 corresponds to “Very unimportant”, “Unimportant”, “Neutral”, “Important”, and “Very important”. The distributions across all groups vary in their shapes, suggesting that participants showed their relative preferences even in the Likert group.

or "Very Important." Despite six out of nine topics had the same median of "Important," the shapes of their distributions were different, suggesting that participants expressed different levels of support based on their relative preferences.

Comparing across the three QV surveys, the response distributions were similar but had subtle differences (Figure 4). Most distributions of all the topics approximated a Normal distribution, consistent with prior work by Quarfoot et al. [65]. The medians of the distributions in QV36 varied less compared to those in QV108 and QV324. As the number of available voice credits increased, we found no decrease in the median of percentage budget usage – all around 98% (for details, refer to Figure 17 in the Appendix). This finding suggested that participants made effective use of the extra credits when completing QV with a large budget up to the order of N^2 (N is the number of options in a survey) with our QV interface despite more complicated calculations.

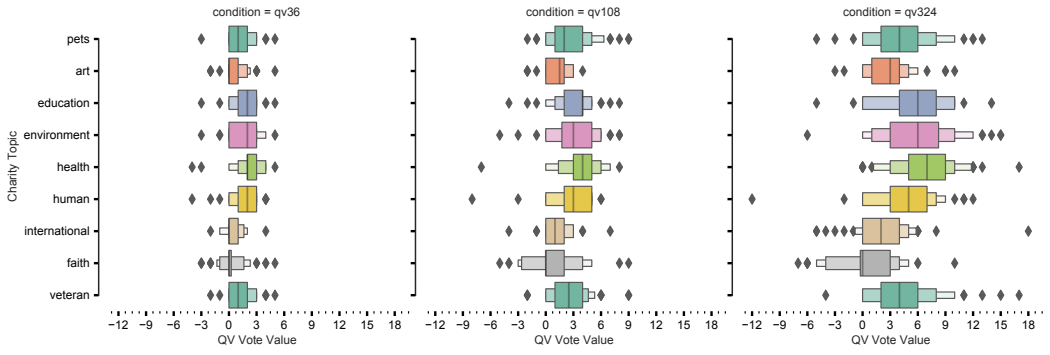


Fig. 4. Distribution of QV responses per societal causes in QV36, QV108 and QV324 in Boxen plots. The maximum possible number of votes on a topic was 6 votes in QV36, 10 votes in QV108, and 18 votes in QV324. Most distributions in all three QV set-ups follow a normal distribution. The medians of the distributions in QV36 varied less compared to those in QV108 and QV324. The tails stretch out longer as the number of total credits increases.

Since we are interested in the alignment between survey responses and donation behaviors, to get an intuitive view of how they correlate with each other, we visualize the correlation between the normalized survey responses (between -1 and 1) and the proportional donation amount of every participant in Figure 5. For all subplots, the points approximately followed a trend line with a positive slope, indicating a potential positive correlation between survey responses and donation behaviors. Within each topic, the slopes of the fitted lines in QV were more positive than that of Likert in most cases, suggesting a stronger correlation between QV survey responses and donation behaviors. To examine rigorously about how aligned the survey responses were with the donation behaviors in different conditions, we present the results from our Bayesian analysis.

4.3 Bayesian Analysis Results

Overall, we concluded from our analysis that survey responses from QV108, QV324 and averaged QV aligned significantly better with the donation results than Likert scale responses with a medium effect size (0.5 - 0.6).

Recall in Section 3.5, we estimated the posterior distributions of the mean (μ_{1-4}), standard deviation (σ_{1-4}), and degrees of freedom ν of the Student-t distributions that characterized the average cosine similarity angle for the four conditions, Likert, QV36, QV108, and QV324. Traceplots of the MCMC chains in Figure 7 show the results of MCMC estimation. The Gelman-Rubin statistic (a measure of MCMC convergence) $\hat{R}$ for all parameters was 1, indicating that the multiple sampling chains converged.

The first graph in the left column of Figure 7 shows that the mean cosine similarity angle of the four conditions varied. In QV108 and QV324 (the overlapping orange and green distributions on the left-most side), the modes of the mean cosine similarity angle (QV108: mode = 44.649 deg, QV324: mode = 44.796 deg) were smaller than those in the other two conditions, indicating a better alignment between the survey results and donation behavior in QV108 and QV324. The modal value of the average angle in QV36 (mode = 49.029 deg, the red distribution in the middle) was slightly higher than that of QV108 and QV324. Likert (the blue line) had the largest model value (52.857 deg) among all four conditions.

To contrast the mean cosine similarity angles between Likert and each of the the three QV budget cases, we constructed the distribution of the absolute difference between the means and the distribution of the corresponding effect size (normalized difference) as shown in Figure 6. The first three columns show the absolute difference (top row) and effect size (bottom row) between Likert and each of the three QV conditions. The fourth column compares the Likert condition with the averaged QV condition, by pooling together the responses from three QV conditions.

We now examine column 2 (Likert vs. QV108) in detail, and the rest of the columns follow a similar logic of interpretation. The second column shows the absolute contrast and its effect size of the mean cosine similarity angle between Likert and QV108. The absolute contrast equals to the angles in Likert group subtracted by the angles in QV108. In the top figure of the second column, the mode of the contrast is 8.2 with the 94% High Posterior Density (HPD) interval of [2.7, 14]. This means that the cosine similarity angles in the Likert group were most frequently 8.2 degrees larger than were the angles in the QV108 group, i.e., Likert condition responses were most frequently 8.2 degrees more misaligned with the donation behaviors than were the responses from QV108 condition. 94% of the difference in misalignment angles lie between [2.7 deg, 14 deg]. Since the HPD lies outside a significant ROPE (Region of Practical Equivalence) of 0 ± 1 deg, there was a significant difference between the alignment levels of the two survey methods. In addition, the bottom figure of the second column shows that the effect size has a modal value of 0.56, a medium-sized effect[9].

[9]We use the conventional standard that an effect of 0.2 is considered small, 0.5 is medium, and 0.8 is large.

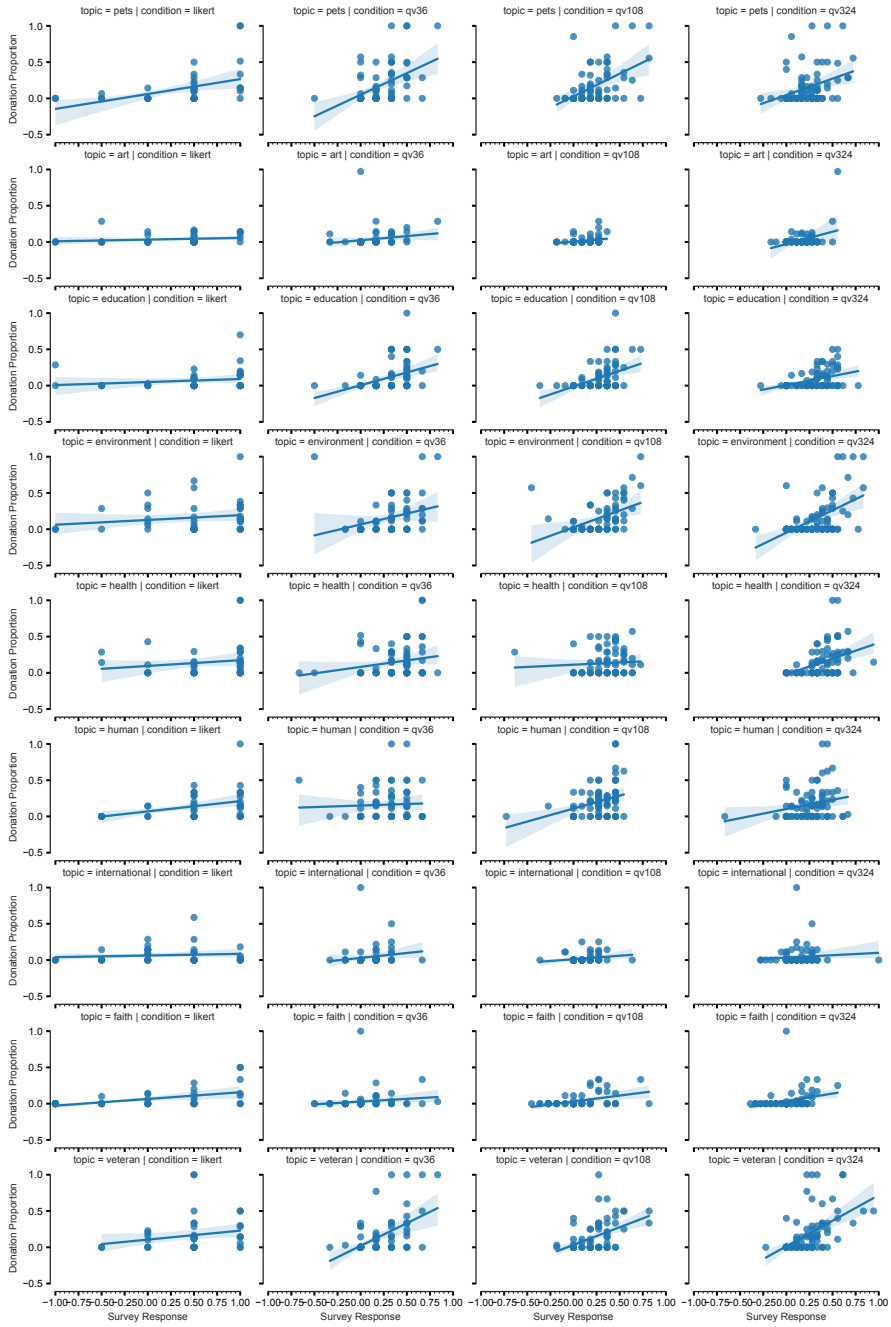


Fig. 5. Scatterplots showing the relationship between participants' normalized survey response and proportional donation amount for nine topics in all four conditions, Likert, QV36, QV108, and QV324. Each row is one topic, and each column is one survey condition. **The main finding** is: notice that for all subplots, the points showed a positive correlation between survey responses and donation behaviors. We also observed that slopes in the QV plots were more positive than that of Likert in most topics. This observation suggests a stronger correlation between QV survey responses and donation behaviors.

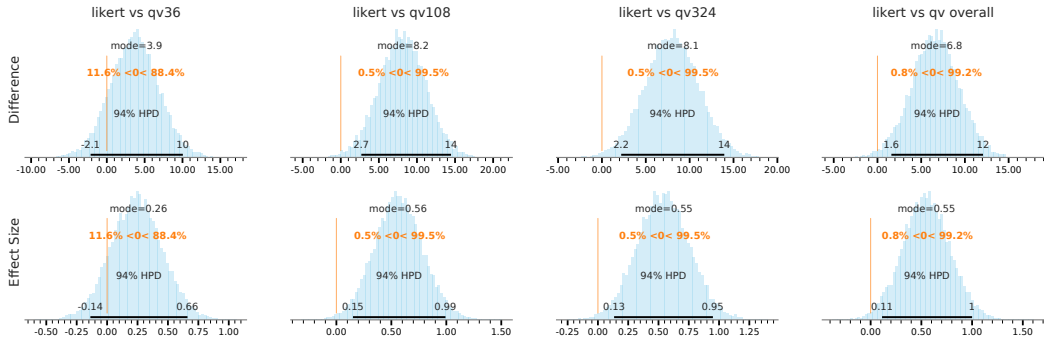


Fig. 6. The figure shows the contrasts distribution of the mean cosine similarity angles between the four experimental conditions. The four columns show the contrasts between the Likert and QV36, QV108, QV324 and the averaged QV condition respectively. The first row shows the absolute difference while the second row is about the effect size. Since we are highlighting contrasts, each sub-figure shows an orange vertical line located at 0. **The main finding** is: survey responses from QV108, QV324 and averaged QV aligned significantly better with the donation behaviors than did Likert response, with a medium effect size.

The HPD interval of the effect size is [0.15, 0.99], not overlapping with the ROPE of $[0 \pm 0.1]$, which consists of half of the small effect size of 0.2, indicating a statistically significant[10], medium to large effect size.

In the third column, we compare the mean cosine similarity angle of the Likert group against that of the QV324. The interpretation of this column is very similar to that of the second column above. The posterior of the effect size distribution has a mode of 8.1 with a 94% High Posterior Density (HPD) interval of [2.2, 14] for the contrast. Again, the HPD lies outside the ROPE of 0 ± 1 which implies that QV324 responses were better aligned to the donation behaviors than were the Likert responses. Similarly, the effect size has a mode of 0.55 with a 94% HPD interval of [0.13, 0.95], indicating a significant, medium to large-sized effect.

The first column shows a different result. It compares the degree of misalignment of the Likert group against that of QV36. The mode of the contrast is 3.9, with a 94% High Posterior Density (HPD) interval of $[-2.1, 10]$. While the mode is positive, the HPD interval overlaps with a ROPE of 0 ± 1, implying that the observed differences are not significant. The corresponding effect size shows a mode of 0.26 and a HPD interval of $[-0.14, 0.66]$. About 11.6% of the HPD is to the left of zero, also indicating an insignificant effect size. The result suggests that QV survey with only 36 voice credits did not outperform Likert scale survey significantly in terms of alignment with the donation behaviors.

At last, the fourth and final column compares the degree of misalignment with donation behaviors between the Likert group and the averaged QV condition, by pooling three groups of QV together. We see a mode of 6.8 with an High Posterior Density (HPD) interval of [1.6, 12]. Following the same interpretation logic, the pooled QV condition aligned significantly closer to the donation behaviors than Likert condition. The modal effect size is 0.55, with an HPD interval of [0.11, 1], indicating a significant medium to large effect size.

[10]We use the phrase "statistically significant" not in the traditional non-Bayesian sense, but to imply that the posterior distribution HPD was outside the ROPE, a significant result in Bayesian data analysis.

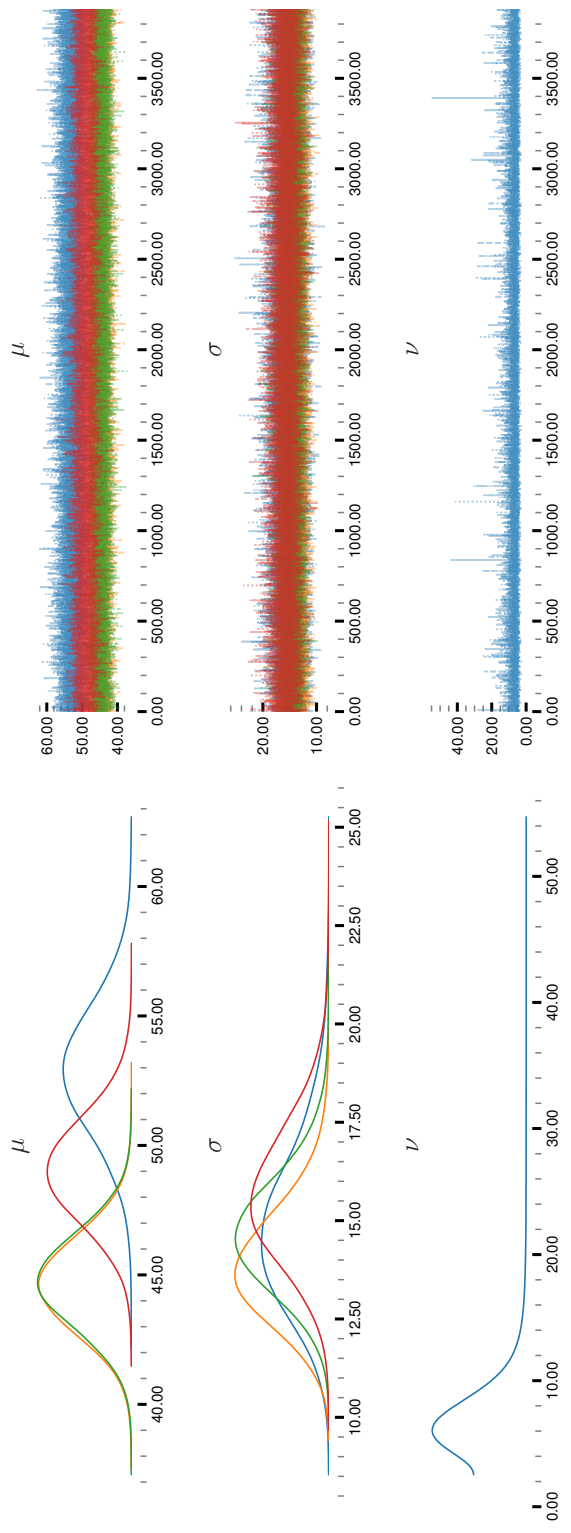


Fig. 7. Traceplot showing the results of the MCMC estimation in experiment one. The left column is the posterior distributions for μ_{1-4}, σ_{1-4}, and ν of the Student-t distribution. The right column shows the corresponding sampling traces. The color mappings are: red (QV36), orange (QV108), green (QV324), blue (Likert). Notice that QV108 and QV324 have similar distributions for μ_{1-4} and are to the left of QV36 and Likert, meaning a better alignment with donation behaviors. Likert shows the worst alignment: as a reminder, the ideal alignment angle should be 0^o. Note also, that the modal value for degrees of freedom $\nu \approx 7$, confirming our choice of the Student-t distribution instead of the Normal distribution, which usually requires $\nu \geq 30$. Furthermore, the Gelman-Rubin statistic $\hat{R}$ for all parameters was 1, indicating convergence of the sampling MCMC chains.

5 METHODS – EXPERIMENT TWO: IMPORTANCE OF VIDEO ELEMENTS

The first experiment answered RQ1 and showed that QV aligned closer to the participants' true preferences compared to Likert scale when choosing among K independent options of the same subject matter. To strengthen our results and examine the generalizability of QV, we designed a second experiment to answer our second research question:

> RQ2: How well do QV responses and Likert-scale responses align with the respondent's true preferences in a survey where the survey respondent chooses among K dependent options that jointly contribute to the same topic?

Since relationships and interactions among ballot options may impact how users make trade-offs, we want to examine if the results from QV also align better with people's true preferences than does the Likert scale in RQ2.

We hypothesized that QV would outperform Likert in accurately eliciting participants' true preferences in the new setting. We changed the application domain from public policy to HCI , an important research area in the CSCW community, since HCI user studies often rely on eliciting preferences from users to inform designs that involve trade-offs and in turn create better user experiences. These characteristics made HCI studies an excellent domain to test out QV. To verify our hypothesis, we designed a within-subjects study that elicited participants' preferences on different video elements using QV, Likert, and a pricing task.

In this section, we first explain how we selected video streaming experience as the HCI study scenario. Then, we demonstrate the experiment workflow accompanied by the interfaces of the experiment. Finally, we explained the analysis approach.

5.1 Choice of HCI study

We set out to find a typical HCI study where UX/UI researchers survey users to understand what features to prioritize. In the end, we decided to use the research scenario of understanding users' preferences among various video and audio elements under network or monetary constraints [57, 62].

Research on video and audio elements of video playback from the lens of HCI is relatively mature. Understanding how users with bandwidth constraints made trade-offs across multiple videos and audio elements [57, 62] is a typical example for which of the K dependent options of the same topic would people prioritize under constraints. Oeldorf-Hirsch et al. [62], for example, conducted a study to examine the differences in participants' perceptions between three video bit rates, three video frame rates, and two audio sampling rates across three types of video content via a 5-point Likert scale.

We proposed a similar user research topic as in Oeldorf-Hirsch et al. [62]'s study. We designed the experiment to answer the following question: "Given a video with unsatisfying quality, under limited bandwidth, how should the bandwidth be allocated to enhance the five visual and audio elements, including motion smoothness [34], audio stability [32], audio quality [41], video resolution [40], and audio-video synchronization [74], to obtain an acceptable video streaming experience from the viewers' perspective?" We selected the five video elements based on prior work. For their detailed definitions, please refer to Appendix C.2. To our knowledge, no prior work has studies the combination of five elements in a single experiment.

Prior work suggested that the type of video affected users' perceptions for video elements. In this experiment, we used a 90-second weather forecast video for the United States. We chose a weather forecast video for two reasons. First, the concept of a weather forecast is generic and universal. The terms used in the weather forecast are usually easy to understand. Second, since we are studying both audio and visual elements, we wanted a video that conveyed information via both visuals

and speech. In a weather forecast video, the meteorologist usually spoke aloud the weather while pointing at visual cues such as icons and numbers.

In the next section, we describe how we conducted the video elements trade-off experiment to compare QV and Likert scale's ability in truthfully reflecting users' preferences across the video elements.

5.2 Experimental Flow

We recruited participants on Mechanical Turk through the CloudResearch platform [50]. Like our first experiment, we used a pre-survey to match the participants' distribution with the U.S. population in age and education based on 2018 US census estimates. The average completion time was 35 minutes 42 seconds and all participants received $6 as base pay and a bonus up to $2. All participants followed six steps. The six steps were (1) demographic survey, (2) tutorials and attention checks, (3) a video playground, (4) Likert and QV surveys, (5) a filler task, and (6) a design task. Now we explain the six steps in detail. We include the experiment flow diagram in the appendix and present the experiment protocol as supplementary materials.

Step 1. Demographic Survey We greeted participants with a consent form. In the consent form, we presented the goal of the study as understanding how people think about the importance of the different elements during video streaming. We did not reveal to the participants that this experiment aimed to compare Likert and QV until they completed the survey. Once participants gave their consent, they would fill out a demographic survey that contained questions identical to those in the first experiment.

Step 2. Tutorials and Attention Checks In step two, we provided two tutorials to the participants. All participants would first read through a tutorial that defined the five video elements used in the experiment, via textual explanation and pairs of video examples. On the next page, they needed to answer five multiple-choice questions about the definitions of video elements and two attention checks designed to see if audio and video played fine on the participant's device. Participants qualified for continuing the experiment only if they answered fewer than two questions incorrectly. This step made sure participants fully understood the terminologies used in the rest of the experiment.

Participants then moved on to a tutorial on how QV works, with a short instructional video supplemented with text. They had a chance to play with a QV interface similar to that in Figure 2. Once the participants were confident that they understood QV, they needed to complete the same QV quiz in experiment one. The system disqualified a participant immediately if they answered two or more questions wrong.

Step 3. Video Playground To increase this experiment's fidelity, we framed the study as a market research initiative by a fictitious company (participants were not aware that the company was fictitious), with a goal to provide a video-streaming product in cars using satellite-based Internet. We first showed the participants the "current prototype" of the company's streaming service, which simulated what a weather forecast video played under limited bandwidth would look like, with all five elements at the worst quality in the range we designed to study[11].

To help participants better understand the impact of various enhancement levels for each element on the current prototype, we led them to a video playground shown in Figure 8. This playground allowed participants to use the control panel to adjust the levels of enhancements for all five video elements and see real-time changes in the video on the top of the page, i.e., the video played in the most recent combination of quality levels. Participants could pause and play the video at any time,

[11]Motion smoothness: 2.5 fps; Audio stability: 20% packet loss rate; Video resolution: 120x90 at 32 kbit/s; Audio quality: 8kHz sampling rate; Audio-video synchronization: visuals play 2000ms ahead of the audio

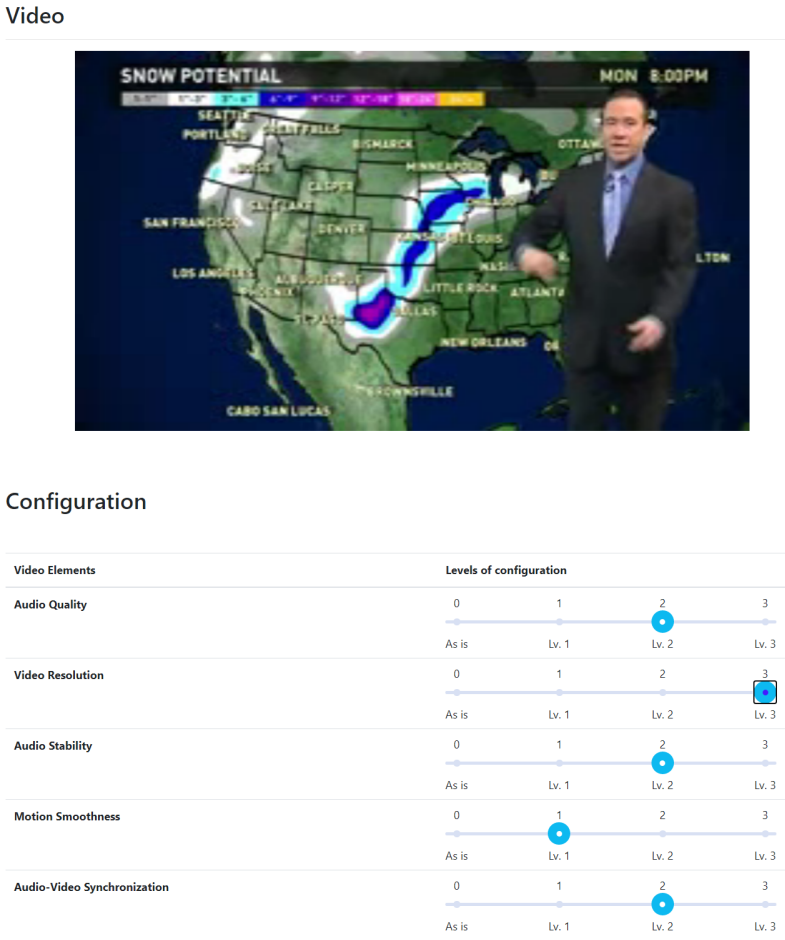


Fig. 8. The real-time video element interface allows participants to adjust video playback elements and understand how differences in the elements impact their viewing experience. We selected four levels of quality settings for each element according to prior research, ranging from an unacceptable quality to a good quality. The technical details of this implementation are described in appendix D.1.

and replay the video as many times as they choose. We encouraged participants to test out different combinations freely in this playground to help them understand the impact of each element and how the five elements interact. We asked participants to describe how changing the elements impacted their experience in a free-form text question to make sure participants did experience different settings.

For each element, we provided a slider with four levels, Level 0 at the lowest quality and Level 4 at the highest. We designed the intermediate levels based on prior research such that the changes between each level of an element had a quasi-linear impact on viewers' perception. The four levels of the five video elements are listed below, from Level 0 to Level 3:

- Motion Smoothness [34]: 2.5 fps, 6.25 fps, 8 fps, and 25 fps
- Audio Stability [32]: 20%, 10%, 5%, and 0% probability in packet loss

- Video Resolution [40]: 120x90 at 32 kbit/s, 168x126 at 64 kbit/s, 240x180 at 96 kbit/s, and 240x180 at 224 kbit/s; encoded in the WMV2 (Windows Media Video 8) codec
- Audio Quality [41, 61]: 8kHz, 16kHz, 24kHz, 32kHz; encoded using the AAC (Advanced Audio Coding) codec
- Audio-Video Synchronization [74]: 2000ms, 1250ms, 750ms, 0ms; visual ahead of audio

Step 4. Surveying Preferences After experiencing how different enhancements on the five video elements impacted their viewing experience, we collected participants' opinions on how critical the improvements on each video element in the current prototype were for them to understand the weather forecast video. Participants completed a QV survey and a Likert scale survey in a randomized order to prevent ordering effect. Since this was a within-subjects study, we randomized the display order of the elements on the surveys to minimize carryover effect and ordering effect. The QV interface was similar to that in experiment one. We designed the credit budget to be 100 voice credits, the best option based on experiment one, i.e., $K^2 \times O$, where $K = 5$ and $O = 4$ in this case.

Step 5. Filler Task Before moving on to the task for eliciting the participant's true preferences, we designed a survey as the filler task that asked participants' about their consumption preferences for subscription tiers across several well-known streaming services, telling them that the survey results would be an important part of the next task. This survey aimed to prevent participants from translating their survey responses to the video elements survey directly to the next task.

Step 6. Design a Product To capture a participant's true preferences on how much they *value* each of the five video elements, we created an incentive-compatible product design task to elicit their true willingness-to-pay for the elements. To create a high fidelity scenario, we told participants that the system assigned them into the designer group, and their job was to design a streaming service for cars via satellite internet. They should design and price their product such that another participant from the buyer group (fictitious, but the participants were not aware), who we claimed to have matched for them based on demographic information and the consumption preference survey, would be willing to purchase the product. We emphasized that their product should be affordable and should allow the buyer to understand the weather forecast video. To incentivize the participants, we offered them 10% of the final price they proposed as their bonus if the buyer decided to purchase their product.

We guided participants to complete the task in two steps. In the first step, participants needed to select one of the two qualities for each video element, a lower quality and a higher quality, and "assemble" the product. Quality one refers to level 0 in the playground, the worst one they had experienced. For quality two, we selected the quality levels in the playground that matched what prior research showed to be the "acceptable level"[12]. Participants saw real-time changes to the video as they updated the qualities. Participants were essentially making a binary decision of whether a video element was "important" or "not important" for them, and equivalently for the buyer that had similar preferences as them.

In the second step, participants set a price they thought to be reasonable for each of the five elements between \$0 - \$4 based on the qualities they selected and how much they thought the buyer would value them. The sum of these prices was the total price of the product. We reminded participants throughout the task that the higher they priced the elements, the more their bonus could be. However, if they overpriced any element from the buyer's perspective, the buyer might not purchase their design and they would fail to gain any bonus.

[12]Motion smoothness: 6.25 fps; Audio stability: 10% packet loss rate; Video resolution: 240x180 at 96 kbit/s; Audio quality: 16kHz sampling rate; Audio-video synchronization: visuals play 750ms ahead of the audio

The goal of this design was to elicit the truthful willingness-to-pay for each participant as the their true preferences. Our set-up was incentive-compatible and the best strategy for the participants was to price based on how they themselves valued each of the elements. If they priced the product higher than their accurate valuations, the "buyer" with similar demographic and consumption preferences as them would reject their proposal, given that the buyer would likely value the elements the same way. Vice versa, if the participants set their prices to be lower than their actual willingness-to-pay, they would lose out on earning a higher bonus.

5.3 System Design

We reused the QV interface from experiment one in our second experiment. To create real-time adjustments in video qualities, we pre-generated video-only and audio-only files of different qualities. When a participant changed an audio or video quality setting, the system served the correct combination of video and audio files. We used JavaScript to adjust the video-audio synchronization in the front-end at real-time.

This design balanced the need for a high network speed to stream every configuration from the server and the need for a powerful client to compute the video and audio qualities. The experiment source code for experiment two is publicly available[13]. More details of the system implementation are provided in Appendix D.1.

5.4 Analysis Method

Since RQ2 is also about the degree of alignment, similar to RQ1, we followed a similar analysis approach as described in Section 3. In experiment two, we compared the alignment between survey responses with the prices set in an incentive-compatible scenario. We used the same definition of "alignment" and the same metric for alignment, cosine similarity angle, as explained in Section 3.4. For the Likert group, we mapped the ordinal responses into a vector where the result for each video element ranges from -2 to 2. For QV, the vector contains the number of votes for each video element. Then, we computed the cosine similarity angle between a Likert or QV vector and the absolute prices set by the same participant,.

Once we obtained the two sets of cosine similarity angles, one for Likert and one for QV, we applied the same Bayesian formulation detailed in section 3.5 to model the mean cosine similarity angle in both groups. Then, we compared the two distributions to see if they how far they are apart.

6 EXPERIMENT TWO RESULTS

In this section, we first describe the participants' demographic in the second experiment. Then, we present descriptive statistics for the raw data. Lastly, we discuss results from our Bayesian analysis.

6.1 Participant Demographics

We collected 101 complete responses in the second experiment[14]. We removed 8 poor-quality responses where participants responded to the qualitative questions facetiously and put the same survey rating across all video elements. Among the remaining 93 participants, 55 of them identified as male and 38 as female. Around 80% of the participants were White, 13% were Black or African American, 5% were Asian, and the remaining 2% preferred not to disclose their racial information.

[13]https://github.com/a2975667/QV-buyback

[14]The experiment data that support the findings of this study are openly available in the Illinois Data Bank at: https://doi.org/10.13012/B2IDB-1928463_V1 [13]. The associated computation notebook for data analysis are openly avaliable at: https://github.com/CrowdDynamicsLab/QV_True_Preference_Analysis.

Table 2. experiment two sample demographics statistics align closely with 2018 US census.

Demographics	**Sample (%)**	Census (%)
Education		
No High School	**4.30**	10.2
High School	**28.0**	27.7
College Associate	**31.2**	33.1
Bachelor's Degree and above	**36.6**	33.1
Age		
18–24	**9.68**	13.7
25–39	**39.8**	30.7
40–54	**28.0**	28.3
55–69	**22.6**	27.3

Similar to experiment one, we aligned the participants' age and education level distribution to match that in the US 2018 census estimates [18], as shown in Table 2.

6.2 Descriptive Statistics

As shown in Figure 9 and Figure 10, on an aggregated level, participants expressed similar relative preferences across the five video elements in both Likert and QV. Audio quality ranked the highest in both cases, while motion smoothness and audio-video synchronization had the lowest ranks. While the aggregated preference were similar between Likert and QV, we will discuss their differences on an individual participant level in the next subsection. Similar to experiment one, a majority of the Likert response distributions skewed to the left; in contrast, QV response distributions approximated a Normal distribution. With a budget of 100 voice credits in QV, 57% of the participants used at least 98% of the budget and 82.8% of them used over 90% of the budget, suggesting that most participants actively made use of the majority of the budget.

Set prices for the five video elements exhibited similar patterns as the survey responses on an aggregated level (Figure 11). Compared to the survey response distributions, elements with a higher average rating of importance in surveys also had higher average prices during product design. Since we instructed participants during price-setting that the buyer would be willing to pay more for the element that they value more, the alignment of preferences on an aggregated level between surveys and prices indicated that the participants kept our instruction in mind when they decided on the prices.

Similar to experiment one, we are interested in the degree of alignment between survey responses and prices set in an incentive-compatible scenario on an individual level. Figure 12 visualizes the estimated correlation between normalized Likert or QV responses and normalized prices. All elements had a trend line with a positive slope, suggesting potential positive correlations. QV responses seemed to have a more positive slope with prices compared to Likert responses, indicating a possibly better alignment. Next, we present statistical support for this phenomenon.

6.3 Bayesian Analysis Results

Overall, our Bayesian analysis for experiment two showed that QV survey responses aligned significantly better with the price setting behaviors than Likert responses with a medium to high effect size (0.5–0.6). The first graph in the left column of Figure 14, which contains traceplots for

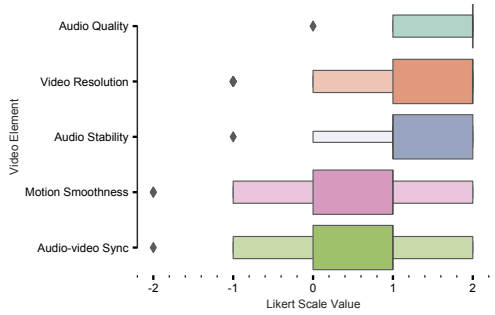


Fig. 9. Distribution of Likert scale survey responses per video elements in Boxen plot. Each level from -2 to 2 corresponds to "Very unimportant", "Unimportant", "Neutral", "Important", and "Very important". The distributions of different elements vary in their shapes, suggesting that participants showed their relative preferences even in the Likert group.

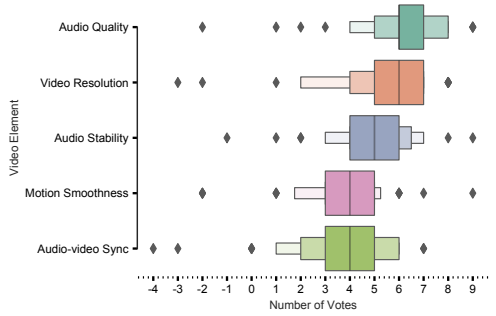


Fig. 10. Distribution of QV responses per video element in QV in Boxen plots. The maximum possible number of votes on an element was 10 votes given 100 voice credits. Most distributions in all three QV set-ups follow a normal distribution.

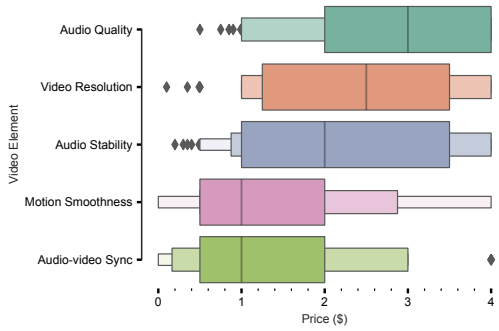


Fig. 11. Distribution of prices set by participants per video elements in Boxen plot. Prices ranged from $0 to $4. Compared to the survey response distributions, elements with a higher rating of importance in surveys also had higher prices during product design.

the MCMC estimations, shows that the distribution of the mean cosine similarity angle for QV (orange line) is to the left of that for Likert (blue line). Since a perfect alignment means a zero angle, QV had better alignment with the set prices relative to Likert.

To confirm if the difference was statistical significant, we constructed the distribution of the absolute difference between the means and the distribution of the corresponding effect size (normalized difference), as shown in Figure 13. In the subfigure on the left, the mode of the contrast is 5.8, meaning that the cosine similarity angles in the Likert group were most frequently 5.8 degrees larger than the angles in the QV group. Since the HPD of [2, 9.7] lies outside a significant ROPE (Region of Practical Equivalence) of 0 ± 1 deg, there was a significant difference between the alignment levels of the two survey methods. The medium to high effect size is significant based on the figure on the right, with a modal value of 0.51 and a HPD interval of [0.16, 0.84].

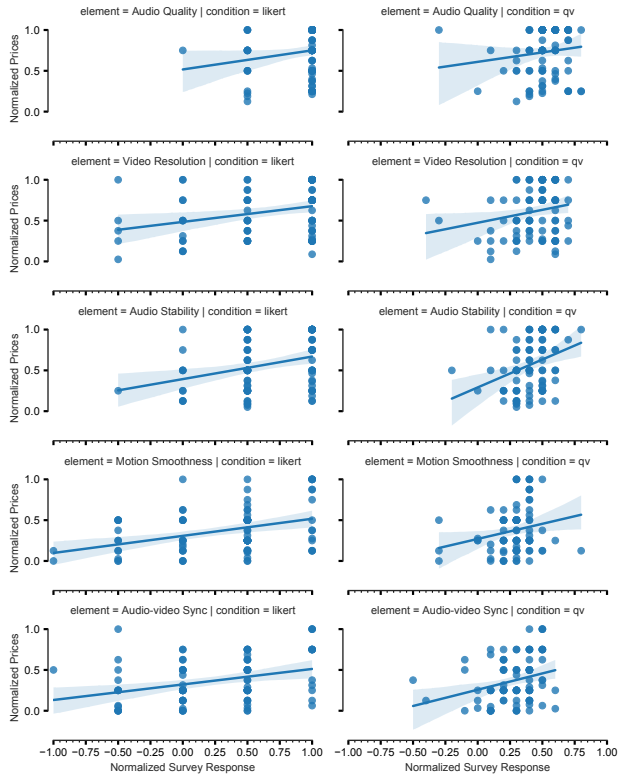


Fig. 12. Scatterplots showing the correlation between participants’ normalized survey response and normalized prices for five video elements in the Likert survey and QV survey. Each row is one topic, and each column is one survey condition. **The main finding** is: all elements had a trend line with a positive slope, indicating potential positive correlations. QV responses seemed to have a more positive slope with prices compared to Likert responses, indicating a possibly better alignment.

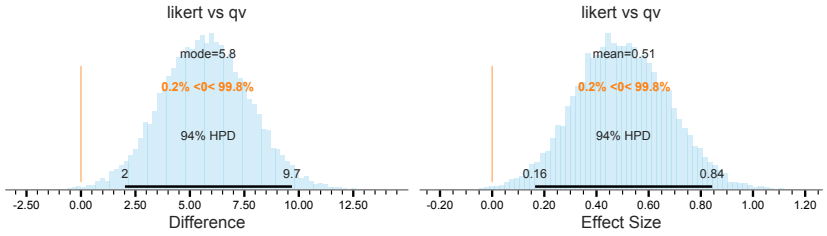


Fig. 13. The figure shows the contrasts distribution of the mean cosine similarity angles between the Likert group and QV group. The subgraph on the left shows the absolute difference while the one on the right is about the effect size. Since we are highlighting contrasts, each sub-figure shows an orange vertical line located at 0. **The main finding is:** survey responses from QV aligned significantly better with the price setting behavior than Likert scale responses with a medium effect size.

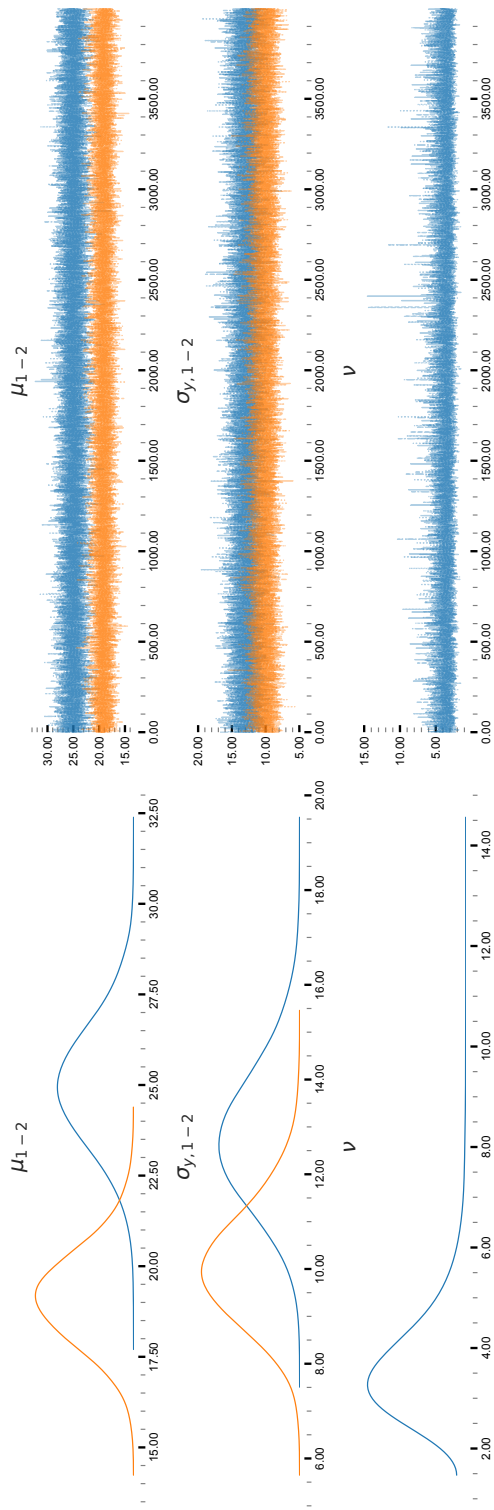


Fig. 14. Traceplot showing the results of the MCMC estimation in experiment two. The left column is the posterior distributions for μ_{1-2}, σ_{1-2}, and ν of the Student-t distribution. The right column shows the corresponding sampling traces. The color mappings are: orange – QV, blue – Likert. Distribution for μ_{1-2} of QV is to the left of Likert, meaning a better alignment with price setting behaviors. Note also, that the modal value for degrees of freedom $\nu \approx 3$, confirming our choice of the Student-t distribution instead of the Normal distribution, which usually requires $\nu \geq 30$. Furthermore, the Gelman-Rubin statistic $\hat{R}$ for all parameters was 1, indicating convergence of the sampling MCMC chains.

7 DISCUSSION

In this section, we discuss the implications of the two experiment results. We first discuss when and why QV aligns better with true preferences and then discuss the impact of the QV voice credit budget. We conclude the section with a discussion around when to use QV.

7.1 When and Why Does QV Align Better with True Preferences?

In both RQ1 and RQ2, we ask how QV survey responses align with people's true preferences compared to Likert scale survey responses in resource-constrained collective decision-making. From the two experiments, we showed that QV aligned significantly better with people's true preferences than a 5-point Likert scale survey when the survey respondents were asked to either (1) choose among K independent options of the same topic or (2) choose among K dependent options that jointly contributed to the same topic. Note that the above conclusion was only true when the QV survey provided a medium ($O(K^{1.5})$) or large ($O(K^2)$) voice credit budget to participants. Our Bayesian analysis showed a medium to high effect size for the difference between the degree of alignment in the QV group and the Likert group in both experiments. This suggests that QV had the potential to elicit true preferences more accurately in collective decision-making. We now discuss two potential reasons that may explain our results.

Costs and scarcity One explanation for QV's better alignment in the above conditions may be the inherent difference between the cost of expressing an opinion in a QV and a Likert scale survey. Expressing opinions using a Likert scale does not carry any cost—an individual freely selects any value, including extreme values, for all K options if they prefer. In QV, participants pay for their votes for each option at a quadratic cost under a limited budget.

Participants in our QV experiments face a *scarcity* of resources when they pay for options via voice credits while working with a limited budget. Shah et al. [71], in their work, "Scarcity Frames Value", found that individuals making a decision under a scarcity constraint more readily perform trade-offs among the choices and were apt to ignore contextual cues, thus becoming more consistent with rational behavior. In conditions where a survey aims to understand how people make trade-offs among K options, as we do in our two experiments, the QV mechanism with the emphasis of a limited budget may make the trade-off heuristic more accessible to participants. Thus QV may nudge participants to *trade-off* among the options and help participants make more rational decisions. We observed qualitative supports for this claim from experiment one. After experiencing a drop in the number of voice credits, one participant commented, "I had fewer credits, so each vote seemed more expensive." This comment suggested that participants experienced the idea of "scarceness" and resource constraint via the limited voice credit budget during the QV survey. While scarcity prompts trade-off thinking, having too much scarcity in QV also limits participants' flexibility of expression, a possible reason for why QV36 underperformed QV108 and QV324 in experiment one. We discuss this issue further in the next two subsections.

Flexibility of expression Another potential reason why QV aligned better with incentive-compatible behaviors is that, with a large enough budget, QV allowed participants more flexibility to express their opinions. A 5-point Likert scale provides only five choices for each question, limiting the way a participant can voice their opinion. One Likert group participant in the first experiment explicitly mentioned "[...] I would answer otherwise, if there were other options, such as not much, or a little bit." Participants wanted to express more fine-grained attitudes while a 5-point Likert scale forced them to map their preferences onto a limited fixed scale. QV, in contrast, allows participants to specify the relative distance between two options with greater flexibility, as long as the total cost does not exceed the given credits. The flexibility may enable participants to stay closer to their incentive-compatible preferences when rating the options.

We compared QV to a 5-point Likert scale in this study because the 5-point Likert scale is one of the most commonly used Likert scales in various fields [53], including public policy and HCI. One may conjecture that a 7-point or 11-point Likert scale may allow more flexibility than a 5-point Likert scale. Debates about the scale format in Likert scale surveys started in 1965 [42]. Some studies found that results among different scale formats were transferable while others found certain differences [16]. We leave the comparison of Likert scale surveys in other scale formats and QV surveys as an open question for future research.

7.2 Effects of the Amount of Voice Credits

In RQ1, we investigated how the number of voice credits available to participants impacted the QV survey results empirically. In experiment one, Bayesian analysis showed that QV results did not align with true preferences significantly better than a 5-point Likert scale until a sufficient amount of voice credits was used. We saw that QV with 36 voice credits ($O(K)$) under-performed QV with 108 ($O(K^{1.5})$) and 324 ($O(K^2)$) voice credits. We found potential explanations through participants' free-form text responses.

When participants had only 36 credits in QV, some of them voiced their need to make hard trade-offs. `P9e5e6` said, "I think I covered the bare basics." and `Pe37f2` said, "Less to go around, so had to knuckle down and allocate the most to what I think is most important." These responses indicate that while extreme scarcity still encouraged participants *ranking* behavior, it also limited their flexibility in expressing their degree of preferences, i.e. *rating* the options.

On the other hand, when participants experienced an increase in voice credits from having only 36 credits, some of them expressed their appreciation for the increased freedom to state their opinions. `Pcc4aa` reported, "with more credits I can show what I really like." and `P2d9da` stated, "Because now that I have a lot more credits, I felt that I could vote on more issues that mean something to me." The different qualitative responses for QV36, QV108 and QV324 explain why QV with more voice credits performed better than QV36 in the first experiment. These qualitative responses also supports the concept that a higher level of flexibility may contribute to why QV outperformed a 5-point Likert scale in the degree of alignment in the previous subsection.

While having fewer voice credits may worsen QV's degree of alignment with participants' incentive-compatible preferences, QV with a stringent budget may better elicit the options participants value most since it encourages harder trade-offs, based on the qualitative responses above. Future research could explore this potential effect of QV more closely.

In the first experiment, we explored up to a budget of $O(K^2)$ voice credits, where K is the number of options in the survey, specifically 324 credits. While QV aligned better than Likert scale up to this amount of voice credits and the percentage of credits used remained high (with a median of 98%), we suspect that too many voice credits may pull participants away from trade-off thinking, or create excessive cognitive loads to participants. Where the threshold of "too many" voice credits lies remains an open question for future work.

7.3 When to Use QV?

Even though one may be tempted to conclude from this study that decision-makers should prefer QV to Likert in all circumstances, the goal of our study is not to claim that one survey method should replace another. QV has its strengths, and also weaknesses.

Though QV better elicits true preferences in comparison to Likert in the two experimental contexts in this paper, QV requires a learning curve for respondents. QV is suited for online surveys when respondents have the time to familiarize themselves with QV and are likely to use their smartphones, tablets, or personal computers to respond to the survey. QV is not well suited for surveys that allow respondents to fill out their responses on paper or survey respondents via

telephone. Therefore, survey creators should carefully consider the pros and cons of QV and select the best suiting survey method in their contexts.

8 LIMITATIONS AND FUTURE WORK

QV is a relatively new area of research. Comparing the alignment of Likert scale and QV with users' true preferences is challenging. During the study, we identified various open questions that we've yet to address. In this subsection, we address our limitations and propose open questions for future research.

8.1 Comparing Ordinal Data with Numerical Data

To compare ordinal Likert scale responses with numerical donation amounts or set prices, we mapped the 5-point Likert scale to integers in the range of $[-2, 2]$. We used the number of votes in QV directly since they are numerical. We made this decision because selecting "Neutral" (mapped value = 0) in the Likert scale had a similar meaning as casting a zero vote in QV.

Whether our approach of mapping Likert data to metric values is the best approach to compare ordinal data with numerical data is debatable. At the same time, identifying the best measure to do so is challenging—the best way to analyze ordinal Likert data is still open to debate [28]. Future research could explore if there are alternatives for such a comparison that circumvent the challenge of mapping ordinal data.

8.2 Comparing QV with Other Surveying Methods

Despite the Likert scale being one of the most-used surveying techniques, one may be curious about how other voting mechanisms that capture the concept of resource constraints or make the trade-off heuristic accessible compare to QV. For example, voting algorithms in participatory budgeting (PB) [9, 30, 29, 48, 6], used by governments to ask citizens to prioritize resource allocation, involve resource constraints. In participatory budgeting, voters are asked to identify a subset I among projects P on which they would like the government to allocate resources. Furthermore, each project $i \in P$ has a fixed implementation cost c_i known to the survey taker. The survey-taker knows the total budget available B to the decision-maker (e.g., local city council) and each respondent picks a subset $I \in P$ of projects such that the total costs $\sum_{i \in I} c_i \leqslant B$ stays within the budget. Examples of PB algorithms include knapsack voting [30] and ranked voting [47].

PB algorithms elicit relative rankings among options, which incurs trade-off thinking like QV, but cannot elicit the *strength* of a participant's preferences, i.e., the degree to which the participant prefers option A over B. PB algorithms that involve implementation costs, such as knapsack voting, may be less easy to adapt to CSCW or social science surveys, since the decision-maker, who creates the survey, has to assign implementation costs to each option and an overall budget, which may be problematic. For example, in a survey about interface design, the decision-maker would need to assign costs to each interface element on which they are eliciting an opinion. Furthermore, unlike the typical participatory budgeting scenario, where project costs involve allocating *the respondent's tax dollars,* in typical social science and CSCW surveys, the implementation of an option involves no obvious cost to a participant.

An alternative option to knapsack voting that avoids the problem of cost assignment might be to use a linear constraint on the vote magnitudes—that is use $\sum_k |n_k| \leqslant B$, where $|n_k|$ is the magnitude of the vote for option k. In this case, the cost of an additional vote is constant, while that of QV increases proportionally with the vote already cast on that option. Comparing the performance of above mechanisms with QV makes an interesting open question.

8.3 Upper Bound of the Number of Options

In experiment one, our participants chose between 9 options, while in our second experiment, participants had 5 options on the survey. In both cases, QV performed well, suggesting that participants could make effective trade-offs in QV across up to 9 options. However, our study did not identify the upper bound of the number of options users can handle comfortably on a QV survey. One can imagine the difficulty for QV survey respondents to vote among dozens of survey options. In fact, work by Iyengar et al. [36] observed that more choices may not necessarily increase participants' satisfaction, suggesting that people were not good at making choices across an extensive array of options. The same phenomenon could happen in our case—is there a limit to how many options could be on a QV survey to maintain high-quality data collection?

8.4 Generalizability to Different Types of Surveys

In this study, we examined QV in two settings. We chose settings that made sense to translate into QV surveys and leveraged prior research. We did not exhaustively examine the type of survey questions that work with QV and those that may not. Hence, readers should take caution in generalizing our results to other survey settings.

We limited our experiments within the scope of resource-constrained surveys. In the first experiment, the survey asked participants to choose among K independent options for the same topic. In experiment two, the survey asked participants to choose among K dependent aspects that jointly contribute to the same topic. Though different, both of these surveys aimed to help us understand relative preferences and trade-offs. We do not yet know if QV would work for a survey in which survey options have a different relationship (e.g., surveys that consist of options that are not on the same topic), or a survey that do not involve any resource constraint.

Similarly, our study only tested QV in the context of public policy and an HCI user study. Many other disciplines make use of Likert scale surveys to help make resource-constrained decisions. Future research can explore if QV better elicits true preferences than Likert scale in other domains.

8.5 User Study on QV

Understanding how individuals learn about, feel about, and use QV is an important topic. Without a doubt, understanding how QV works requires more cognitive load than traditional voting and surveying techniques, and using QV takes more time and effort. In addition, examining respondents' mental models of QV and how the models impact individuals decision-making process may help decision-makers further understand the effectiveness of QV empirically and design better QV surveys and interfaces. Our study only scratched the surface of the above questions by using the responses from a few free-form text questions. Thus, future work may conduct a rigorous user study to explore these questions.

8.6 Interface Design for QV

The final open question is designing a simple, intuitive QV interface for empirical use. QV involves more complicated calculations than Likert. A well-designed interface should reduce a user's cognitive load to help them make accurate decisions easily. Currently, after our iterative-design process, we provided participants information such as the number of votes per option, voice credits used and voice credits remaining, and how they allocate the voice credits to each option.

Different interface designs could nudge users to behave in specific ways. How the interface should provide voters with this information in an optimal way remains an open question. Finally, we need to investigate QV interface designs for mobile and tablet devices.

8.7 Donation as True Preferences

That individuals donate to public goods, including charities, is a puzzle to economists since game theory predicts that rational individuals will free-ride, and thus there ought to be no contributions to public goods, including charities. Much of the experimental work in behavioral economics on public goods, as summarized by Fehr and Gintis [20], indicates that about 30% of the participants in these experiments were free-riders—they *do not* contribute to public goods. This is consistent with our own experimental finding where 27% of the participants (see Figure 15) *did not* contribute to public goods (i.e., charities) across all experimental groups. Andreoni [3] further developed the theory of "impure altruism" that involved "warm-glow giving", which suggests that individuals donate to public-goods in order to receive social acclaim or to avoid scorn, and is consistent with empirical observations of charitable giving. Thus the theory of 'impure altruism' may explain the donations across charities in our experiment.

While prior work [77, 26, 66, 7, 25] has used binding voluntary donations with real monetary consequences to elicit participants' incentive-compatible preferences, we acknowledge the limitations of such an approach. For example, factors, including the prior experience with a charity, may influence the donation amount. To check for possible bias, we first surveyed how each participant viewed (either favorably, neutral view or dis-favorably) the charitable organizations used in our experiments, prior to completing their donation tasks, to ensure that there was no systematic bias. All experimental groups exhibited similar favorability distributions across charities.

One other possible confound is that MTurkers may be less likely to donate since they have a strong incentive to earn money. Thus, we might expect little or no donation from MTurkers. Assuming that MTurk workers reduce their donation amount equally across all causes, we minimized this limitation by focusing on the *proportion* of donations across the charities instead of using the absolute donation value. In other words, we were looking at how much more one was willing to donate to a charity *relative* to the other organizations.

8.8 Quality of Data Collection via MTurk

Our experiments, like all other experiments conducted on MTurk, suffered from the limitation that not all participants joined the study with good faith or participated in the incentive-compatible tasks with a rational mindset. To mitigate the risk, we implemented multiple tutorials, quizzes, and attention checks, filtered out responses with facetious free-form text responses, and introduced incentive-compatible tasks with financial incentives. To ensure response quality, we confirmed that most participants used up almost all QV budgets and donated to more than one charities as shown in Appendix B.1 and Appendix B.2.

8.9 Generalizability to Non-US Population

While our experiment samples covered a wide range of ages and education levels, we targeted only the US population due to limited resources. Prior studies have found that cultural background affected response patterns in Likert scale surveys [15]. Thus, future research needs to explore if cultural background also impacts people's interaction with QV and whether our results hold under those circumstances.

9 CONCLUSION

In this paper, we examined Quadratic Voting, a computational-powered survey method that combines ratings and ranking surveying approaches, in the setting of resource-constrained collective decision-making. Through two randomized controlled experiments and Bayesian analysis, we showed empirically that a QV survey with sufficient voice credits better elicits participants' true

preferences than a Likert scale survey, with a medium to high effect size. Furthermore, our study provided the first example of applying QV in a prototypical HCI user study for the CSCW community. While our study demonstrated the potential of QV as a computational tool to facilitate truthful preference elicitation in online, resource-constrained surveys, the goal of this research is *not* to convince decision-makers to replace their Likert scale surveys with QV-based surveys. Instead, we hope to spark an interest among the CSCW community to explore a rich set of promising future research directions of QV, such as to compare QV with surveying methods besides the Likert scale survey, better understand the generalizability of QV, and improve the interface design for QV. In conclusion, we encourage decision-makers to consider QV as a promising online alternative to the Likert scale in resource-constrained scenarios where it's beneficial to elicit both the respondents' ratings and rankings preferences.

ACKNOWLEDGMENTS

We thank the voluntary pretest participants who helped us improved the experiment design and system. A big thanks to the experiment participants. Additional thanks to Vinay Koshy, Ziang Xiao, Yu-Chun Grace Yen, Chi-Hsien Eric Yen, Silas Hsu, Sneha Krishna, Meng Huang, Cian Lin and the anonymous reviewers who provided valuable feedback to this work. Last but not least, we like to thank meteorologist Howard Bernstein for his awesome weather forecasts that we used in one of our experiments. This work was partially supported by Microsoft, Facebook and Capital One Financial Corporation.

A EXPERIMENT ONE METHODS

A.1 Definitions of the Nine Societal Causes

In this subsection, we detailed the definitions of the nine societal causes used in the first experiment. We derived these causes from the categorization of charity groups on Amazon Smile [15], to ensure that the nine societal causes covered a broad spectrum of categories. The nine categories were defined as:

(1) Pets and Animals: Animal Rights, Welfare, and Services; Wildlife Conservation; Zoos and Aquariums
(2) Arts, Culture, Humanities: Libraries, Historical Societies, and Landmark Preservation; Museums; Performing Arts; Public Broadcasting and Media
(3) Education: Early Childhood Programs and Services; Youth Education Programs and Services; Adult Education Programs and Services; Special Education; Education Policy and Reform; Scholarship and Financial Support
(4) Environment: Environmental Protection and Conservation; Botanical Gardens, Parks and Nature Centers
(5) Health: Diseases, Disorders, and Disciplines; Patient and Family Support; Treatment and Prevention Services; Medical Research
(6) Human Services: Children's and Family Services; Youth Development, Shelter, and Crisis Services; Food Banks, Food Pantries, and Food Distribution; Multipurpose Human Service Organizations; Homeless Services; Social Services
(7) International: Development and Relief Services; International Peace, Security, and Affairs; Humanitarian Relief Supplies
(8) Faith and Spiritual: Religious Activities; Religious Media and Broadcasting
(9) Veterans: Wounded Troops Services, Military Social Services, Military Family Support

The participants saw the same definitions when completing the surveys during the study.

[15] https://smile.amazon.com/

B EXPERIMENT ONE RESULTS

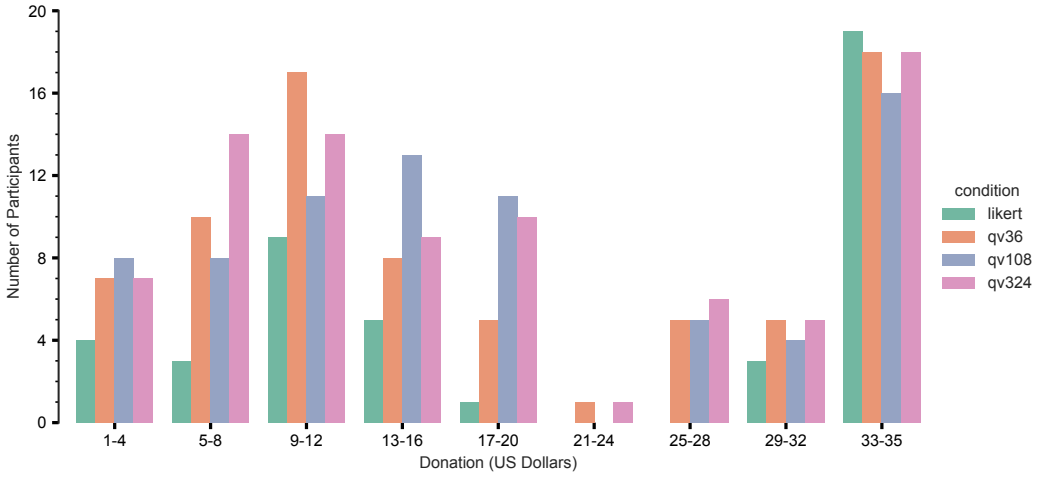


Fig. 15. Distributions of the total amount donated by participants across four surveying methods. This figure also included participants that did not donate any amount, whom we excluded from our analysis. We saw two distributions, one centered by \$9 – 12 and the other centered by \$33 to \$35. We also see more Likert participants donate almost most of their donation quota compared to the QV Groups.

B.1 Total Donation Amount

Figure 15 also demonstrates two clusters for the total donation amount. The first cluster centered around \$9 – 12 with the majority in the range of \$5 – 20. This group of people, making up about 60% of the entire sample, donated part of the lottery winning amount but still kept a significant portion for themselves. The other clustered around \$33 to \$35, suggesting that this group of participants contributed almost the full amount of the lottery prize. There were approximately 25% of the participants who behaved this way. The total donation amount distribution across four surveying methods were relatively consistent, except that almost twice the proportion of participants in the Likert condition donated almost the full amount compared to the other QV conditions. One possible explanation for the difference is the Likert group required less effort compared to that of QV, and participants felt less tempted to earn an extra reward for their time spent in the Likert condition.

B.2 How Participants Donated

To ensure that participants distributed their donation amount, we extracted the information on how each individual donated. Figure 16 shows the distribution of how the participants contributed to each group. On an aggregated level, only 18.57% of participants donated to a single charity, and 59.29% of participants donated to three or more charity.

B.3 QV Budget Usage

To understand how participants used their budgets, we examined the percentage of credits consumed. Participants do not need to use up all their budgets in QV. We found no decrease in the median of percentage budget usage – all around 98%, as available voice credits increased. QV324 did exhibit a longer tail for percentage budget usage. Participants, in general, used up their budgets as much as they can while they make trade-offs between options, under budget constraints.

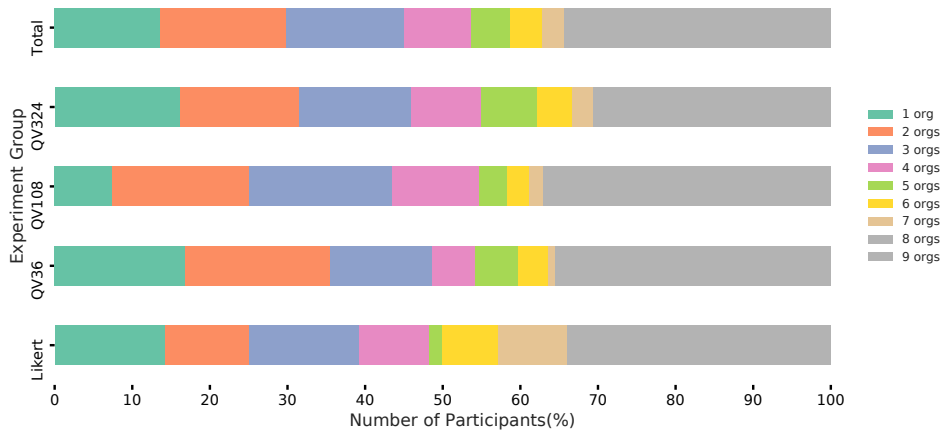


Fig. 16. The distributions of how participants donated within each experiment group for experiment one. This plot removed participants that donated to zero organizations since they were excluded from our original analysis. We see that about 80% of participants donated to more than two or more organizations across all experiment groups. More than half of all participants donated to more than three organizations.

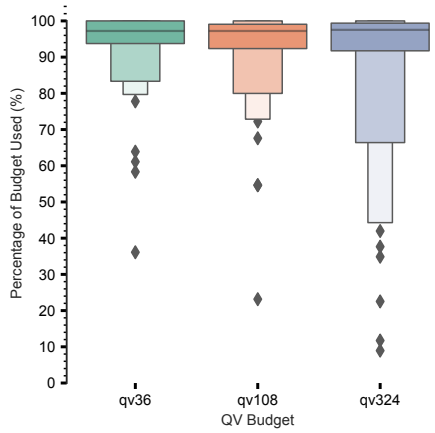


Fig. 17. Distribution of Percentage Budget Used in QV36, QV108 and QV324. Percentage budget used is the percentage of voice credits used out of the total voice credits budget available. The medians for all three QVs are around 98%.

C EXPERIMENT TWO DESIGN

C.1 Experiment 2 Flow Chart

We present the experiment flow chart for experiment two in Figure 18. Specifically, Figure 19 provides the two interfaces involved in Step 6.

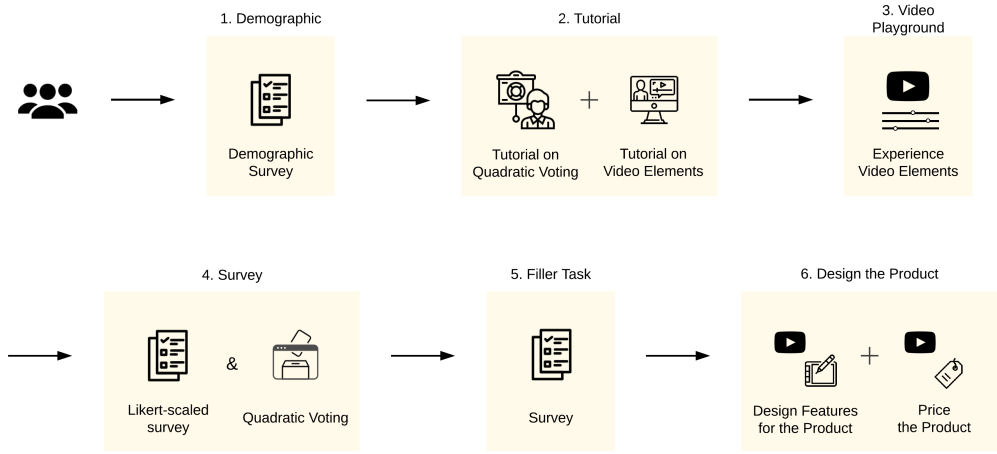


Fig. 18. We used a within-subjects design for experiment two. We randomly assigned participants into two groups: one group would complete the Likert scale first and then quadratic voting; the other group experienced them in a reversed order.

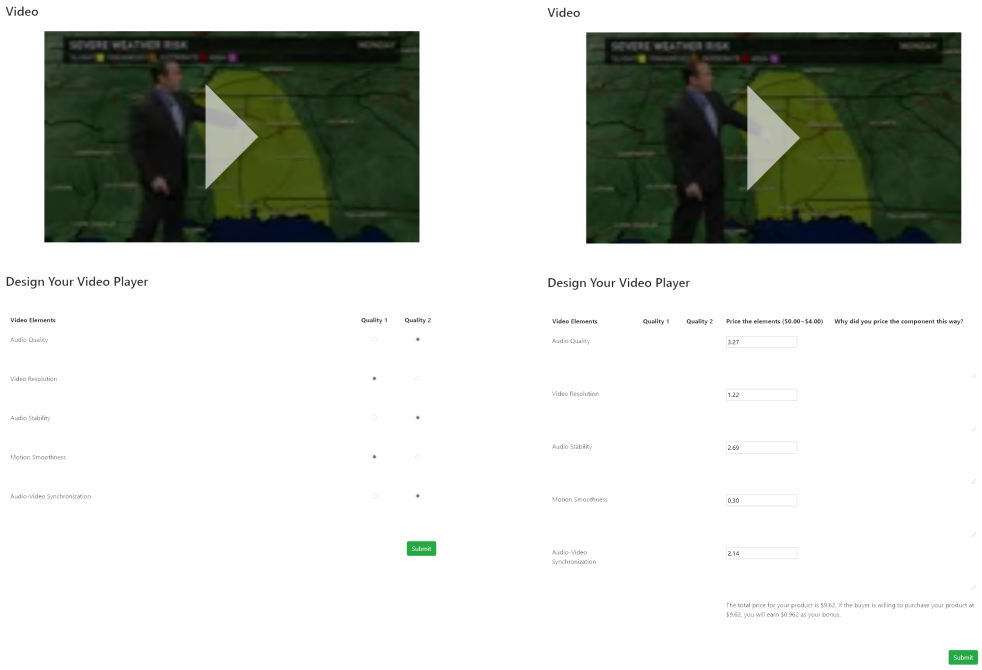


Fig. 19. The two steps participants encountered when completing Step 6 of the survey. Participants would first need to select which one of the two qualities they would include in their video streaming product. Participants could see real-time changes to the video as they updated the qualities. Once they made their decisions, participants would price each of the elements between $0 and $4. Participants would receive a commission worth 10% of the total price if the buyer accepted their product at their set prices.

C.2 Definition of the Five Video Elements for Experiment Two

In experiment two, we designed the research scenario to answer the following question: "Given a video with unsatisfying quality, under limited bandwidth, how should the bandwidth be allocated to enhance the five video and audio elements, including the motion smoothness [34], audio stability [32], audio quality [41], video resolution [40], and audio-video synchronization [74], to obtain an acceptable video streaming experience from the viewers' perspective?" We selected the five video playback elements based on prior work, and below are their definitions:

- Motion Smoothness [34, 62]: refers to how smooth the visuals of the video are. The number of frames transferred from the server to the viewer per second may be impacted under limited bandwidth. Having a low frame rate means that the video feels jerky and slow.
- Audio Stability [32]: refers to how smoothly the audio of the video plays. With limited bandwidth, there may be lost audio packets. This creates short intervals of silence during playback, undermining the interpretability of the audio. The higher probability an audio packet may be lost, the more stuttered the audio sounds.
- Video Resolution [62, 40]: refers to how sharp the visuals in the video look. With limited bandwidth, one may reduce the video's size by providing a lower resolution. At a lower resolution, the video imagery becomes pixelated and unclear.
- Audio Quality [62, 61]: refers to how clear and crisp the audio sounds. A lower audio sampling rate needs a lower bandwidth to transmit. With a lower audio sampling rate, the audio sounds more muffled and unclear.
- Audio-Video Synchronization [74]: refers to how well video visuals are matched with the audio playback. Our experiment focused only on the type of asynchronization where the audio plays ahead of the video. Under bandwidth constraint, visuals and audio may be out of sync due to packet loss in visuals or audio.

D SYSTEM DESIGN DETAILS

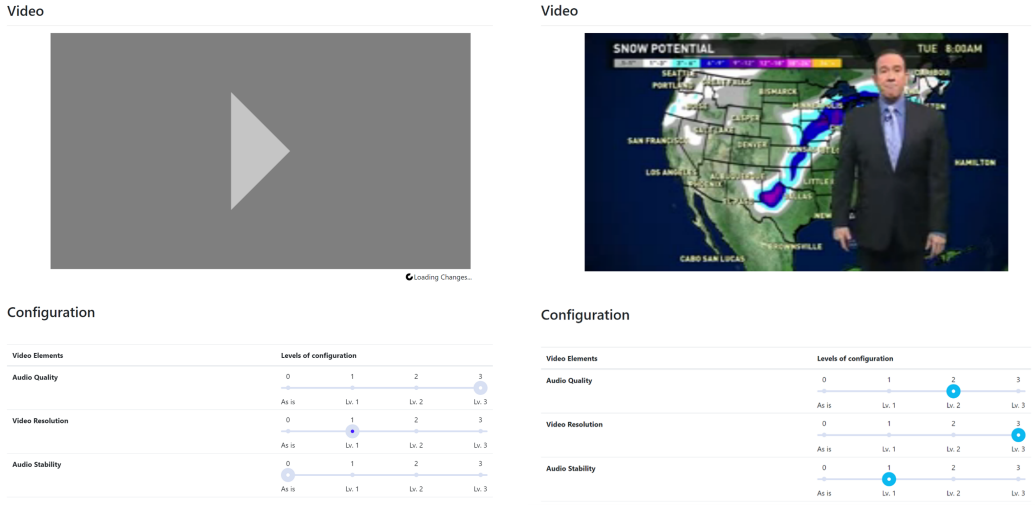
D.1 Experiment two video interface implementation detials

For this experimental setup, we used AngularJS and bootstrap for the front-end implementation powered by Flask web framework written in Python. We used MongoDB Atlas to store data and Heroku to serve the system. Survey.js rendered all types of surveys besides the QV interface in our experiment. The experiment source code is publicly available [16] and so is the standalone QV interface [17].

In this experiment, the most challenging component is to design a stable, reliable, and real-time video rendering interface for participants to experience how changes in different video element quality contribute to their overall experience. Since we provided four possible levels of adjustments for each element, there will be $4^5 = 1024$ possible combinations, which is impossible to pre-generate and serve to the participants. Therefore, we need a real-time rendering video playback system in our experiment.

To the best of our knowledge, no video player supports real-time rendering of different video qualities ,and that there is little work that degrades video playback purposely. Thus, to achieve our experiment goal, we need to implement our own video playback system Figure 8. We broke the video clip into two parts: (1) a video without audio and (2) audio. In our final experiment, we pre-generated $4^2 = 32$ files. $4^2 = 16$ of which are different levels of video quality ($N = 4$) and frame rates ($N = 4$) while the other 4^2 versions are varying levels of audio quality ($N = 4$) and audio stability ($N = 4$). We decided not to use the server nor the browser to render the video quality

[16] https://github.com/a2975667/QV-buyback
[17] https://github.com/hank0982/SimpleQV



(a) Video interface when loading

(b) Video interface when playing

Fig. 20. To assure video playback consistency, the interface would signal loading when switching video and audio files upon participants changing the toggles. The image to the left shows how the cue was presented. The image to the right shows how the system works under normal conditions.

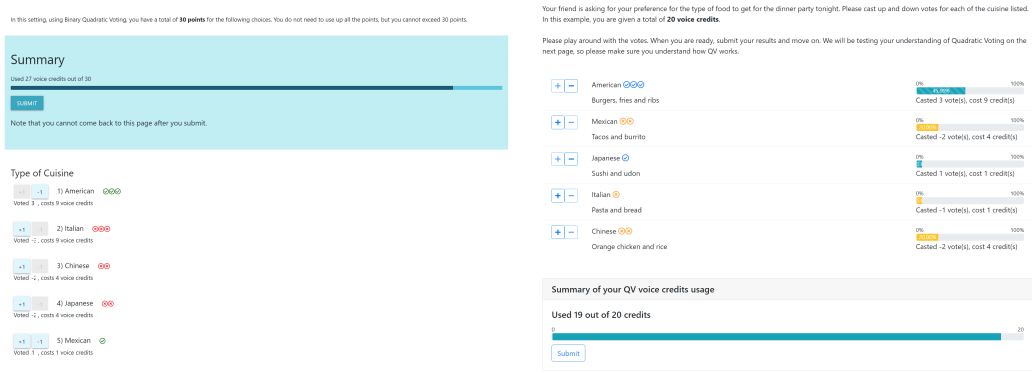
or frame rates in real-time to control what the participants see even in areas of lower internet bandwidth or when there is congestion on the server. We pre-generated the audio files because of similar reasons and pre-generation made sure the locations of packet loss were consistent across participants. FFMPEG was used to generate all 16 files. Video files were first decoded and encoded at the desired resolution, bitrate and framerate. Audio files were first decoded and and encoded at the designated bit rate. We simulated audio stability by randomly losing 40 milliseconds of packets according to the probability listed in 5.1.

Once these files were pre-generated, even the most distinct environment could see the same video and audio files. Understanding that network environments might delay the transmission of video and audio files, we implemented a spinning wheel in the interface to signal while the files are still loading (Figure 20a). Once the client received the correct combination of files, it will play both files simultaneously according to the corresponding time anchor (Figure 20b). A front-end JavaScript determined this time anchor, simulating the video-audio synchronization levels by playing the video and audio files from different start times.

D.2 QV interface iterations

The current QV interface was designed over multiple iterations. The goal of the interface is to assist participants' voting process using visual information to reduce their cognitive load. Figure 21 portraits the draft QV interface with the current design used in our experiments. Both interfaces featured a voting panel that contained a list of options to vote on. To the left of each option, participants can use the plus and minus buttons to vote for or against an option. Buttons for an item were automatically disabled if the number of voice credits remaining did not permit an additional vote for that item.

The major difference lies in the information presented to the participants. In the new interface, we provided a bar with the proportion of voice credits contributed to that option rather than



(a) The initial QV interface

(b) The current QV interface

Fig. 21. The QV interface was redesigned multiple times. The image on the left showcased the same QV playground in an early iteration of the interface design. The new design provided much more information to reduce participant's cognitive loads.

simple text under each option. Comparing the two interfaces, the new interface also allowed more description to be present under each of the options. In addition, we floated the summary panel at the bottom of the page at any time to ensure visibility, which provided information on the number of voice credits the participants have and have not used out of the total budget. Finally, we also changed the color scheme of the interface for accessibility reasons.

REFERENCES

[1] 2019. $120 million in requests and $40 million in the bank. how an obscure theory helped prioritize the colorado budget. Retrieved Oct. 15, 2020 from https://coloradosun.com/2019/05/28/quadratic-voting-colorado-house-budget/.

[2] Duane F Alwin and Jon A Krosnick. 1985. The measurement of values in surveys: a comparison of ratings and rankings. *Public Opinion Quarterly*, 49, 4, 535–552.

[3] James Andreoni. 1989. Giving with impure altruism: applications to charity and ricardian equivalence. *Journal of political Economy*, 97, 6, 1447–1458.

[4] Theo Araujo, Anke Wonneberger, Peter Neijens, and Claes de Vreese. 2017. How much time do you spend online? understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures*, 11, 3, 173–190.

[5] John H Batchelor and Chao Miao. 2016. Extreme response style: a meta-analysis. *Journal of Organizational Psychology*, 16, 2.

[6] Gerdus Benade, Swaprava Nath, Ariel D Procaccia, and Nisarg Shah. 2020. Preference elicitation for participatory budgeting. *Management Science*.

[7] Matthias Benz and Stephan Meier. 2008. Do people behave in experiments as in the field?—evidence from donations. *Experimental economics*, 11, 3, 268–281.

[8] Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. 2012. Evaluating online labor markets for experimental research: amazon. com's mechanical turk. *Political analysis*, 20, 3, 351–368.

[9] Yves Cabannes. 2004. Participatory budgeting: a significant contribution to participatory democracy. *Environment and urbanization*, 16, 1, 27–46.

[10] Charlotte Cavaillé, Daniel L Chen, and Karine Van Der Straeten. 2018. Towards a general theory of survey response: likert scales vs. quadratic voting for attitudinal research.

[11] Patricia A Champ, Richard C Bishop, Thomas C Brown, and Daniel W McCollum. 1997. Using donation mechanisms to value nonuse benefits from public goods. *Journal of environmental economics and management*, 33, 2, 151–162.

[12] Yuan-Chia Chang, Hao-Chuan Wang, Hung-kuo Chu, Shung-Ying Lin, and Shuo-Ping Wang. 2017. Alpharead: support unambiguous referencing in remote collaboration with readable object annotation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2246–2259.

[13] Ti-Chung Cheng, Tiffany Wenting Li, Karrie Karahalios, and Hari Sundaram. 2021. Dataset for ' "I can show what I really like.": Eliciting Preferences via Quadratic Voting'. (2021).

[14] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 1217–1230.

[15] Heather M Davey, Alexandra L Barratt, Phyllis N Butow, and Jonathan J Deeks. 2007. A one-item question with a likert or visual analog scale adequately measured current anxiety. *Journal of clinical epidemiology*, 60, 4, 356–360.

[16] John Dawes. 2008. Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales. *International journal of market research*, 50, 1, 61–104.

[17] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 135–143.

[18] 2019. Educational attainment in the united states: 2018. Retrieved Oct. 13, 2020 from https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html.

[19] Jon X Eguia, Nicole Immorlica, Katrina Ligett, E Glen Weyl, and Dimitrios Xefteris. 2019. Quadratic voting with multiple alternatives. *Available at SSRN 3319508*.

[20] Ernst Fehr and Herbert Gintis. 2007. Human motivation and social cooperation: experimental and analytical foundations. English. *Annual Review of Sociology*, 33, pp. 43–64.

[21] Ernst Fehr and Herbert Gintis. 2007. Human motivation and social cooperation: experimental and analytical foundations. *Annu. Rev. Sociol.*, 33, 43–64.

[22] Kraig Finstad. 2010. Response interpolation and scale sensitivity: evidence against 5-point scales. *Journal of Usability Studies*, 5, 3, 104–110.

[23] Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and individual differences*, 7, 3, 385–400.

[24] Kate J Garland and Jan M Noyes. 2008. Computer attitude scales: how relevant today? *Computers in Human Behavior*, 24, 2, 563–575.

[25] Philip Gendall and Benjamin Healey. 2010. Effect of a promised donation to charity on survey response. *International Journal of Market Research*, 52, 5, 565–577.

[26] Michael Getzner. 2000. Hypothetical and real economic commitments, and social status, in valuing a species protection programme. *Journal of Environmental planning and Management*, 43, 4, 541–559.

[27] P Glaser. 2008. Response rates. encyclopedia of survey research methods. (2008).

[28] Rainer Göb, Christopher McCollin, and Maria Fernanda Ramalhoto. 2007. Ordinal methodology in the analysis of likert scales. *Quality & Quantity*, 41, 5, 601–626.

[29] Ashish Goel, Anilesh K Krishnaswamy, and Sukolsak Sakshuwong. 2016. Budget aggregation via knapsack voting: welfare-maximization and strategy-proofness. *Collective Intelligence*, 783–809.

[30] Ashish Goel, Anilesh K Krishnaswamy, Sukolsak Sakshuwong, and Tanja Aitamurto. [n. d.] Knapsack voting.

[31] Sandy JJ Gould, Anna L Cox, Duncan P Brumby, and Sarah Wiseman. 2015. Home is where the lab is: a comparison of online and lab data from a time-sensitive study of interruption. *Human Computation*, 2, 1, 45–67.

[32] Vicky Hardman, Martina Angela Sasse, and Isidor Kouvelas. 1998. Successful multiparty audio communication over the internet. *Communications of the ACM*, 41, 5, 74–80.

[33] David R Hodge and David Gillespie. 2003. Phrase completions: an alternative to likert scales.(note on research methodology). *Social Work Research*, 27, 1, 45–56.

[34] Quan Huynh-Thu and Mohammed Ghanbari. 2008. Temporal aspect of perceived quality in mobile video broadcasting. *IEEE Transactions on Broadcasting*, 54, 3, 641–651.

[35] John PA Ioannidis. 2005. Why most published research findings are false. *PLoS medicine*, 2, 8, e124.

[36] Sheena S Iyengar and Mark R Lepper. 2000. When choice is demotivating: can one desire too much of a good thing? *Journal of personality and social psychology*, 79, 6, 995.

[37] Jeremy P Jamieson and Stephen G Harkins. 2011. The intervening task method: implications for measuring mediation. *Personality and Social Psychology Bulletin*, 37, 5, 652–661.

[38] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: explored and explained. *Current Journal of Applied Science and Technology*, 396–403.

[39] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: why bayesian statistics better fit the culture and incentives of hci. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4521–4532.

[40] Hendrik Knoche, John D McCarthy, and M Angela Sasse. 2005. Can small be beautiful? assessing image resolution requirements for mobile tv. In *Proceedings of the 13th annual ACM international conference on Multimedia*, 829–838.

[41] Hendrik Knoche, John D Mccarthy, and M Angela Sasse. 2008. How low can you go? the effect of low resolutions on shot types in mobile tv. *Multimedia Tools and Applications*, 36, 1-2, 145–166.

[42] Samuel S Komorita and William K Graham. 1965. Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, 25, 4, 987–995.

[43] John K Kruschke. 2010. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1, 5, 658–676.

[44] Ozan Kuru and Josh Pasek. 2016. Improving social media measurement in surveys: avoiding acquiescence bias in facebook research. *Computers in Human Behavior*, 57, 82–92.

[45] Steven P Lalley and E Glen Weyl. 2018. Quadratic voting: how mechanism design can radicalize democracy. In *AEA Papers and Proceedings*. Vol. 108, 33–37.

[46] Steven P Lalley and E Glen Weyl. 2018. Quadratic voting: how mechanism design can radicalize democracy. In *AEA Papers and Proceedings*. Vol. 108, 33–37.

[47] David Ledo, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. Evaluation strategies for hci toolkit research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–17.

[48] David Timothy Lee, Ashish Goel, Tanja Aitamurto, and Helene Landemore. 2014. Crowdsourcing for participatory democracies: efficient elicitation of social choice functions. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[49] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

[50] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. Turkprime. com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, 49, 2, 433–442.

[51] 2019. Little public support for reductions in federal spending. Retrieved Oct. 15, 2020 from https://www.pewresearch.org/politics/2019/04/11/little-public-support-for-reductions-in-federal-spending/.

[52] Xiaojuan Ma and Nan Cao. 2017. Video-based evanescent, anonymous, asynchronous social interaction: motivation and adaption to medium. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 770–782.

[53] Naresh K Malhotra and Mark Peterson. 2006. *Basic marketing research: A decision-making approach*. Prentice hall.

[54] Richard McElreath. 2015. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

[55] Stephan Meier. 2007. Do subsidies increase charitable giving in the long run? matching donations in a field experiment. *Journal of the European Economic Association*, 5, 6, 1203–1222.

[56] Gerhard Meisenberg and Amandy Williams. 2008. Are acquiescent and extreme response styles related to low intelligence and education? *Personality and individual differences*, 44, 7, 1539–1550.

[57] Andreea Molnar and Cristina Hava Muntean. 2013. Comedy: viewer trade off between multimedia quality and monetary benefits. In *2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 1–7.

[58] Guy Moors, Ingrid Vriens, John PTM Gelissen, and Jeroen K Vermunt. 2016. Two of a kind. similarities between ranking and rating data in measuring values. In *Survey Research Methods* number 1. Vol. 10, 15–33.

[59] Ryan Naylor et al. 2017. First year student conceptions of success: what really matters? *Student Success*, 8, 2, 9–19.

[60] Olga Nikolayeva, Esteban Buz, Linda Liu, Andrew Watts, and Tim Florian Jaeger. 2015. Web based tutorial on experimental design. https://www.hlp.rochester.edu/resources/BCS152-Tutorial/.

[61] Peter Noll. 1993. Wideband speech and audio coding. *IEEE Communications Magazine*, 31, 11, 34–44.

[62] Anne Oeldorf-Hirsch, Jonathan Donner, and Edward Cutrell. 2012. How bad is good enough?: exploring mobile video quality trade-offs for bandwidth-constrained consumers. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*. ACM, 49–58.

[63] Sharon Oviatt. 2006. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th ACM international conference on Multimedia*, 871–880.

[64] Eric A Posner and E Glen Weyl. 2018. *Radical markets: Uprooting capitalism and democracy for a just society*. Princeton University Press.

[65] David Quarfoot, 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.

[66] Richard C Ready, Patricia A Champ, and Jennifer L Lawton. 2010. Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment. *Land Economics*, 86, 2, 363–381.

[67] Catherine A Roster, Lorenzo Lucianetti, and Gerald Albaum. 2015. Exploring slider vs. categorical response formats in web-based surveys. *Journal of Research Practice*, 11, 1, 1.

[68] Alvin E Roth. 1982. Incentive compatibility in a market with indivisible goods. *Economics letters*, 9, 2, 127–132.

[69] Tim Roughgarden. 2010. Algorithmic game theory. *Communications of the ACM*, 53, 7, 78–86.

[70] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. 2016. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2, e55.

[71] Anuj K. Shah, Eldar Shafir, and Sendhil Mullainathan. 2015. Scarcity frames value. *Psychological Science*, 26, 4, 402–412.

[72] 2020. Shelf solutions. Retrieved Oct. 12, 2020 from https://www.nielsen.com/apac/en/solutions/measurement/shelf-solutions/.
[73] Amit Singhal et al. 2001. Modern information retrieval: a brief overview. *IEEE Data Eng. Bull.*, 24, 4, 35–43.
[74] Ralf Steinmetz. 1996. Human perception of jitter and media synchronization. *IEEE Journal on selected Areas in Communications*, 14, 1, 61–72.
[75] Lynn Vavreck et al. 2007. The exaggerated effects of advertising on turnout: the dangers of self-reports. *Quarterly Journal of Political Science*, 2, 4, 325–343.
[76] Ziang Xiao, Po-Shiun Ho, Xinran Wang, Karrie Karahalios, and Hari Sundaram. 2019. Should we use an abstract comic form to persuade? experiments with online charitable donation. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW, 1–28.
[77] Ewa Zawojska, Mikołaj Czajkowski, et al. 2015. Re-examining empirical evidence on contingent valuation–Importance of incentive compatibility. Tech. rep.
[78] Amy X Zhang, Lea Verou, and David Karger. 2017. Wikum: bridging discussion forums and wikis using recursive summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2082–2096.

Received October 2020; revised January 2021; accepted January 2021