"…I can show what I really like.": How Quadratic Voting better align true preferences than Likert Scale Surveys

A clear and well-documented LaTeX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the "acmart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

CCS Concepts: • **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
. 2020. "…I can show what I really like.": How Quadratic Voting better align true preferences than Likert Scale Surveys. In *CSCW '20: The 23rd ACM Conference on Computer-Supported Cooperative Work and Social Computing, Oct 17 – 21, 2020, Virtual.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Likert scale survey is one of the most widely used methods to obtain the participant's opinion in the realm of human-computer interaction. Survey participants would express a rating across a series of measurements — *Very agree to very disagree* or *On a scale of 1 to 5* — for a listed statement. Very often, these opinions help researchers or decision-makers uncover consenses across a group of people.

However, there had been findings of how researchers can easily misuse Likert scale surveys either applying incorrect analysis methods [3] or misinterpreting the analysis results [12, 24] leading to questionable findings. In addition, many research papers do not explain the rational behind the use of Likert scale surveys. In a community that adopted Likert scale surveys almost as the defacto standard, we ask a fundamental question: "Is Likert-scale survey the ideal method to measure collective attitudes for decision making?"

We begin by exploring one type of question in collective decision making that aims To elicit user preferences among K options. Research agencies, industry labs or independent researchers often want to understand how to better allocate resources. For example, ordinal scale polls were designed to understand public opinions on government policy [1] because there is limited funding. Companies deploy online surveys to understand how product users feel about the features and services that needs further improvements because companies have limited time to develop the next release. Physical surveys can be found in shopping centers to collect an individual's experiences for products on the shelf because there are limited shelves. All these examples demonstrated how surveys are often tied to making decisions by gathering consensus from surveying individual's attitudes.

CSCW '20, Oct 17 – 21, 2020, Virtual

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

In this study, we look at an alternative method called Quadratic Voting (QV). Published in 2015, Weyl at al. [25] proposed Quadratic voting as a voting mechanism with approximate Pareto efficiency. Under this voting mechanism, voters were initially given a fixed amount of voice credits (VC). With the credits, individuals can purchase any number of votes to support any of the statements listed on the ballot. However, the cost of each vote increases quadratically when voted toward the same option. The authors proved that this mechanism is more efficient at making a collective decision because it minimizes welfare loss. Since 2015, a few studies compared Likert scaled surveys with QV empirically and theoretically [19, 26]. Cavaille et al. argues that QV outperforms Likert-scale surveys among a set of political and economic issues [4]. Despite these findings, we are not aware of related works that compare Likert scale surveys and QV with participants' underlying true preferences. Therefore, it is unclear whether or not and in what degree does QV results align with participants' behaviors. In addition, no current work, to the best of our knowledge, deployed QV in the area of HCI.

To be more specific, we ask the following research questions:

RQ1. How does results from QV, Likert scaled survey align with people's behavior when surveying societal issues?
RQ2. How does results from QV, Likert-scale align with people's behavior when placed in an HCI context?
RQ3. How do different amounts of voice credits impact results of QV empirically?
RQ4. What are some qualitative insights that can be observed when participants vote under QV?

To answer these research questions, we designed two experiments. The first experiment, designed to answer RQ1 and RQ3, is a between-subject study where participants express their attitudes among a set of societal causes using QV and Likert-scaled surveys and then donate to organizations relevant to these organizations. The second experiment created an HCI study environment, aimed to answer RQ2, where participants were asked about opinions among different video elements and their opinions using QV and Likert-scaled surveys. Our results showed that both experiments support QV in providing a clean and efficient way compared to Likert scale surveys at eliciting participant's true preferences.

Contributions Our work made several contributions to the research community. First, we proved empirically the use of QV outperforms Likert scale survey when conducting "choosing one in K" experiments. Second, we showed that the usability of QV is transferrable from a generic domain to HCI. Third, we designed a bayesian model that facilitates the comparison of Likert scale surveys, QV, and behaviors. Fourth, we developed an online experiment to mimic real-life HCI-related decision making. And finally, we provided the source code of our easy to deploy, interactive web platform for QV to the community.

Design Implication TODO. Talk about interface, future work and insights.

2 RELATED WORKS

In this section, we first explain the challenges that Likert faces and then describe related works of QV.

2.1 Likert Scale Surveys

Likert-scale survey, is a commonly used method to collects participant's opinions because of its ease of use. These surveys are deployed to validate findings or clarify hypotheses [15, 22] in HCI. They are also used to verify or uncover the user's needs. The original Likert scale surveys were invented by Rensis Likert in 1932, which utilizes step-intervals from one attitude to the next on the scale [16]. Researchers today design 3, 5, 7, or even 12-point Likert scale evaluation surveys to accommodate different uses [9, 10]. In addition, these surveys can also use verbal descriptions to demonstrate an ordinal scale. Some researchers even developed alternative forms of Likert scale such as slider scales [27] or phrase completions [11], which aimed to circumvent some of the shortcomings of the traditional Likert scale.

There are, however, widespread controversies in the community on when, why, and how to use it [3]. For instance, researchers can misuse statistical methods , such as using mean and standard deviations [12], to understand outcomes when working with an ordinal metric. Quantifying aggregated Likert scale surveys, such as explaining what "agree and a half" means can also be unclear. Besides, as options on the survey can be stepwise, one should not assume scales to be equally divided, which can be confusing. In other words, strongly agree and agree can be different compared to neutral and agree [7, 12].

An empirical study by Quarfoot et al. identified another challenge where people exaggerate their views when filling out political surveys [26]. In this study, participants often express strong polarized opinions or have no opinion at all, making it hard to form optimal conclusions[25]. This occurrence was theoretically proved by Cavaille et al. [4], where respondents tend to overstate their values if they want to influence the results through the survey. These challenges motivated us to understand whether QV can fill the gap and provide a more accurate measurement for collective decision making.

2.2 Quadratic Voting

Quadratic voting originated with the argument that current one-person-one-vote system can easily bias toward the majority's opinion and omitting the minority's votes[25]. This phenomenon is termed as the tyranny of the majority where the democratic decision does not take care of those in need. In other words, these types of voting does not allow fine-grain responses to the options they were to vote [23]. Some voting mechanisms tried to resolve this by introducing rank-based voting in which voters would decide how they rank the options while submitting their opinions. This mechanism can however suffer from Condorcet's paradox where results can be suboptimal because the ranks of the voters might not be transitive[23]. Many other voting mechanisms suffer from similar issues.

This triggers the development of Quadratic Voting created by Weyl at al. to overcome traditional voting challenges [25]. QV tries to capture a cost for the voter when he or she made a particular decision by voting toward specific options. This "price-taking" equilibrium helps participants maximizes their utility using their votes. This is theoretically proven by Lalley et al. [14] and showed that there exists an approximate structure of Bayes-Nash equilibria.

In order for QV to capture the voice of the minority and allocate the correct cost to the votes QV has the following mechanism: Consider collecting N participants voting, each participant is entitled to K voice credits. Participants can express their binary opinion (for or against) each optiono_i within a set of options O listed on the ballot. Participants can purchase any number of votes $v_i \in \mathbb{R}$ vetted toward any of the options o_i. However, to vote v_i votes toward o_i, participants have to spend v_i^2 voice credits, billed toward their k credits. The outcome of QV would be the ranks of the sum among the total votes for any option $\sum V_{oi}$ across all N participants.

To use an analogy, Suppose every voter has a bank account with a fixed amount of money, say 100 dollars. On a ballot, there are ten statements. Voters can now buy votes using their money in the account. However, for each statement, the cost of the votes increases quadratically. For example, voting two fors on the first option would cost the voter four votes; voting three against on the fifth option costs the voter nine votes, and so on. This means that the more votes one devote to an option, the more costly it is to do so, forgoing the opportunity the voter had to vote for other options.

2.3 QV in the wild

After QV was proposed, Quarfoot et al. conducted an empirical study to understand how QV results compare to Likert scale surveys. They recruited 4500 participants to survey an individual's opinions across ten public policy using either Likert-style questionnaire, QV survey, or both. The study found that the number of people who voted for the same number of votes for the options distributed normally and consistently across all options. This differs from results from Likert scaled surveys completed from the same group of participants, where results are either heavily skewed or polarized "W-shaped"distribution. Researchers also saw individuals spent more time expressing their opinion and reveal a more fine-grain attitude toward the policies. Thus, the study concludes that QV provides a clearer picture of public opinion to policymakers [26].

The work by Quarfoot et al., however, only used mean and z-scores, to compare the final aggregated results across the two methods. In addition, the design on the policies have a strong tendency for voters to agree or disagree on the extreme, such as one's opinion on "same-sex marriage". Thus, little do we know if QV produces different results than Likert scale surveys if the options are less competitive, for example, choosing one's favorite ice cream flavor.

Another empirical study was applied to education by Ryan Naylor [19]. The author again used QV and Likert to understand essential elements among a list of factor that impacts students' success in universities. Results showed that QV provided more insights, such as distinguishing good-to-have factors from must-have elements. These factors are not heated debated controversies compared to public policies in the previous studies and they are independent elements that do not require students to make trade-offs. For example, students can have a sense of "belonging" and a sense of "achievement" at the same time.

To the best of our knowledge, there does not exist an empirical study that focused on investigating how QV and Likert perform under the condition of selecting one in K options. This setting was recently discussed in a theoretical work by Eguia et al. [8], who claims that QV is still in favor of resolving budget-constrained for risk natural agents to figure out an efficient decision across multiple alternatives as a collective choice problem. We aim to complete this missing piece of the puzzle. Further, we are not aware of any work that studies alignment between participants' actual beliefs and QV surveys. Existing research pointed out possible fallacy exists with self-reporting [2, 29]. Thus, we aim to understand how QV and Likert scale surveys align with the agent's true beliefs. We also want to test whether the total number of voice credits impacted the results of QV. Finally, we believe that no HCI research utilized QV to form design decisions.

3 METHODS

We designed two experiments to investigate our research questions. The first experiment applied both QV and Likert-scaled surveys to measure people's preferences among societal issues. We then deployed a donation task to match the results of the surveys to people's preference. The second experiment extends upon the first one with a focus in the context of HCI survey.

3.1 Demographics

3.2 Experiment 1

The first experiment aims to answer the first research question. We try to understand how QV aligns with people's true preference compared to Likert-scaled surveys when a group of people is selecting n items among k options. This experiment also aims to answer the third research question: trying to observe if and how different numbers of voice credit impact participants QV responses. Conducted between subjects, the first experiment is made up of three primary segment: demographics, surveys and donation. This process is demonstrated in graph X(a).

Both group of participants will fill out a demographic survey after agreeing the consent form. This demographic survey captures participant's basic information such as age, gender, income, ethnicity, profession and so on. Participants are divided into two groups: QV group and Likert group. Participants in the QV group will first go through a tutorial on what QV is and how QV works using a pre-recorded video. To make sure that the participants understand the concepts correctly, they have to correctly answer questions regarding the concepts in order to move on. They would experience 4 QV surveys shown as graphic X(b).

The first two QV surveys asked participants to vote with QV among 9 different societal causes based on the causes they think requires more resource allocation. The only difference between the first two surveys are the number of voice credits the participants have to express their votes. The two number of credits a participants can have are equally drawn from two of the three possible number of credits: 36, 108 and 324. [Explain the three voice credits here] The last two surveys are designed to look similar to the first two. They differ at the set of nine societal causes presented to the participants. These nine causes have no direct purpose to the experiment. They are designed to distract the participants and prevent them making connections to the donation task which follows. Note that the selection of voice credits for the first two QV survey would be the same for the latter two QV survey. The second group of participants completes two Likert-scaled surveys. The two surveys mirror the nine societal causes listed on the first and third QV surveys. Both surveys are provided in the supplementary materials.

A donation task is deployed to the participants after the surveys were completed to measure the true preferences based on participant's behaviors. Participants are told that for every 70 participants, one participant would win 35 US dollars. Assuming winning the 35 US dollars, the participants were asked if they would want to donate some money to some of the nine charity groups. These nine charity groups mirrors the nine societal causes listed on the first two QV survey and the first likert-scaled survey. Participants are aware that the research team will match one dollar to each one dollar they donated to an organization. Participants are also aware that they get to keep the amount of money not donated to any organizations if winning the lottery.

3.2.1 Selection of the societal causes.

3.2.2 System Architecture and Interface. The voting system is constructed using Python Flask for the back-end, Angular for front-end and MongoDB for database storage. The experiment source code is publicly available [1] and the QV interface is also provided as a stand-alone repository [2]. In this subsection, we focus on the QV interface.

The QV interface, shown in graph Y, consists of three major sections. The first section contains definitions of QV and the prompt of the task. The second section shows a list of option with a plus and minus button to its left. Buttons are disabled if the number of voice credit does not permit the next vote. A bar on the right of the option shows the proportion of voice credits used to that option. The final section is a floating summary at sticks to the bottom of the page. It contains a visualization of the total number of credits and the remaining credits.

3.3 Experiment 2

The second experiment extends upon the first one, in which it examines whether Quadratic Voting betters at aligning people's actual preferences compared to a Likert-scaled survey in an HCI setting. Different from political and public-opinion surveys, testing participants' preference in interface design and user experience is much more non-trivial. Thus we developed a buy-back mechanism

[1]Not yet public
[2]https://github.com/hank0982/QV-app

and observe participants' behaviors as their true preference. This experiment also acts as a concrete example as to how QV can be incorporated in HCI.

3.3.1 Choice of HCI Research Question. Research on video and audio quality from the lens of HCI has been a relatively mature. Contributions has been made to fields like multi-media conferencing [30], video-audio perception [5, 18] and more specifically trade-offs between video and audio elements under network monetary constraints [17, 21].

Oeldorf-Hirsch et al. [21] conducted a study, covering the widest range of elements to the best of our knowledge, to understand how users with bandwidth constraints made trade-offs between video and audio elements. They examined participants' attitude between three video bit rates, three video frame rates and two audio sampling rates across three types of video content. Participants were asked to rate the overall quality, video quality, audio quality and enjoyment level on a 5-point Likert scale in each condition. Conclusion were drawn using mean and standard deviation of the survey results. This is a typical study where the goal is to find 1 or some of the K elements to choose from when under constraint. In our second experiment, we expand this study to collect people's preference among a wider range of video and audio elements and compare how Likert-scaled survey and QV reflects people's true perception preferences.

3.3.2 Experiment 2 Design. In our experiment, we included a total of five video and audio element that will impact a video. These elements include video and audio package loss rate, determining whether the audio or video stutters; video resolution and audio sampling rate effecting the quality of video and audio; and video-audio synchronization. We selected a few segment of weather broadcasting from a news channel as the content of our video. Weather broadcasts usually convey information via both visual and audio channels, appeal to a wide array of audiences, and do not require prior knowledge to understand.

To ensure the ecological validity of the experiment, we situated the comparison of different video and audio elements in a hypothetical scenario in which the participant is a manager of a weather reporting news station. As the manager, the participant was asked to rate the importance of each video and audio elements with the goal to maximize customer understanding of the context where network is of low bandwidth and that the weather broadcast cannot be shown in its best quality.

We designed a between-subject study with three groups of 60 participants. After the participants agreed with the consent form, all three group of participant were presented an example weather broadcast segment with controls of the five video and audio elements under the video shown in figure M. All five elements were set to sub-optimal by default, making the content near incomprehensible. Participants can alternate the five elements in any combination, to see how elements impact to the video.

Once participants think that they had a grasp of how different elements impact a video, the first group of participants then completed a 5-point Likert-scaled survey while the second group of participants completed a QV survey with K voice credits [3], asking their opinions on the importance of the 5 video and audio elements in a weather broadcast under a low bandwidth environment. The third group of participants were asked to perform a buy-back task for a bad-quality advertising video.

The buy-back task mimics a rational customer's behavior: buying essential tools to complete some given task. Participants were told that as the manager of the weather broadcast agency, they need to verify if their viewer can understand the content of the video. Therefore, the goal of their task is to correctly answer a set of multiple choice questions to make sure that they correctly comprehend the video. Given the video with sub-optimal video, participants were given a

[3] K is decided from experiment 1

budget of $30 to purchase some or all of the features back. To ensure incentive-compatibility of the participants' buy-back actions, we offered to pay the participants their own remaining amount from the $30 budget through a lottery under the condition that they answered 80% of the multiple choice questions correctly, [missing probability] version of a new weather broadcast video adjusted by their buy-back choices. These questions contained factual questions such as, "What is the weather of Chicago?", "What is the highs and lows of San Diego", "Which of the follow cities got colder?". Participants were shown three example questions before the buy-back task to assist their decision. Participants can replay the video with their adjustments while answering the questions to ensure that participants do not require memorization. There will be a 5 minute timer to minimize the impact of replaying the video. With this design, participants would try their best to make the video comprehensible based on their opinions on which feature(s) was most needed at the lowest cost.

In the given weather broadcast video, there were 4 levels of quality for each of the 5 elements. By default, the video set to the lowest level for all elements before any adjustment occurred.

(1) Audio Package Loss Repaired with Silence (package loss rate) [30]: 20%, 10%, 5%, 0%
(2) Video Package Loss (package loss rate) [6]: 20%, 8%, 4%, 0% (20, 8.3, 3.3, 0)
(3) Audio Sampling Rate [20, 21]: 8kHz, 11kHz, 16kHz, 48kHz
(4) Video Resolution [13, 21]: 120x90, 168x126, 208x156, 240x180
(5) Video-audio Synchronization (time video behind audio) [28]: 240ms, 200ms, 160ms, 0ms (new: 1850, 1615, 1050, 0)

In the buy-back task, each level of improvement for one feature costs $2. It would cost the entire budget of $30 to buy all levels of every feature back. Hence, the option of buying back everything was given to the participants, in return, there would be no extra payoff remaining for the participant.

Similar to the first experiment, the money spent on each feature during the buy-back task are considered as the true preference the population had towards the 5 video and audio features. The results from the Likert-scaled surveys and QV survey were then compared to the population's true preference to see how different they were.

4 RESULTS

4.1 Experiment 1

4.1.1 report results...

REFERENCES

[1] 2018. 2018 Midterm Voters: Issues and Political Values. *Pew research center* (2018).
[2] 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 (2017), 173–190.
[3] Phillip A Bishop and Robert L Herron. 2015. Use and misuse of the Likert item responses and other ordinal measures. *International journal of exercise science* 8, 3 (2015), 297.
[4] 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. (2018).
[5] Sherry Y Chen, Gheorghita Ghinea, and Robert D Macredie. 2006. A cognitive approach to user perception of multimedia quality: An empirical investigation. *International Journal of Human-Computer Studies* 64, 12 (2006), 1200–1213.
[6] Mark Claypool and Jonathan Tanner. 1999. The effects of jitter on the peceptual quality of video. In *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*. ACM, 115–118.
[7] Diane R Edmondson. 2005. Likert scales: A history. In *Proceedings of the 12th conference on historical analysis and research in marketing (CHARM)*. 127–133.
[8] Jon X Eguia, Nicole Immorlica, Katrina Ligett, E Glen Weyl, and Dimitrios Xefteris. 2019. Quadratic voting with multiple alternatives. *Available at SSRN 3319508* (2019).
[9] Kraig Finstad. 2010. Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies* 5, 3 (2010), 104–110.

[10] Kate J Garland and Jan M Noyes. 2008. Computer attitude scales: How relevant today? *Computers in Human Behavior* 24, 2 (2008), 563–575.
[11] David R Hodge and David Gillespie. 2003. Phrase completions: An alternative to likert scales.(Note on Research Methodology). *Social Work Research* 27, 1 (2003), 45–56.
[12] Susan Jamieson et al. [n.d.]. Likert scales: how to (ab) use them. *Medical education* 38, 12 ([n. d.]), 1217–1218.
[13] 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 (2008), 145–166.
[14] 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.
[15] 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.
[16] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).
[17] 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.
[18] Andreea Molnar and Cristina Hava Muntean. 2015. Assessing learning achievements when reducing mobile video quality. *Journal of Universal Computer Science* 21, 7 (2015), 959–975.
[19] Ryan Naylor et al. 2017. First year student conceptions of success: What really matters? *Student Success* 8, 2 (2017), 9–19.
[20] Peter Noll. 1993. Wideband speech and audio coding. *IEEE Communications Magazine* 31, 11 (1993), 34–44.
[21] 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.
[22] A Ant Ozok. 2009. Survey design and implementation in HCI. *Human-Computer Interaction: Development Process* 253 (2009).
[23] Eric Pacuit. 2019. Voting Methods. In *The Stanford Encyclopedia of Philosophy* (fall 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
[24] Godfrey Pell. 2005. Use and misuse of Likert scales. (2005).
[25] Eric A Posner and E Glen Weyl. 2018. *Radical markets: Uprooting capitalism and democracy for a just society*. Princeton University Press.
[26] 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 (2017), 283–303.
[27] 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 (2015), 1.
[28] Ralf Steinmetz. 1996. Human perception of jitter and media synchronization. *IEEE Journal on selected Areas in Communications* 14, 1 (1996), 61–72.
[29] Lynn Vavreck et al. 2007. The exaggerated effects of advertising on turnout: The dangers of self-reports. *Quarterly Journal of Political Science* 2, 4 (2007), 325–343.
[30] Anna Watson and Martina Angela Sasse. 1996. Evaluating audio and video quality in low-cost multimedia conferencing systems. *Interacting with Computers* 8, 3 (1996), 255–275.