Organize, Then Vote: Exploring Cognitive Load in Quadratic Survey Interfaces

ANONYMOUS AUTHOR(S)*

Quadratic Surveys (QSs) elicit more accurate preferences than traditional methods like Likert-scale surveys. However, the cognitive load associated with QSs has hindered their adoption in digital surveys for collective decision-making. We introduce a two-phase "organize-then-vote" QS to reduce cognitive load. As interface design significantly impacts survey results and accuracy, our design scaffolds survey takers' decision-making while managing the cognitive load imposed by QS. In a 2x2 between-subject in-lab study on public resource allotment, we compared our interface with a traditional text interface across a QS with 6 (short) and 24 (long) options. Two-phase interface participants spent more time per option and exhibited shorter voting edit distances. We qualitatively observed shifts in cognitive effort from mechanical operations to constructing more comprehensive preferences. We conclude that this interface promoted deeper engagement, potentially reducing satisficing behaviors caused by cognitive overload in longer QSs. This research clarifies how human-centered design improves preference elicitation tools for collective decision-making.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; **Collaborative and social computing design and evaluation methods**; User studies; **HCI design and evaluation methods**; **Interactive systems and tools**; **Empirical studies in interaction design**.

Additional Key Words and Phrases: Quadratic Survey; Survey Response Format; User Interface; Preference Construction; Cognitive Load

ACM Reference Format:
Anonymous Author(s). 2024. Organize, Then Vote: Exploring Cognitive Load in Quadratic Survey Interfaces. *Proc. ACM Hum.-Comput. Interact.* 1, 1 (December 2024), 54 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Designing intuitive survey interfaces is crucial for accurately capturing respondents' preferences, which directly impact the quality and reliability of the data collected. Recent Human-Computer Interaction (HCI) studies highlight how certain survey response formats can increase errors [1, 2] and influence survey effectiveness [3]. In this paper, our goal is to introduce an effective interface for a **Quadratic Survey (QS)**, a survey tool designed to elicit preferences more accurately than traditional methods [4]. Despite the promise of QSs, there has been no research on designing interfaces to support their unique quadratic mechanisms [5], where participants must rank and rate items — a task that poses significant cognitive challenges. To popularize QSs and ensure high-quality data, this paper addresses the question: *How can we design interfaces to support participants in completing Quadratic Surveys (QSs) more effectively?*

We envision an effective interface that navigates participants through the complex mechanism and preference construction process, tailored to a QS. A QS improves accuracy in individual preference elicitation compared to traditional methods like Likert scales by requiring participants to make trade-offs using a fixed budget of credits, where purchasing k votes for an option in QS costs k^2 credits [6, 4]. This quadratic cost structure forces respondents to carefully evaluate their preferences, balancing the strength of their support or opposition against the limited budget. However,

Manuscript submitted to ACM

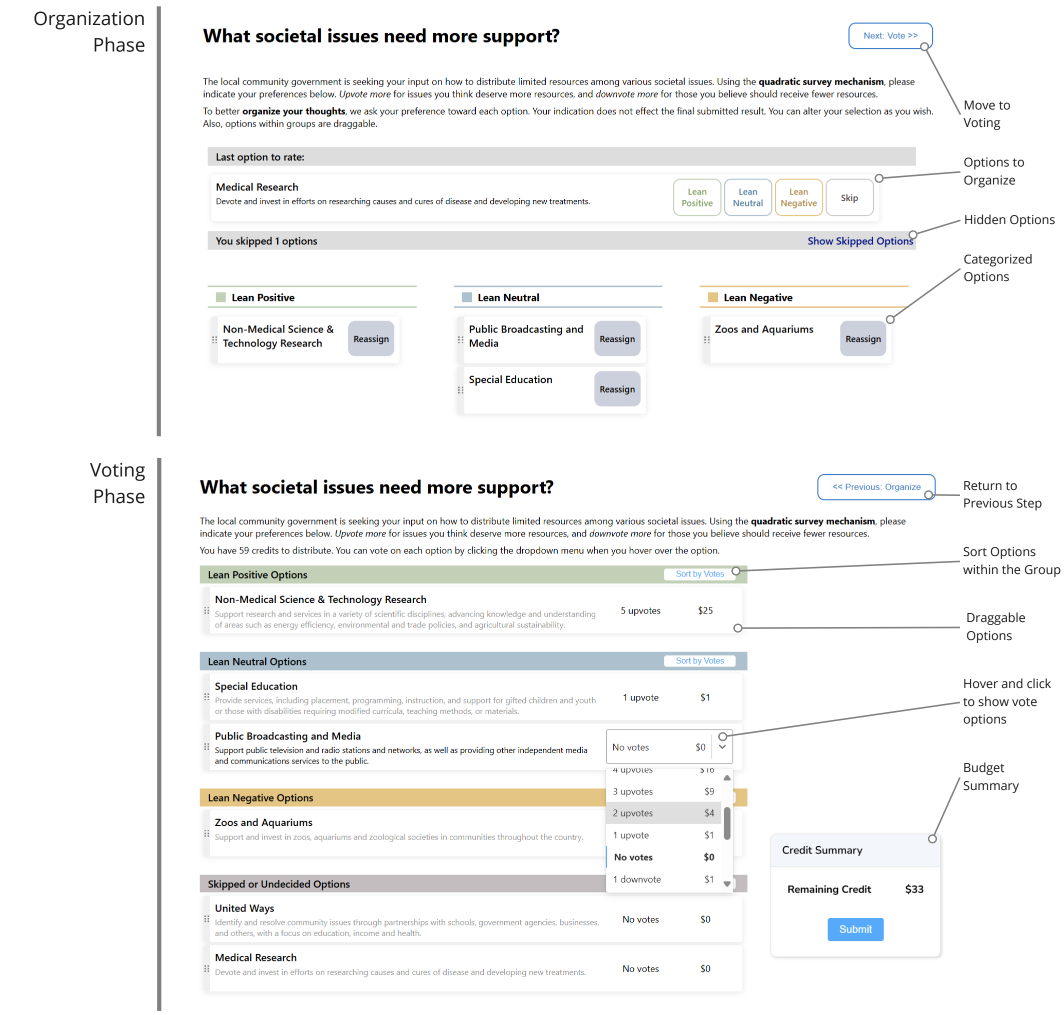


Fig. 1. The Two-Phase Interface: The interface consists of two phases. Survey respondents can navigate between phases using the top right button. In the organization phase, the interface presents one option at a time to the respondents, and they chose one of four positional choices: “Lean Positive”, “Lean Neutral”, “Lean Negative”, or “Skip”. Skipped options are hidden and can be evaluated later. The chosen options then appear below. Items can be dragged and dropped across categories or returned to the stack. In the voting phase, options are listed in the order of the four categories. When hovering over each option, respondents can select a vote for that option using a dropdown menu. Each dropdown menu contains the cost associated with the vote. A sort button allows ascending sorting within each category. A summary box tracks the remaining credit balance.

the process of making these thoughtful trade-offs introduces challenges. As individual preferences are often constructed when presented with the options [7], the act of weighing costs, evaluating options, and constructing rankings increases cognitive load. Moreover, QSs, often referred to as Quadratic Voting (QV) in voting scenarios, can involve hundreds of options [8, 9], increasing the risk of cognitive overload and the taking of mental shortcuts [10, 11, 12].

To date, existing quadratic mechanism-powered applications simply present options, allow vote adjustments and automatically calculate votes, costs, and budget usage. Such designs focused heavily on the mechanics operating the tool, rather than supporting possible challenges these application users faced. Survey interface literature, while addressing decision-making and usability, focuses on traditional surveys that do not share the unique option-to-option trade-offs that a QS introduces [13, 14, 15, 16, 17, 1]. Prior research in HCI and beyond explored techniques to manage cognitive load [18, 19, 16, 20, 21] and scaffold challenging tasks [22, 23, 24, 25] showing promise in supporting preference construction with a QS. Thus, this study aims to bridge this gap.

We propose a novel interactive two-phase "organize-then-vote" QS interface (referred to as the two-phase interface for short, Figure 1), which was developed through multiple iterations. It aims to facilitate preference construction and reduce cognitive load when making trade-offs through three key elements. First, the interface scaffolds the preference construction process by having participants initially categorize the survey options into "Lean Positive," "Lean Neutral," or "Lean Negative." This serves as a cognitive warm-up, easing participants into the more complex QS voting task. Second, the interface arranges the options according to these categorizations, providing a structured visual layout. Third, participants can refine the positions of these options using drag-and-drop functionality, giving them greater control and agency in the preference-construction process.

To explore how these interface elements affect cognitive load and support preference construction in QSs, we pose the following research questions:

- RQ1. How does the number of options in Quadratic Surveys impact respondents' cognitive load?
- RQ2a. How does the two-phase interface impact respondents' cognitive load compared to a single-phase text interface?
- RQ2b. What are the similarities and differences in sources of cognitive load across the two interfaces?
- RQ3. What are the differences in Quadratic Survey respondents' behaviors when coping with long lists of options across the two-phase interface and the single-phase text interface?

We invited 41 participants to a lab study comparing our two-phase interface with a baseline to understand how different interface designs and option lengths (6 options or 24 options) impact cognitive load.

Self-reported cognitive load using the NASA Task Load Index (NASA-TLX) and semi-structured interviews identified common challenges in QS, such as preference construction and budget management, while highlighting differences between text and two-phase interfaces. The two-phase interface fostered more strategic engagement with survey options, encouraging consideration of broader impacts in the long QS, reducing time pressure in the short QS, and eliciting greater affirmative satisfaction (e.g., "feeling good"). Quantitative results support these observations: participants in the two-phase interface—particularly in long surveys—traversed the list less frequently but maintained the same number of edits while spending more time per option. This suggests that reduced traversal did not diminish engagement. Together, these findings highlight the organizing phase's role in fostering deeper engagement with survey options.

Contributions. We contribute to the body of knowledge in the HCI community by proposing the first interface specifically designed for QS and QV-like applications, aimed at reducing cognitive challenges and scaffolding preference construction through a two-phase interface with direct manipulation. Before our work, no research had explored QS interfaces. This is particularly important for long QSs prone, which are prone to cognitive overload. Few studies in HCI address interfaces for surveys and questionnaires. Our study demonstrated how user interfaces can facilitate preference construction in situ and promote deeper engagement with survey options through interface elements. Additionally, this paper offers the first in-depth qualitative analysis of user experiences among Quadratic Mechanism applications, identifying usability challenges and key factors contributing to cognitive load. The impact of our contribution extends beyond QSs, offering design implications for other preference-elicitation tools in multi-option scenarios. By making QSs easier to use and more accurate, our design also encourages wider adoption among researchers and practitioners. Finally, our work lays the groundwork for future quadratic mechanisms interface design to better facilitate individuals in communicating their preferences.

2 Related Work

This research lies at the intersection of three core areas: quadratic surveys, existing QV interfaces and choice overload along with cognitive challenges. In this section, we review the related works in each of these areas.

2.1 Quadratic Survey and the Quadratic Mechanism

We introduce the term **Quadratic Survey (QS)** to describe surveys that utilize the quadratic mechanism to collect individual attitudes. The **quadratic mechanism** is a theoretical framework designed to encourage the truthful revelation of individual preferences through a quadratic cost function [5]. This framework gained popularity through **Quadratic Voting (QV)**, also known as plural voting, which uses a quadratic cost function in a voting framework to facilitate collective decision-making [26].

To illustrate how QS works, we formally define the mechanism: each survey respondent is allocated a fixed budget, denoted by B, to distribute among various options. Participants can cast n votes for or against option k. The cost c_k for each option k is derived as:

$$c_k = n_k^2 \quad \text{where} \quad n_k \in \mathbb{Z}$$

The total cost of all votes must not exceed the participant's budget:

$$\sum_k c_k \leq B$$

Survey results are determined by summing the total votes for each option:

$$\text{Total Votes for Option } k = \sum_{i=1}^{S} n_{i,k}$$

where S represents the total number of participants, and $n_{i,k}$ is the number of votes cast by participant i for option k. Each additional vote for each option increases the marginal cost linearly, encouraging participants to vote proportionally to their level of concern for an issue [27].

QS adapts these strengths of the quadratic mechanism in *voting* to encourage truthful expression of preferences in *surveys*. Unlike traditional surveys that elicit either rankings *or* ratings, QS allows for *both*, enabling participants to cast multiple votes for or against options, incurring a quadratic cost. Cheng et al. [4] showed that this mechanism aligns individual preferences with behaviors more accurately than Likert Scale surveys, particularly in resource-constrained scenarios like prioritizing user feedback on user experiences.

In recent years, empirical studies on QV have expanded into various domains [28, 29]. Applications based on the quadratic mechanism have also grown, including Quadratic Funding, which redistributes funds based on outcomes from consensus made using the quadratic mechanism [30, 31]. Recent work by South et al. [32] applies the quadratic mechanism to networked authority management, later used in Gov4git [33]. Despite the increasing breadth and depth of applications utilizing the quadratic mechanism, little attention has been paid to user experience and interface design, which support individuals in expressing their preference intensity. Our work aims to address this by designing interfaces supporting quadratic mechanisms.

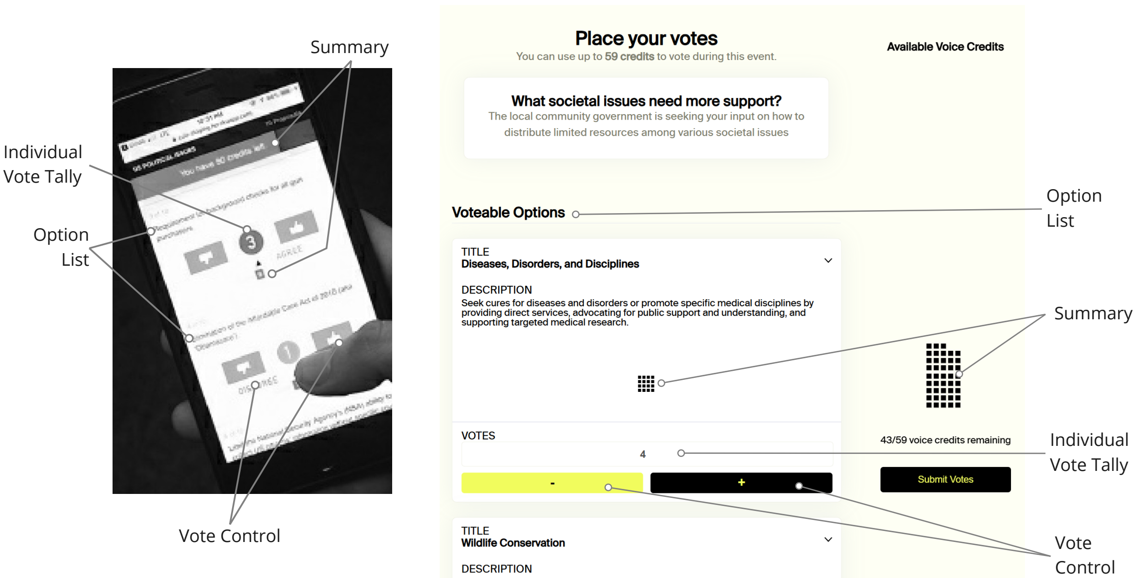


Fig. 2. A selection of two QV interfaces. The interface on the left was used in the first empirical QV research [6]. Little information is available about the software, except for an image from Posner and Weyl [27]. The interface on the right is an open-sourced QV interface [34] forked from GitCoin [35], used by the RadicalxChange community [36]. Both interfaces share the common elements with different visual representations.

2.2 Existing QV Interfaces

Since QS shares QV's underlying mechanism, we used snowball sampling to identify publicly available QV applications mentioned in news and academic sources. Currently, no widely adopted QV interface is tied to a single vendor or platform. Fig. 2 shows two variations of existing interfaces, with all QV interfaces employing a single-step approach with different visual representations of common elements. [33, 37, 4, 38] All QV interfaces generally include:

- Option list: A list of options for voting.
- Vote controls: Buttons to increase or decrease votes for each option.
- Individual vote tally: A display of the votes cast per option.
- Summary: An auto-generated summary of costs and the remaining budget.

These components let users operate QV mechanically, providing little understanding of voters' usability needs nor offering cognitive support. In addition, HCI research on survey interfaces is limited [39, 40] with most efforts focusing on alternative input modalities like bots, voice interfaces, or virtual reality [41, 42, 2, 43].

2.3 Cognitive Challenges and Choice Overload

The challenge of respondents making difficult decisions using quadratic mechanisms remains unexplored in the literature. Lichtenstein and Slovic [7] identified three key elements that make decisions difficult. These elements include making decisions in unfamiliar contexts, quantifying the value of one's opinions, and being forced to make trade-offs due to conflicting choices. QS fits at least two of the three elements: participants may encounter a selection of unfamiliar options by the survey designer; they are asked to quantify the difference between option preferences through a numerical vote; and the budget constraint enforces trade-offs under a non-linear function, which means that a vote decrease for one option is not necessary equivalent to an increase for another, making iterative adjustment and evaluating tradeoffs difficult. Thus, we believe QS introduces a high cognitive load.

Cognitive load refers to the demands placed on a user's working memory during the interaction process, which significantly influences the usability of the system [44, 45]. Cognitive overload can adversely affect performance [46], leading individuals to rely on heuristics rather than deliberate, logical decision-making [47]. When presented with excessive information, such as too many options, individuals 'satisfice', settling for a 'good enough' solution rather than an optimal one [10, 11, 12]. Subsequently, too many options can overwhelm individuals, resulting in decision paralysis, demotivation, and dissatisfaction [48].

Additionally, Alwin and Krosnick [49] highlighted that the use of ranking techniques in surveys can be time-consuming and potentially more costly to administer. These challenges are compounded when ranking numerous items, requiring substantial cognitive sophistication and concentration from survey respondents [50].

Notable applications of QV include the 2019 Colorado House, which considered 107 bills [51], and the 2019 Taiwan Presidential Hackathon, which featured 136 proposals [52]; both used a single QV question with hundreds of options. These empirical applications of QV suggest the importance of understanding QS with many options' impact on cognitive load and support developing interfaces for practical uses.

3 Quadratic Survey Interface Design

This section presents our QS interface. Drawing on existing QV interfaces described in Section 2.1 and prior literature, we iterated through paper prototypes and three design pre-tests, detailed in Appendix A. Initially, participants struggled to *rank* relative preferences among options and *rate* the degree of trade-offs between them. In this study, we focus on addressing the former challenge, which pertains to preference construction.

3.1 'Organize-then-Vote': The Two-Phase Interface

3.1.1 Justifying a two-phase approach. The main objective of the two-phase interface is to facilitate preference construction and reduce cognitive load. As shown in Figure 1, the interface consists of two steps: an organization phase and a voting phase. In both phases, survey respondents can drag and drop options across the presented list.

A two-phase approach. Preferences are shaped through a series of decision-making processes [7]. Two decision-making theories inspired this two-step interaction interface design: Montgomery [53]'s Search for a Dominance Structure Theory (Dominance Theory) and Svenson [54]'s Differentiation and Consolidation Theory (Diff-Con Theory). The former suggested that decision-makers prioritize creating dominant choices to minimize cognitive effort by focusing on evidently superior options [53]. The latter described a two-phase process where decisions are formed by initially *differentiating* among alternatives and then *consolidating* these distinctions to form a stable preference [54]. Pre-tests showed participants puzzled by ranking all options before voting. These theories suggest decisions emerge by eliminating choices, not by fully ranking them. Therefore, the organize-then-vote design makes this natural process more explicit. Phase one focused on differentiating and identifying dominant options, enabling survey respondents to preliminarily categorize and prioritize their choices. Phase two presented these categorized options in a comparable manner, with drag-and-drop functionality, enhancing one's ability to consolidate preferences. This structured approach aimed to construct a clear decision-making procedure that reduced cognitive load and enhanced clarity and confidence in the decisions made.

Phase 1: Organization Phase. The goal of the organization phase was to support participants in identifying clearly superior options or partitioning choices into distinguishable groups. In this section, we first describe how the interaction works, then we detail the reasons for the implemented design decisions.

The organizing interface, depicted on the top half of Figure 1, sequentially presents each survey option. Participants select a response among three ordinal categories – "Lean Positive", "Lean Negative", or "Lean Neutral". Once selected, the system moves that option to the respective category. Participants can skip the option if they do not want to indicate a preference. Options within the groups are draggable and rearrangeable to other groups should the participants wish.

To support preference formation, respondents are shown one option at a time, allowing them to either recall a prior judgment or construct a new one based on the presented choices [55]. Limiting the information presented this way also helps reduce cognitive load by preventing overload from too many options [56]. This incremental process ensures that participants form opinions on individual options.

The three possible options — Lean Positive, Lean Neutral, and Lean Negative — aim to scaffold participants in constructing their own choice architecture [57, 58], which strategically segments options into diverse and alternative choice presentations while avoiding biases from defaults. We believed that these three categories were sufficient for participants to segment the options. We do not limit the number of options one can place in each category to prioritize user agency, allowing participants full control over how they organize their preferences [59]. Immediate feedback displays the placement of options and allows participants to rearrange them via drag-and-drop, adhering to key interface design principles [59]. At the same time, it allows finer-grain control for individuals to surface dominating options and create differentiating groups of options.

Phase 2: Interactive Voting Phase. The objective of the voting phase is to facilitate the consolidation of differentiated options through interactive elements while reinforcing the differentiation across options constructed by participants in the previous phase. This facilitation is achieved by retaining the drag-and-drop functionality for direct manipulation of position and enabling sorting within each category.

Options are displayed as they are categorized within each category from the previous step and in the following section — Lean Positive, Lean Neutral, Lean Negative, and Skipped or Undecided — as detailed on the bottom half of Figure 1. The Skipped or Undecided category contains options left in the organization queue, possibly because survey respondents have a pre-existing preference or chose not to organize their thoughts further. The original order within these categories is preserved to maintain and reinforce the differentiated options. This ordering sequence mitigated early prototype concerns where uncategorized options were left at the top of the voting interface confusing survey respondents. Respondents have the flexibility to return to the organization interface at any point during the survey to revise their choices.

In the voting interface, options are draggable, allowing participants to modify or reinforce their preference decisions as needed. Each category features a sort-by-vote function for reordering within the group, which, although it doesn't affect the final outcome, supports information organization and consolidation. Both features aim to group similar options automatically and emphasize proximity, reducing cognitive load by following the proximity compatibility principle to enhance decision-making [60].

While multiple interaction mechanisms exist, drag-and-drop has been extensively explored in rank-based surveys. For instance, Krosnick et al. [61] demonstrated that replacing drag-and-drop with traditional number-filling rank-based questions improved participants' satisfaction with little trade-off in their time. Similarly, Timbrook [62] found that integrating drag-and-drop into the ranking process, despite potentially reducing outcome stability, was justified by the increased satisfaction and ease of use reported by respondents. The trade-off was deemed worthwhile as QSs did not use the final position of options as part of the outcome if it significantly enhanced user satisfaction and usability [63]. Together, these design decisions led to our belief that a two-phase interface with direct interface manipulation could reduce the cognitive load for survey respondents to form preference decisions when completing QSs.

In addition, we made three aesthetic design decisions considering existing QV-based interfaces. First, we removed visual elements like icons, emojis, progress bars, and vote visualizations, as prior research indicated that emojis could influence survey interpretations and reduce user satisfaction [64, 16]. While effective visualizations can aid decision-making, this study does not aim to address that question. Second, all options are visible on the screen simultaneously. Prior research recommends placing all items on the voting screen to prevent overlooked votes [65]. This echoes the proverb "out of sight, out of mind," reducing where individuals might be biased toward visible options, and additional effort is required for individuals to retrieve specific information if options are hidden. Last, use a dropdown positioned to the right of each survey option for ease of access to the budget summary when determining the votes. The layout of the votes and cost was inspired by online shopping cart checkout interfaces where quantities are supplied next to the itemized costs followed by the total checkout amount. Figure 3 shows the two alternatives—click-based buttons (participants disliked multiple clicks) and a wheel-based design (unfamiliar to some)—and settled on the dropdown.

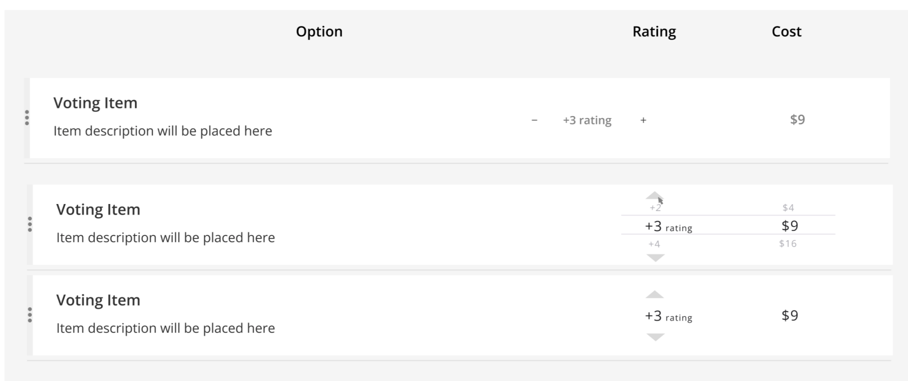


Fig. 3. Alternative vote control. The click-based design (upper) mirrors traditional vote control used in other QV interfaces, where each click controls one vote. The wheel-based design (the latter two) allows control through both clicks and mouse wheel rotation.

3.2 Baseline Interface: Single-Phase Text Interface

We created a single-phase text interface (referred to text interface for short, Figure 4) as a control, enabling us to see how organizational features affect cognitive load and behavior. Like existing interfaces, it uses static lists, a summary box, and a vote control. To ensure a fair comparison, we applied the same design principles: no extraneous visuals, all

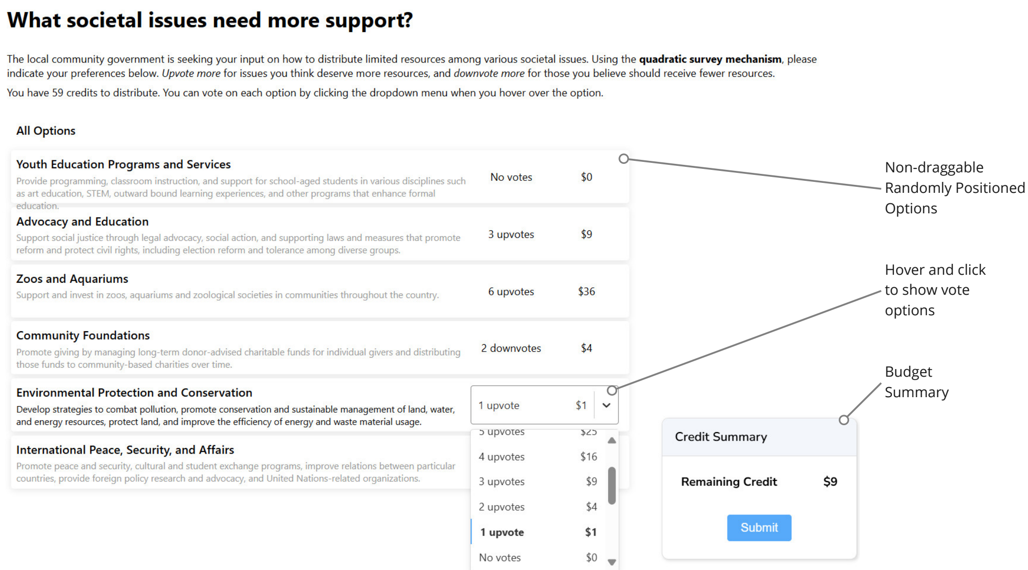


Fig. 4. The text-based interface: This interface is based on the two-phase version but does not include the organization phase and lacks the drag-and-drop functionality. Options are randomly positioned.

options on one screen, and dropdown-based voting. The prompt appears at the top, followed by a randomly ordered list to prevent ordering bias [66, 67]. Costs and the credits summary appear on the right.

Both experimental interfaces were developed with a ReactJS frontend and a NextJS backend powered by MongoDB. We open-source both interfaces.[1]

4 Experiment Design

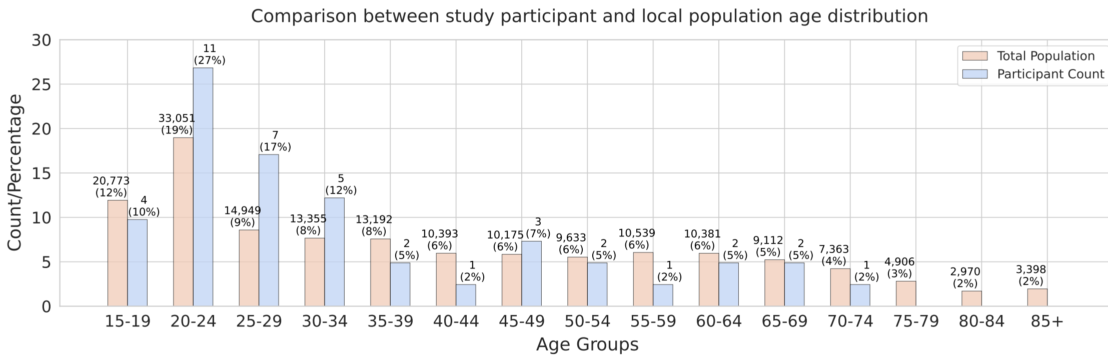
In this section, we describe our experiment design. The study was approved by the university's Institutional Review Board (IRB).

4.1 Recruitment and Participants

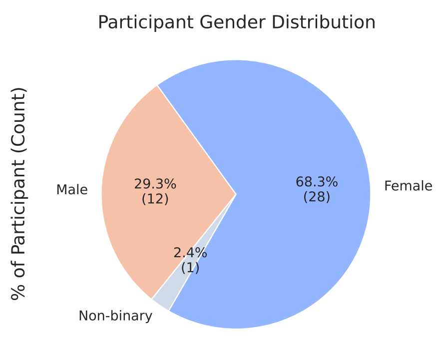
We recruited 41 participants from a United States college town using online ads, digital bulletins, social media posts, email newsletters, and physical flyers in public spaces beyond campus. We advertised the study as focusing on societal attitudes to mitigate potential response bias. One participant was excluded due to data quality concerns[2].

To ensure diversity, we prioritized non-students by selectively accepting them and monitoring demographic distribution. The mean participant age was 34.63 years, with an age distribution similar to the county's demographic profile (Figure 5a), although there was a slightly higher representation of younger adults. Gender and race demographics are presented in Figures 5b and 5c. Demographic differences between groups were reasonably balanced, although

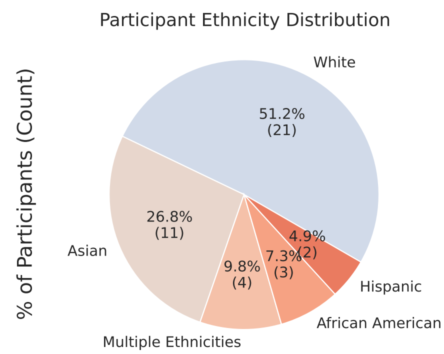
[1]link-to-github
[2]The participant reported not completing the survey seriously, as they believed the experiment was fake.



(a) Age distribution of the study participants were similar to the locale's demographic profile.



(b) Gender distribution of our participants skewed towards female participants.



(c) Ethnicity distribution remains diverse with fewer Hispanic and African American participants.

Fig. 5. Demographic distributions: Age, Gender, and Ethnicity

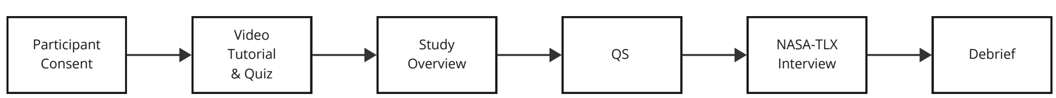


Fig. 6. Study protocol: Participants are asked to learn about the mechanism of QSs after consenting to the study. The researcher explained the study overview and asked participants to complete the QS. A NASA-TLX survey followed by interviews to understand participants' cognitive load. We debriefed participants after the study.

participants using the short text interface skewed slightly younger ($\mu = 32.1$), and those in the long two-phase interface group had a broader age range ($\mu = 38.8$, $\sigma = 19.6$). Full details are provided in Appendix C.

4.2 Experiment Design

We implemented a between-subject design to avoid learning effects and minimize participants' fatigue from potential complexity of QSs. The experiment focused on public resource allotment, following the methodology of Cheng et al. [4], in which participants expressed preferences across societal issues. These issues are relevant to all citizens and effectively highlight the need to prioritize limited public resources. Participants received a survey with options randomly drawn from the 26 societal topics[3] evaluated by Charity Navigator [68], an organization that assesses over 20,000 charities in the United States. Randomly selecting the options each participant saw aimed to control for potential systematic content biases introduced by specific voting options across surveys of different lengths. Participants were randomly assigned to one of four groups:

- Short Text (ST): A text interface with 6 options. ($N = 10$)
- Short Two-Phase (S2P): A two-phase interface 6 options. ($N = 10$)
- Long Text (LT): A text-based interface 24 options. ($N = 10$)
- Long Two-Phase (L2P): A two-phase interface with 24 options. ($N = 10$)

The choice of 6 and 24 options, representing short and long lists, was guided by prior research. Studies recommend fewer than 10 options for constant-sum surveys [69] and fewer than 7 for the Analytic Hierarchy Process [70]. Classic cognitive load research [71, 72] suggests the use of 7 ± 2 items. A meta-analysis by Chernev et al. [73] identified 6 and 24 as common values for short and long lists in choice overload studies, which are rooted in the original experiment by Iyengar and Lepper [48].

4.3 Experiment Procedure

Participant's spent on average 40 minutes (range: $27 - 68$, $\sigma = 9$) in the lab. Figure 6 visually represents the study protocol detailed in the following subsections.

4.3.1 Consent, Instructions, and Quiz. Participants were invited to the lab to control for external influences and used a 32-inch vertical monitor to display all options. After consenting, participants watched a video explaining the quadratic mechanism without any mention of the interface's operation, followed by a quiz to ensure understanding. Participants rewatched the video or consulted the researcher until they successfully selected the correct answers. Each participant's screen was captured throughout the study.

4.3.2 Quadratic Survey. The researcher informed participants that the study aimed to help local community organizers understand preferences on societal issues to improve resource allocation. Aware that their screens were being recorded, participants completed the survey independently inside a semi-enclosed space in the lab, without the researcher's presence. Once they completed the survey, participants notified the researcher.

4.3.3 NASA-TLX Survey and Interview. The researcher joins study participant with a paper-based weighted NASA Task Load Index (NASA TLX), followed by a semi-structured interview after being informed that the researcher would begin audio recording with their laptop. We adopted the paper-based weighted NASA Task Load Index (NASA TLX), a widely used multidimensional tool that averages six subscale scores to measure overall workload after task completion [74, 75, 76]. NASA-TLX is favored for its low cost and ease of administration [77], and it exhibits less variability compared to one-dimensional workload scores [78], making it suitable for our study. While cognitive load can be assessed through

[3]See Appendix D for the full list.

performance, psychophysiological, subjective, and analytical measures [77], the length and complexity of QSs make some of these impractical. Performance and analytical measures require task switching or interruptions, which risk increasing overall cognitive load and experiment time. Psychophysiological measures, such as pupil size [79] and ECG [80], are costly, sensitive to external factors, and often require participants to wear additional equipment.

4.3.4 Demographic, Debrief, and Compensation. After the interview, the researcher collected participant's demographics and debriefed them, explaining that the study's goal was to understand interface design and cognitive load. Participants received a $15 cash compensation.

5 Result: Self-Reported Cognitive Load in Quadratic Surveys

This section presents findings on cognitive load in QSs, focusing on how the number of options and different interfaces influence it (**RQ1**, **RQ2a**). We analyze similarities and differences in cognitive load sources across conditions (**RQ2b**).

Qualitative findings are based on an inductive thematic analysis [81], conducted after transcribing the interviews. Snippets were coded according to the research questions and merged into overarching themes. Differences across conditions were refined and validated using a deductive coding process.

Quantitative findings are derived from a Bayesian approach, which enhances transparency by interpreting posterior distributions and moving beyond binary thresholds [82]. Bayesian methods suit various sample sizes, leveraging maximum entropy priors to ensure conservative and robust inferences [83].

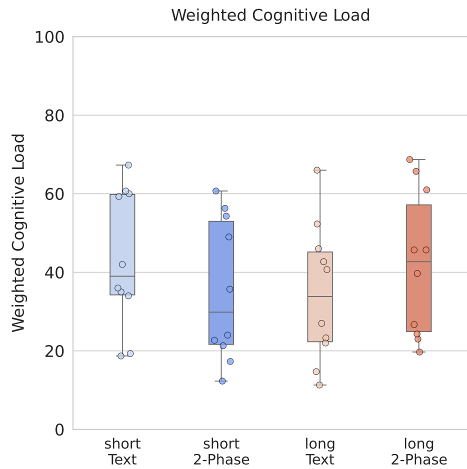
5.1 Overall Cognitive Load from NASA-TLX

Weighted NASA-TLX uses a continuous 0-100 score, with higher values denoting greater cognitive load. We use predefined mappings of NASA-TLX scores to cognitive levels: low, medium, somewhat high, high, and very high, as described by Hart and Staveland [74]. Results are shown in Figure 7, with value interpretations presented in Figure 7b.

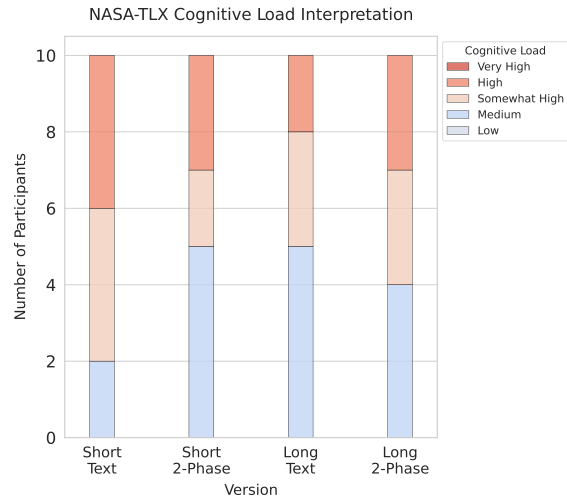
Given the sparsity of the data, we modeled the weighted NASA-TLX scores using cognitive levels as ordinal outcome variables. Then, we developed a hierarchical Bayesian ordinal regression model to analyze ordinal response data. The model includes length as an ordinal predictor, and interface type as a categorical predictor modeled with hierarchical priors to allow partial pooling across categories. Interaction effects between length and interface are captured using a non-centered parameterization constrained by an LKJ prior to account for correlations [83]. We use the same model for the NASA-TLX subscales. Given that subscales do not have cognitive level interpretations, we constructed weighted bins to facilitate the ordinal regression model. We present details of this model and additional subscale results in Appendix H.

In Bayesian analysis, the 94% high-density interval (HDI) represents the range where the true parameter is most likely to lie. While the results (Figure 8) are not statistically significant because 0 is within this range, the HDI quantifies probabilistic trends and accounts for uncertainty in a transparent manner.

- Increased option length with text interface trends to *reduced* cognitive load with a posterior probability of approximately 69.8%. This reflects a median cognitive load of 33.85 (mean = 34.60, SD = 17.69) compared to a median of 39.00 (mean = 43.23, SD = 17.65).
- Within short QSs, the two-phase interface trends to *reduced* cognitive load, with a posterior probability of 71.7% supporting the reduction. Participants report a median cognitive load of 29.85 (mean = 35.36, SD = 18.17) under the two-phase interface compared to a median of 39.00 (mean = 43.23, SD = 17.65) under the text interface.



(a) NASA-TLX Weight Score: The Long Two-Phase Interface exhibits the highest weighted cognitive load with a median of 42.70, a mean of 42.02. This is higher than the long text interface, which has a median cognitive load of 33.85 and a mean of 34.60. However, the short text interface demonstrates a higher cognitive load with a median of 39.00, a mean of 43.23, compared to the short two-phase interface, which has a median of 29.85, a mean of 35.36. The standard deviation is similar across groups at around 18.



(b) NASA-TLX Cognitive Interpretation: More participants in the short text interface, totaling 8, reported a somewhat high or above cognitive load, which is significantly higher compared to the 5 participants who reported similarly for the short two-phase interface. However, the long two-phase interface saw slightly more participants, 6 in total, reporting somewhat high or above cognitive load compared to the long text interface.

Fig. 7. This figure shows the box plot results for weighted NASA-TLX scores across experiment groups and participant counts based on individual score interpretations. In 7a, we observe a downward trend in cognitive load for the short QS, while the long QS shows an upward trend. Interestingly, there is a counterintuitive downward trend between short and long text interfaces. In 7b, these trends are clearer when NASA-TLX scores are grouped into five tiers.

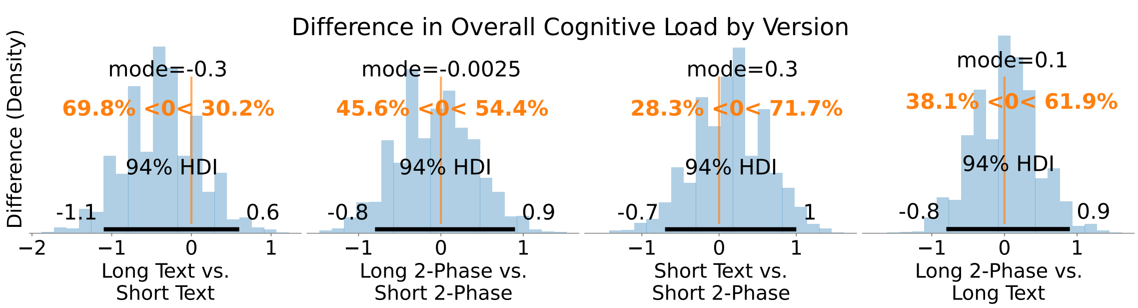


Fig. 8. The figure shows the contrast distribution of the average posterior ordinal category between experimental conditions. **The main takeaway:** while our Bayesian model does not indicate statistically significant differences, longer text interfaces are more likely to reduce cognitive load, and the two-phase interface has a higher probability of lowering cognitive load.

- For the long QSs, there trends an *increase* in cognitive load with a posterior probability of 61.9%. The median cognitive load is 42.70 (mean = 42.02, SD = 18.48) under the two-phase interface compared to 33.85 (mean = 34.60, SD = 17.69) in the text interface.

This result contradicts our hypothesis that more options would increase cognitive load and that interfaces can reduce it. Thus, we explore qualitative results to identify possible explanations. To understand the similarities and differences in sources of cognitive load (**RQ2b**), we analyze qualitative results across the six NASA-TLX subscales: mental demand, physical demand, temporal demand, effort, frustration, and performance. Detailed breakdown of each subscale are provided in Appendix E.

5.2 Qualitative Analysis: Common Sources of Cognitive Load

Our analysis reveals several themes across different cognitive load subscales. We identify three themes common to all experimental conditions.

Preference Construction is cited by 97.5% (N=39) of participants as a significant source of mental demand, consistent with prior literature suggesting that preferences are often constructed in context rather than fixed [7]. Specific tasks contributing to this demand include evaluating the relative importance between options (e.g., `S002` *Figuring out[...] how much I prioritize option 1 over option 2* , 40% ($N = 16$)), making trade-offs due to limited resources (e.g., `S005` *[...] very hard to take decisions ... I felt that multiple options deserve equal amounts of credit ... but you have given very limited credit.* , 42.5% ($N = 17$)), and deciding the exact number of votes (e.g., `S023` *[...] having to pick how many upvotes would go to each one* , 70% ($N = 30$)).

Budget Management emerges as a source of both mental and temporal demand. 25% (N=10) of participants describe the challenge of working with limited credits while trying to maximize their allocation (e.g., `S032` *[...] for certain societal issues, you had to ... take away from other issues you could support*). An equal percentage of participants find it mentally taxing to keep track of remaining credits (e.g., `S006` *[...] looking at the remaining credits, I'm trying to mentally divide that up before I start allocating*).

Operational Actions refer to reactive efforts addressing immediate, tactical needs. These actions involve direct task execution, responding to constraints without reflection on broader, long-term implications. Examples include adjusting choices to stay within budget (e.g., `S003` *I had to alter [...] I kept going under budget*), re-reading options (e.g., `S010` *I just had to reread it again*), completing questions efficiently (e.g., `S010` *I was trying to be efficient in responding to the question*), and interacting with the survey interface (e.g., `S018` *Like (deciding) one upvote or two upvotes[...]*). 40% (N=16) of participants attribute Operational actions to temporal demand. Additionally, 37.5% (N=15) attribute this cause to frustration, and 32.5% (N=13) attribute it to performance. While this is a commonly cited source across experiment conditions, there are different distributions.

5.3 Qualitative Analysis: Different Sources of Cognitive Load

There are several notable differences between the text and two-phase interfaces.

First, regardless of length, when analyzing performance, which refers to a person's perception of their success in completing a task, participants describe their performances differently. We categorize them into indications of satisficing behaviors("good enough"), exhausting their effort (i.e., "done their best,"), or feeling positive (i.e., "feeling good.") There are almost twice as many participants using the two-phase interface to report a positive feeling about their final submission (55% v.s 30% (N=11 vs. 6)).

Second, 70% (N=14) of text interface participants attribute operational actions as contributors to effort, double the percentage observed in the two-phase interface group (35%, N=7). This partially echoes the finding that 90% (N=18) of text interface participants report mental demand from deciding the exact number of votes, compared to 60% (N=12) in the two-phase interface group.

The distinction between the text and two-phase interfaces becomes more pronounced in the context of the long survey. 80% of the long text interface participants (N=8) attribute operational actions to effort, compared to only 20% (N=2) in the long two-phase interfaces. Conversely, 90% of long two-phase interface participants (N=8) attribute effort to strategic actions, compared to 50% (N=5) in the text interface.

We also found differences in how preference construction differs in contributing to their mental demand and sources of effort. Opposite to operational actions, **strategic considerations** refer to considering about long term goals, determining priorities, considering broader implications, and considering option's more holistically.

reflective decisions oriented toward long-term goals. They focus on determining priorities, considering broader implications, and aligning actions with overarching objectives. Consider the following quotes:

> *Trying to figure out what upvotes I should give [...] went back and forth between those two. [...] it was very mentally tasking for me.* `S015 (LT)`

> *[...] really having to think, especially with so many different societal issues. How do I personally prioritize them? And to what extent do I prioritize them?* `S009 (L2P)`

`S015` describes the operation of locating tasks to find the right vote, in contrast to `S009`'s focus on aligning higher-order values holistically. Regarding mental demand, 80% of participants in the long text interface focused on a narrower scope, comparing fewer options ($N = 8$), while only 30% did so in the two-phase interface ($N = 3$). Conversely, 90% of participants in the long two-phase interface considered broader societal impacts and evaluated more options simultaneously ($N = 9$), compared to 30% in the text interface ($N = 3$). Similar distinctions were evident in sources related to effort.

These differences highlight variations in **levels of engagement** with the survey content. Participants using the two-phase interface expressed higher satisfaction with their performance. For the long survey, they engaged with broader aspects across different options and strategically allocated their credits.

5.4 Qualitative Analysis: Instances of Satisficing

When individuals cannot process all available information, prior research has found that people exhibit *satisficing behaviors*, which refers to settling for *good enough* rather than *optimal* decisions [84]. While we did not explicitly ask participants if they 'satisficed,' nor did we measure it quantitatively, we identified satisficing behaviors based on participants' explanations of how they completed the survey. For example,

> *[...] you thought of enough things, you know, and so it wasn't the most effort I could put in because again, that would have been diminishing returns. I tried to think of enough things [...] and then move on. [...]* `S032 (ST)`
>
> *I felt like that (the response) was satisfied, but not perfect. Cause perfect is not a reality.* `S036 (ST)`

This quote illustrates satisficing decision-making, where participants chose to settle for suboptimal outcomes.Satisficing was observed primarily at the beginning and end of the survey, where participants allocated large amounts of credit initially and then managed the remaining credits to confirm their final vote allocations. For instance,

> *[...] Because that (the credit) was what was left. [Laughter] I probably wouldn't use that on <optionA> instead of <optionB>. [...]* `S015 (LT)`
>
> *I tried to use them [...] it went negative, and then I just settled for just $6 remaining. [...] I don't think it's perfect. But I think I'm satisfied. Yeah, I'm satisfied.* `S033 (LT)`
>
> *[...] when I had first started like looking at the first few, I was just doing it kinda like willy nilly, I'm not really paying that much attention to necessarily how many credits I had, or how many categories there were.* `S041 (LT)`

Participants also exhibited satisficing behaviors regarding *defaults*, particularly when constructing their preferences. For example, participant S003, described how default placements influenced their final decisions:

> *Honestly, if medical research [...] was the first one I saw, I think it would automatically give it a lot more.* S003 (ST)

Our qualitative analysis found that 60% of short-text participants ($N = 6$) and 50% of long-text participants ($N = 5$) expressed instances of satisficing behaviors when describing how they completed the survey, compared to none of the short two-phase participants and 30% of long-text participants ($N = 3$). These qualitative results highlighted potential satisficing behavior from QS participants.

6 Clickstream data: Interface reduces edit distance in long surveys

Following our findings on cognitive load, we analyze voting behaviors to identify differences in how participants cope with survey lengths, how interfaces influence their behavior, and why the long text interface might exhibit lower cognitive load. All data are publicly available[4] to ensure transparency and support further research. This measure reveals trends in participants' navigation and engagement with survey options. We examine three dimensions of this measure: edit distance per option, edit distance per action, and cumulative edit distance throughout the survey.

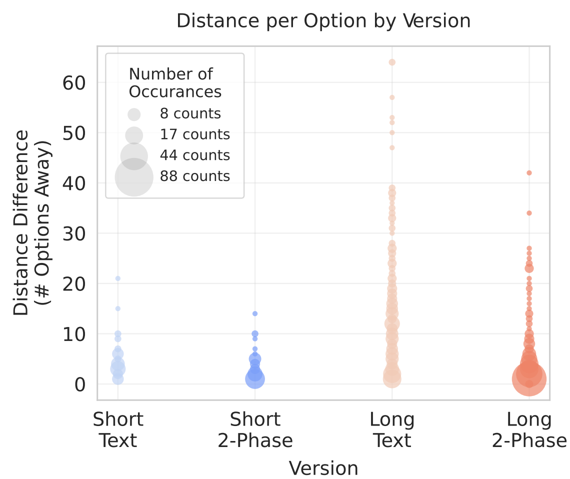


Fig. 9. Edit Distance Per Option: We sum the total number of edit distances for each option, with the figure using the radius to indicate how often a specific edit distance occurred within an experimental condition. **The main takeaway:** Participants in the two-phase interface completed their votes for more options with fewer edit distances, whereas the Long Text interface shows a long tail of options requiring a wider range of edit distances.

Edit distance per option: We sum up all the distances a participant moves while adjusting values for a single option. Figure 9 illustrates differences across the four experimental conditions, with the long text interface showing the largest variance in the distance traveled and the highest mean. We implement a hierarchical Bayesian framework to model edit distance differences across experimental conditions. The observed distance differences are modeled using an exponential distribution, where the scale parameter is linked to survey length (treated as an ordinal variable), interface type (treated as a categorical variable), interaction effects between length and interface, and controlling for individual user variability. The linear predictor includes a global intercept and slope for length, random effects for each interface

[4]link-to-github

condition with an LKJ prior that captures the correlations among interface categories, and user-specific random effects to account for individual heterogeneity. Detailed mathematical formulations of the model are provided in Appendix J.1.

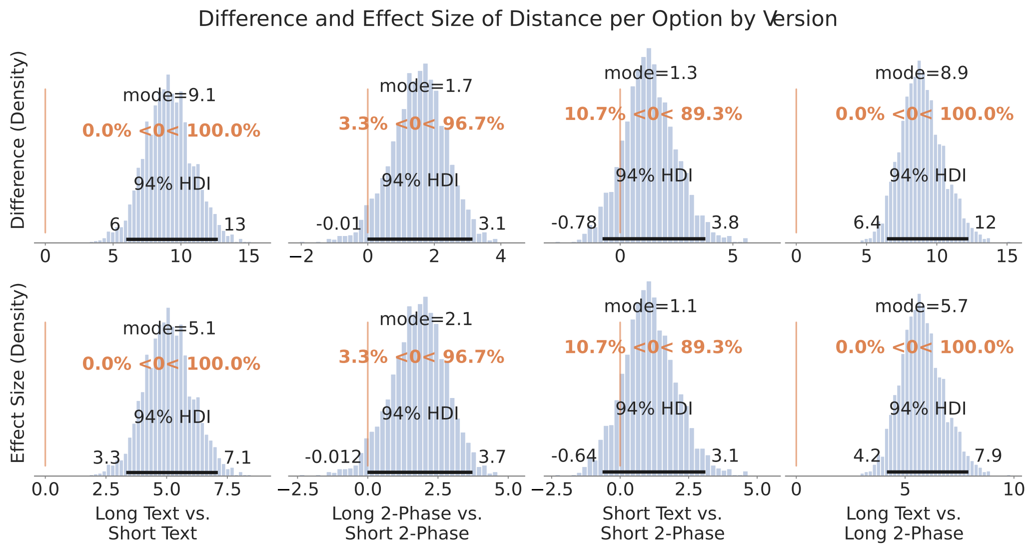


Fig. 10. The figure shows the contrast distributions of the mean edit distance per option between pairwise experimental conditions, with the first row representing absolute differences and the second row depicting effect sizes. **The main takeaway:** is that participants in the long text estimated more edit distance per option compared to those in the short text and the long two-phase condition. Notably, the long two-phase interface required estimated only slightly more edit distances despite the longer survey length.

Figure 10 illustrates the pairwise posterior distributions for differences in edit distances across experimental conditions. For example, the difference in edit distances between the short and long static interfaces has a mode of 9.1, with a 94% highest density interval (HDI) of [6, 13]. This indicates that participants in the long text interface move approximately 9.1 steps more than those in the short text interface, with a high degree of confidence. The effect size is large (mode = 5.1, 94% HDI = [3.3, 7.1]), suggesting a statistically significant difference, which is expected due to the greater number of options in the long text interface.

Similarly, participants using the two-phase interface make approximately 8.9 fewer steps per option (mode = 8.9, 94% HDI = [6.4, 12]) than those in the long text interface, with a large effect size (mode = 5.7, 94% HDI = [4.2, 7.9]). The increase in edit distances between the short and long two-phase interfaces is substantially smaller (mode = 1.7, 94% HDI = [-0.01, 3.1]) compared to their static counterparts. Comparing the short text and short two-phase interfaces shows limited difference (mode = 1.3, 94% HDI = [-0.78, 3.8]), though the posterior distribution favors fewer steps for the two-phase interface (89.3% probability). The model suggests that the two-phase interface reduces edit distance per option, particularly for the long QS.

Edit distance per action: Building on the statistical disparities observed in the previous analysis and the unique patterns exhibited by long text interface participants, we present analyses focusing on edit distance per action and cumulative edit distance throughout the survey between the long text and long two-phase interfaces. Edit distance per action measures how far participants move during each adjustment while completing the survey. Figure 11 illustrates how, at each step, the number of participants moving a given distance (represented by the size of the dots) varies across

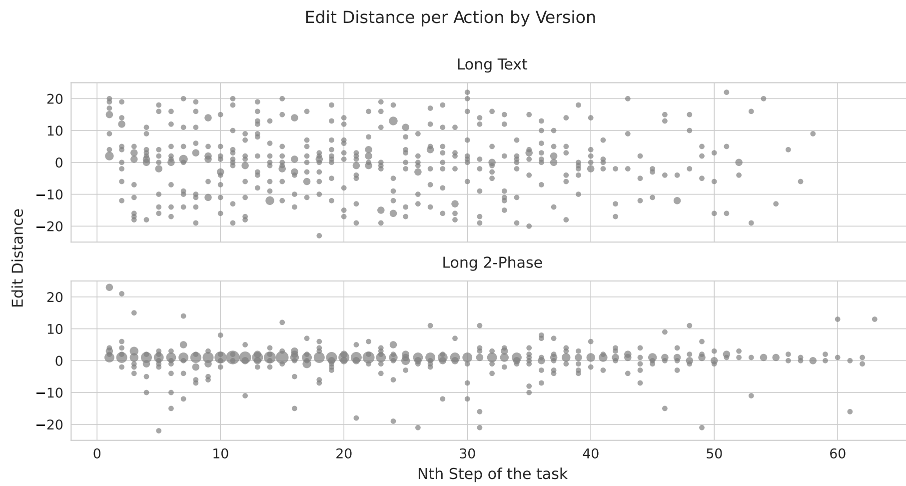


Fig. 11. Edit Distance Per Action: This plot shows the frequency of specific edit distances at each step across the text interface and two-phase interface. **Main takeaway:** Participants in the long two-phase interface tend to make adjustments closer to their previous actions, resulting in visually less variance in edit distances throughout the entire survey.

experimental conditions. Visually, participants move less on average per option within the two-phase interface, with lower variance at smaller scales. This indicates that participants are making local edits, meaning their adjustments tend to occur near their previous edits in terms of edit distance. This also highlights that the organization phase effectively adjusts option positions for easier access, despite participants still having the freedom to move across the interface as all options are presented to them.

In contrast to earlier analyses, we use a hierarchical Bayesian model (detailed in Appendix J.2) to jointly estimate the mean and variance of edit distances across experimental conditions. The model assumes that edit distances are continuous and follow a Normal likelihood. This approach accounts for both central tendencies and variability, using separate predictors for the mean and variance. The model includes hierarchical effects for survey length, interface type, interactions between length and interface, and user-level random effects. Non-centered parametrization is used for survey length and interface type to improve convergence, while interaction effects are modeled with an LKJ prior to capture the correlations between factors. User-level random effects reflect individual differences in behavior, incorporating variability into the model.

Figure 12 illustrates the posterior variance distributions, confirming our hypothesis. Participants in the long text interface exhibit greater variance in movement, frequently navigating across the interface, compared to those in the long two-phase interface. This is evidenced by a variance difference mode of 76 (95% HDI = [59, 99]) and a large effect size (mode = 7.1, 95% HDI = [5.5, 9.2]).

Cumulative edit distance for a participant: Figure 13 illustrates how the two-phase interface reduces per-action distance, accumulating over time. Some long text participants traverse double the amount of distance to complete the task compared to the long two-phase participants. We model this growth rate using a hierarchical Bayesian regression model (Detailed in Appendix J.3), with cumulative distance as the predictive variable. The experimental variables include interface type as a categorical variable, individual users modeled with random effects, and steps taken as a

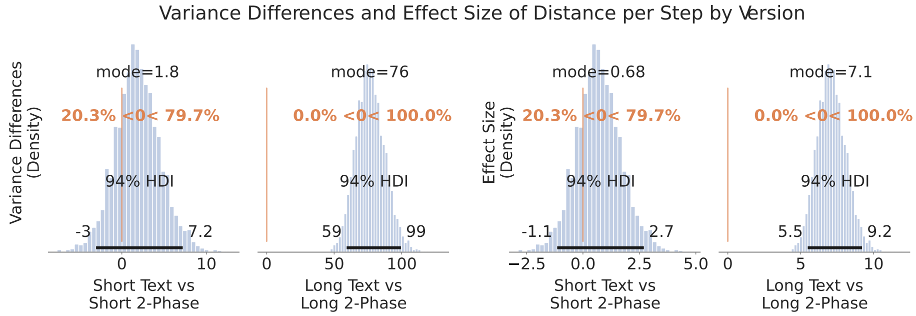


Fig. 12. The figure shows the contrast distributions of the mean edit distance per step between the two-phase interface and text interface for different survey lengths. The left two subplots represent absolute differences, while the right two depict effect sizes. **Main takeaway:** is that participants in the long text condition exhibited greater variance in edit distance per step compared to those in the long two-phase interface. Similarly, the short text condition showed higher differences, although these were not statistically significant in Bayesian terms.

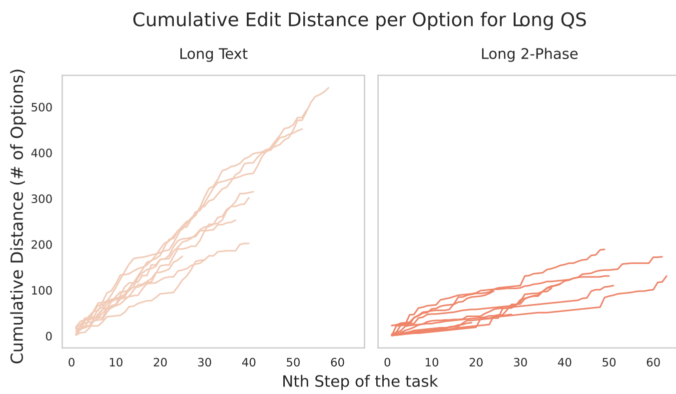


Fig. 13. This plot shows how the cumulative edit distances gained over the course of the survey between long text and long two-phase groups. **Main takeaway:** Participants in the long two-phase interface tend to make smaller, more incremental adjustments, resulting in a visually flatter slope compared to the text interface.

continuous variable. A truncated normal likelihood constrains cumulative distances to positive values and varies these distances across steps for each participant while masking incomplete data.

Figure 14 shows that the slope for the long text interface is approximately 4.7, meaning each step by the text interface would add 4.7 edit distance (94% HDI = [4.2, 5.4]), compared to the long two-phase interface, which shows a statistically significant difference with a mode of 1.4 (94% HDI = [1.3, 1.7]). These results explain that the variance in edit distance per action and the increase in per option edit distance are consistent across participants between the two groups, showing that the organization phase allows participants to focus on adjusting options within proximity without having to navigate the interface to locate and make adjustments during the voting phase.

Evidence from qualitative analysis: Recall the differences in sources of cognitive load between the two experimental conditions: while two-phase interface participants make localized adjustments with nearby options, they experience cognitive demand from preference construction due to broader considerations that involve more options and

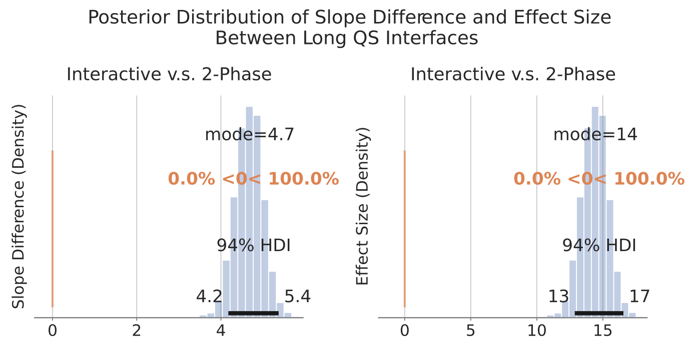


Fig. 14. The figure shows the contrast distributions of slope differences in cumulative edit distance between the two-phase interface and text interface for long QSs. The left subplots show absolute differences, while the right depict effect sizes. **Main takeaway:** Participants in the long text interface exhibited a steeper slope, indicating a faster increase in cumulative edit distance compared to the long two-phase interface.

higher-order values. Similarly, the qualitative results highlight that long text interface participants construct narrower preferences, yet their edit distance indicates broader movements across options.

Fewer participants (60%, N=6) in the long two-phase interface report precise resource allocation as source of demand compared to 90% in the long text interface (N=9). We interpret this as former participants construct preliminary preferences during the organization phase, easing them to focus on deciding their votes as they focus more on deliberate preference building rather than mere completion. Conveniently position options with another option of similar preferences further reduced need to looking for an option and traverse the interface, allowing participants to remaining engaged in vote adjustments.

7 Clickstream data: Interface participants' time spent

In addition to distance, participants in the short survey took an average of 2.7 minutes (short-text: $\mu = 2.3$, $\sigma = 1.27$; short two-phase: $\mu = 3$, $\sigma = 1.02$), while those in the long survey took 9.7 minutes (long-text: $\mu = 7.5$, $\sigma = 3.45$; long two-phase: $\mu = 11.95$, $\sigma = 2.73$). For a fairer comparison of interaction patterns, we analysis total **time-spend-per-option** using the QS system logs in this section. For participants in the two-phase interface conditions, this includes both organization and voting times for that option. The results are visualized in Figure 15.

Overall, participants spend slightly more time per option in the two-phase interface than in the text interface. To quantify these observations, we model the time data as predictive variables of separate Gamma distributions to characterize the continuous response times observed under distinct experimental conditions defined by survey length and interface type (Detailed in Appendix I). Each of the four resulting subsets of data is modeled independently, with separate Gamma-distributed parameters governing the shape and rate of each group's time distributions.

We calculated the posterior differences between the two-phase and text interfaces for all pairwise comparisons of the four groups. The results in Figure 16 indicate that participants using the two-phase interface consistently spend more time per option than those using the text interface, regardless of survey length. For both the short and long QSs, participants most likely spend 6.1 seconds (94% HDI = [1.0, 11.0]) and 6.7 seconds (94% HDI = [3.7, 9.4]) more per option, respectively, with medium effect sizes of $d = 0.49$ (94% HDI = [0.077, 0.89]) and $d = 0.41$ (94% HDI = [0.24, 0.59]). In

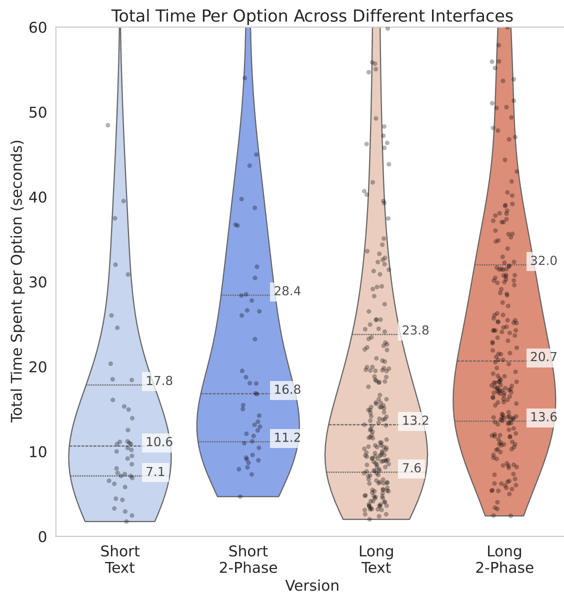


Fig. 15. Total Time per Option. Each dot in the plot represents the total time it took for a participant to complete an option. The shape of the plot presents how the dots distributed within that group. The wider it is, the more dots there are. The three horizontal lines indicate the 25th, 50th, and 75th percentile annotated with value. The two-phase interface skewed slightly higher than the text interface **Main takeaway:** Two-phase interface participants spend longer time per option compared to its counterparts.

both cases, the intervals lie outside the ROPE of 0 ± 1, indicating statistical significance. These findings suggest that the two-phase interface encourages longer deliberation, particularly for longer lists of options.

Some literature points to increased time can lead to cognitive fatigue [85, 86], which can impair decision-making. Other decision science literature suggests that longer decision times can indicate deeper cognitive processing [87, 47]. Our qualitative analysis points to the latter.

Descriptively, participants in the long two-phase condition remained actively engaged during the voting phase, editing their votes an average of 39.3 times per participant (σ =39.3, range=$19-63$) compared to 39.1 times (σ =13.29, range=$15-58$) in the long text condition. This suggests that the two-phase interface does not reduce engagement despite the additional organization step.

Quantitatively, other than the difference in operational thinking and strategic consideration discussed in Section 5.3, we find that 37.5% of participants (N=15) who attribute time to *Decision Making* as a source of temporal demand frame such demand differently. We label a participant as *affirmative* if they describe the pressure to make decisions as a source of temporal demand. For example, `S022` *So it didn't take too much time, but obviously there were a lot of things to consider, so there was some temporal demand.* is an affirmative statement. Conversely, we label a participant as *negative* if they express concern about the time and effort they have already invested. For example, `S024` *maybe I should just hurry up and make a decision.* is a negative statement.

50% of participants (N=5) in the long two-phase group describe the pressure to make decisions affirmatively and none negatively. This suggests that their pressure stems from having too many remaining decisions to make, rather than from the time already invested. This is reflected in their higher average time spent per option and overall time

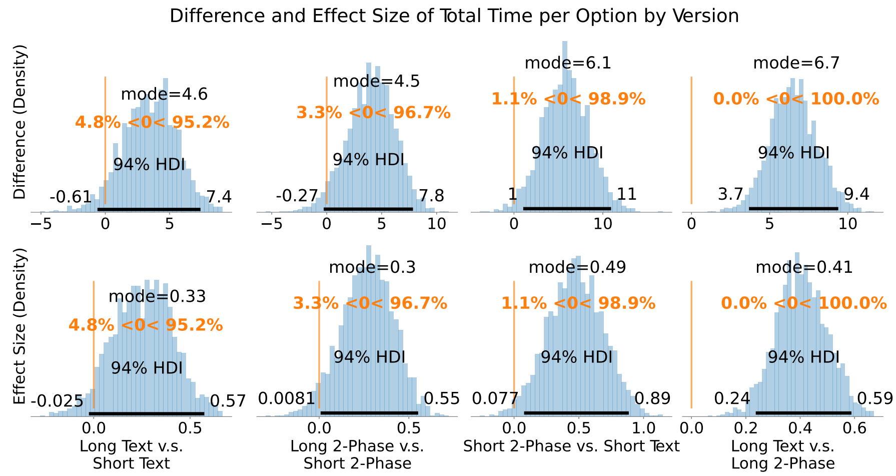


Fig. 16. The figure shows the contrast distributions of the mean time to complete per option between pairwise experimental conditions, with the first row representing absolute differences and the second row depicting effect sizes. **The main takeaway:** is that participants in the long two-phase condition spent more time per option compared to those in the long text and short two-phase conditions. Additionally, short two-phase participants took longer per option than short text participants.

spent (μ = 716.86 seconds, σ = 164.04 seconds) completing the QS compared to the long text group (μ = 449.64 seconds, σ = 206.97 seconds). We interpret results that participants are thoughtfully engaged in constructing their preferences and choose to invest additional time, rather than being driven by decision-related pressures or experiencing urgency.

Conversely, in the short text group, 50% of participants (N=5) express concern about the time and effort they have already invested (S024 *maybe I should just hurry up and make a decision.*) and none frame it affirmatively. Descriptively, participants in the short text group spend comparatively less time than those in the long QS (short text: μ = 139.83 seconds, σ = 76.43 seconds; short two-phase: μ = 178.78 seconds, σ = 61.07 seconds). This suggests that participants in the short text group expect themselves to complete the task sooner than they actually do.

Surprisingly, participants in the long text interface exhibit lower temporal demand compared to both the short text and long two-phase interfaces (Figure 17). Bayesian analysis (Appendix H.1.8) supports this finding, with posterior probabilities of 86.1% and 86.7%, respectively. This result is notable considering participants spent more time per option compared to those in the short text interface and traversed the longest distance among all three groups (Section 6). In addition, only 30% of participants (N=3) mention the time spent making a decision as a source of temporal demand. One possible explanation is that some participants are satisficing, as we pointed out in Section 5.4.

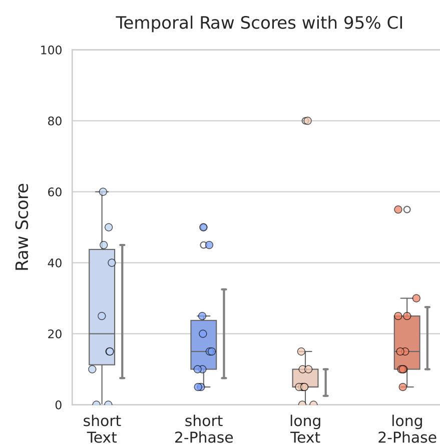


Fig. 17. Temporal Demand Raw Score: Each dot represents a participant's subscale response. **Main takeaway:** Long text interface participants seem to express less temporal demand compared to the other experiment conditions.

In summary, we interpret the result that participants in the two-phase interface spend more time per option as a sign of deeper cognitive processing. This is further supported by examining participants' nuanced voting behaviors under budget constraint conditions for the long QS, which we omit for brevity. Notably, two-phase interface participants make more small vote adjustments (i.e., adding or removing at most 2 votes on an option) when they have fewer remaining credits, further supporting our claim that they experience deeper engagement with preference construction, which we elaborate on further in Appendix G.

8 Discussion and Future Work

In this section, we interpret our findings on cognitive load and respondent behavior in a QS. We highlight the rationale and elements behind the two-phase interface for preference construction and its potential to mitigate satisficing. We also offer usage and design recommendations for practitioners and outline future directions for improving QS interfaces.

8.1 Two-phase interface: a worthwhile trade-off

Survey designers seek thoughtful responses from participants. This means the interface should balance survey usability, respondent satisfaction, and the effort individuals invest in their responses. Our results indicate that the two-phase interface encouraged deeper participant engagement with the options and reduced satisficing behaviors, despite its increased time per option and higher cognitive load for the long QS.

8.1.1 Analysis through the lens of cognitive load theory. Cognitive load theory [56], when applied to QSs, identifies three components of cognitive load: intrinsic load (the cognitive demand required to understand questions and response options), germane load (associated with deeper processing and preference evaluation), and extraneous load (stemming from navigating and operating the survey interface).

Participants were randomly assigned to experimental conditions, with survey lengths containing options randomly drawn from a common pool to control intrinsic load within the same group.

When a QS is short, participants can engage with all options simultaneously. Participants using the two-phase interface traded a slightly longer survey response time for a potential reduction in cognitive load and edit distance. We interpret this as participants freeing up cognitive demand from extraneous load for germane load, prompting them to better construct and express their preferences.

When a QS is long, participants face more options, resulting in a higher intrinsic load at the start of the survey. We believe the two-phase interface traded longer survey response time and a potential increase in cognitive load for deeper engagement with the survey. This heightened cognitive load likely stemmed from making comparisons in both the organization and voting phases. Quantitatively, participants spent more time per option, suggesting deeper engagement while exerting limited extraneous load, as evidenced by shorter traversals during voting. Qualitatively, participants reported experiencing demand primarily from strategic considerations (germane load) rather than operational actions (extraneous load), which were more common among text interface participants.

While some might argue that the additional organizing phase offers participants more opportunities to familiarize themselves with the options compared to text interface participants, the greater overall edit distance and high variance in edit distance per option suggest that text interface participants traversed the list frequently. This finding is further supported by qualitative data, where 70% of long-text participants (N=7) reported scanning the list while voting. This behavior suggests that while long-text participants had opportunities to familiarize themselves with the options, the explicit organization phase encouraged deeper reflection on their preferences.

The effect of the two-phase interface shows nuanced differences influencing cognitive load outcomes; however, both analyses suggest that the two-phase interface *shifted* participants' cognitive focus when completing QS.

8.1.2 Potential in limiting Satisficing. Qualitative findings (Section 5.4) on potential satisficing behavior highlight the importance of careful consideration when deploying a long QS. However, the two-phase interface appeared to limit satisficing behaviors, as evidenced by fewer observations compared to the long text interface for the long QS and none for the short QS. We believe the potential reasons lie in the design of the two-phase interface, which scaffolds the preference construction process.

The deliberate one-option-at-a-time presentation during the voting task in the two-phase interface reduced reliance on defaults and encouraged deeper reflection using cognitive strategies such as *problem decomposition* [88] and *dimension reduction*, both of which are known to reduce cognitive overload.

When asked about their experience with the interface, four participants highlighted how the organization phase supported their preference construction.`S013` illustrated how the one-option-at-a-time approach reduced the dimensions of decision-making:

> *[...] it (organization phase) gives you time to just focus on that single thing and rank it based on how you feel at that moment.*
> `S013 (S2P)`

This focused mode enabled deeper reflection. When considering relative preferences among QS options,`S013` described how it structurally decomposed the problem:

> *[...] to have a preliminary categorization of all the topics [...] (allowed me) to think about and process [...] digest all the information prior to actually allocating the budget [...]* `S009 (L2P)`

This quote highlighted how participants' deliberation occurred during the organization phase, enabling them to focus on constructing preferences without worrying about budget management—both of which are cited sources of cognitive load. Although direct measurement of satisficing behavior reduction is challenging, qualitative data and participant feedback suggest that the two-phase interface has the potential to limit such behaviors. Based on this evidence, we recommend that long QSs be implemented with a two-phase interface and sufficient time for participants to complete the process. We suggest future research investigate the mental processes underlying satisficing behaviors in long QSs.

In summary, we argue that the trade-off of a longer completion time and potentially higher cognitive load in the two-phase interface is justified. Drawing on cognitive load theory, we propose that the interface fosters deeper engagement with the options. Additionally, our qualitative findings and participant feedback suggest that the interface may reduce satisficing, aligning with decision-makers' goals of obtaining thoughtful and deliberate responses from participants.

8.2 Preference Construction guided by Organize, Then Vote

Completing a QS involves a series of in-situ difficult decision tasks Lichtenstein and Slovic [7]. As one participant reflected when completing the survey with options they had never considered before:

> *Oh, there are other aspects that I never care about. [...] Why (should) I spend money on that?*
> `S037 (L2P)`

When processing these unfamiliar options, we believe the two-phase interface supported participants' preference construction process.

First, 40% of long-text participants (N=3) found it challenging to facilitate differentiation without organization tools that would allow grouping or drag-and-drop, as `S025` said:

> *I would like to be able to like, click and drag the categories themselves so I could maybe reorder them to like my priorities. [. . .] make myself categories and subcategories out of this list . . . If I could organize it.* S025 (LT)

In contrast, 60% (N=6) of long two-phase participants appreciated the upfront introduction of all options, which enabled them to organize and use drag-and-drop features to facilitate QS completion. Not only did participants use drag-and-drop options post-voting to reflect and ensure correct vote allocation, but drag-and-drop also enabled participants, like S039, to make fine-grained comparisons between options:

> *I think the system was actually really helpful because I could just drag them. [. . .] I can really compare them, I can drag this one up here, and then compare it to the top one [. . .]* S039 (S2P)

This supports our intention of applying Svenson [54]'s differentiation and consolidation theory, where participants attempt to identify differences and eliminate less favorable options. The organization phase and the drag-and-drop supported some degree of differentiation process.

> *[. . .] the hardest part deciding in which category of place (prefernce bin) each issue is.* S021 (L2P)

This quote by S021 best represents the potential of the organization phase in separating part of the difficult decisions one needs to make when differentiating their preferences during preference construction. With the selected options, the shorter edit distance of long two-phase interface participants suggested that they were consolidating their identified preferences through votes.

8.3 What We Learned: Quadratic Survey Usage and Design Recommendations

This study represents a crucial step toward developing better interfaces to support individuals responding to QSs, by providing a deeper understanding of how survey respondents interact with QSs and the sources of cognitive load. In this subsection, we outline usage and design recommendations applicable to all applications of the quadratic mechanism.

8.3.1 QS: Prioritizing Fewer Options or High-Stakes Evaluations. We recommend deploying a QS with smaller sets of options or for critical evaluations, such as eliciting stakeholders' preferences before making investment decisions in hospital infrastructure. Our findings indicate that cognitive challenges and time requirements increase significantly as the number of options grows. For a long QS, while the two-phase interface helps mitigate some challenges, it does not eliminate them entirely, making adequate deliberation time essential. If a two-phase interface is unavailable, survey designers should present options in advance to allow participants to familiarize themselves and reflect before completing the QS.

8.3.2 Facilitate Quadratic Mechanism Applications through Categorization, Not Ranking. In a QS, the final ranking of preferences is typically a byproduct of vote allocation rather than a deliberate ranking effort. Participants did not explicitly rank options; instead, their preferences emerged dynamically through the voting process. To better support this preference construction, future quadratic mechanism interface designs should focus on helping participants categorize options effectively rather than ranking them directly. Facilitating differentiation among options is more critical than enabling precise manipulation for fine-tuning. We believe this approach extends beyond QSs to other ranking-based survey tools, such as ranked-choice voting and constant-sum surveys. Further research should examine how implementing such functionality influences survey respondents' mental models.

8.4 Future work: Opportunities for Better Budget Management

Budget management emerged as one of the most prominent issues in our study, which the two-phase interface did not address. 35% of participants ($N = 14$) emphasized the ability of current quadratic mechanism applications to perform automated calculations, but noted that this is not sufficient. We identified three key challenges for future work:

First, participants struggled to decide on an initial vote allocation. Some distributed credits equally across options, while others used 1, 2, or 3 votes as starting points. A few anchored their decisions to the tutorial's example of four upvotes. This suggests a need to better understand whether individuals have absolute value preferences among options. Second, 12.5% of participants ($N = 5$) expressed confusion about the relationship between budget, votes, and outcomes, despite understanding their definitions. They struggled to make trade-offs between votes and budget, leading to frustration and hampered decision-making. Third, determining the absolute amount of credits in a QS is highly demanding. Designing interfaces and interactions to address the cold start challenge and help participants decide on the absolute vote value, while also considering ways to limit direct influences, remains an open question.

We believe that, with well-designed interface backed by real-time computing and a better understanding of how individuals calculate trade-offs, we can provide innovative solutions to help participants more easily express their preferences using QSs.

9 Limitations

Evaluating the QS interface is challenging because of its novelty. During the study, we identified several limitations that warrant further research.

Individual differences in cognitive capacity. Variations in individual cognitive capacity influenced participants' performance and cognitive scores. For example, participants with greater experience in decision-making may be better able to manage multiple options. A within-subject study could clarify shifts in cognitive load, but deconstructing established preferences and altering options introduces additional complexity. Therefore, we opted for this in-depth, between-subject study, although the small sample size may introduce noise, potentially distorting the measurement of cognitive load. Future research should aim to quantify the impact of different QS interfaces on cognitive load at a larger scale. Furthermore, participants completed this study in a controlled laboratory environment, with options displayed on a large screen. Future work should also investigate how individuals respond to QSs on smaller devices and in less controlled environments.

Limited experience with QSs. Participants lacked prior experience with the QS interface. After completing a tutorial and quiz, participants proceeded to perform tasks using the QS interface. While participants understood the mechanics of QSs, their familiarity with the interface likely influenced their strategies and cognitive load. As quadratic mechanisms become more prevalent, future research could compare the performance of novices and experts.

Limitations of Time and Distance as Proxies for Decision-Making Effort. While time and distance are common metrics for quantifying the effort involved in decision-making, they do not capture without noise. Participants may considered multiple options simultaneously. We acknowledge that these metrics are approximate indicators of decision-making effort. Despite these limitations, this approach provides valuable insights into decision-making within our experimental constraints.

10 Conclusion

This study introduces and evaluates a two-phase "Organize-then-Vote" interface to help QS respondents construct their preferences. We examined how the interface affected cognitive load and response behaviors across societal issues of varying lengths through in-lab study, NASA-TLX and interviews. The interface's organization and voting phases, designed to reduce cognitive overload by structuring the decision-making process, allowed respondents to differentiate between options before voting. Results revealed that the two-phase design reduced participant's edit distance between vote adjustments throughout the survey despite spending more time per option. Qualitative insights highlighted two-phase interface encouraged more iterative and reflective preference construction and it's potential at reducing satisficing behaviors even though it did not clearly reduce overall cognitive load for the longer QS. Nonetheless, this design shift promoted deeper engagement and strategic thinking compared to the text-based interface, by distributing cognitive effort more effectively. By integrating the organization and drag-and-drop functions, the interface facilitated both preference differentiation and consolidation, making it easier for respondents to refine their decisions. This two-phase interface design supports the development of future software tools that facilitate preference construction and promote the broader adoption of QSs. Future research should explore how to better support individuals in deciding the allocation of budget and design interfaces for smaller devices.

References

[1] Martin Pielot and Mario Callegaro. 2024. Did You Misclick? Reversing 5-Point Satisfaction Scales Causes Unintended Responses. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, (May 2024), 1–7. doi: 10.1145/3613904.3642397.

[2] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, (May 2019), 1–12. doi: 10.1145/3290605.3300316.

[3] Muhsin Ugur, Dvijesh Shastri, Panagiotis Tsiamyrtzis, Malcolm Dcosta, Allison Kalpakci, Carla Sharp, and Ioannis Pavlidis. 2015. Evaluating smartphone-based user interface designs for a 2d psychological questionnaire. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 275–282.

[4] Ti-Chung Cheng, Tiffany Li, Yi-Hung Chou, Karrie Karahalios, and Hari Sundaram. 2021. "I can show what I really like.": Eliciting Preferences via Quadratic Voting. *Proceedings of the ACM on Human-Computer Interaction*, 5, (Apr. 2021), 1–43. doi: 10.1145/3449281.

[5] Theodore Groves and John Ledyard. 1977. Optimal Allocation of Public Goods: A Solution to the "Free Rider" Problem. *Econometrica*, 45, 4, 783–809. JSTOR: 1912672. doi: 10.2307/1912672.

[6] 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.

[7] Sarah Lichtenstein and Paul Slovic, eds. 2006. *The Construction of Preference*. (1. publ ed.). Cambridge University Press, Cambridge.

[8] Adam Rogers. 2019. Colorado Tried a New Way to Vote: Make People Pay—Quadratically | WIRED. *Wired*, (Apr. 2019). Retrieved June 22, 2024 from.

[9] Internet Team. [n. d.] Taiwan Digital Minister highlights country's use of technology to bolster democracy in FT interview. https://www.roc-taiwan.org/uk_en/post/6295.html. (). Retrieved June 13, 2024 from.

[10] Herbert A. Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69, 1, 99–118. JSTOR: 1884852. doi: 10.2307/1884852.

[11] John W. Payne, James R. Bettman, and Eric J. Johnson. 1988. Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 3, (July 1988), 534–552. doi: 10.1037/0278-7393.14.3.534.

[12] Amos Tversky and Daniel Kahneman. [n. d.] Judgments of and by Representativeness.

[13] Erik J Engstrom and Jason M Roberts. 2020. *The Politics of Ballot Design: How States Shape American Democracy*. Cambridge University Press.

[14] Bert Weijters, Elke Cabooter, and Niels Schillewaert. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27, 3, (Sept. 2010), 236–247. doi: 10.1016/j.ijresmar.2010.02.004.

[15] N. D. Kieruj and G. Moors. 2010. Variations in Response Style Behavior by Response Scale Format in Attitude Research. *International Journal of Public Opinion Research*, 22, 3, (Sept. 2010), 320–342. doi: 10.1093/ijpor/edq001.

[16] Vera Toepoel, Brenda Vermeeren, and Baran Metin. 2019. Smileys, Stars, Hearts, Buttons, Tiles or Grids: Influence of Response Format on Substantive Response, Questionnaire Experience and Response Time. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 142, 1, (Apr. 2019), 57–74. doi: 10.1177/0759106319834665.

[17] Habiba Farzand, David Al Baiaty Suarez, Thomas Goodge, Shaun Alexander Macdonald, Karola Marky, Mohamed Khamis, and Paul Cairns. 2024. Beyond Aesthetics: Evaluating Response Widgets for Reliability & Construct Validity of Scale Questionnaires. In *Extended Abstracts of the 2024*

CHI Conference on Human Factors in Computing Systems (CHI EA '24). Association for Computing Machinery, New York, NY, USA, (May 2024), 1–7. DOI: 10.1145/3613905.3650751.
[18] Christian Jilek Paula Gauselmann Yannick Runge and Tobias Tempel. 2023. A relief from mental overload in a digitalized world: How context-sensitive user interfaces can enhance cognitive performance. *International Journal of Human–Computer Interaction*, 39, 1, 140–150. eprint: https://doi.org/10.1080/10447318.2022.2041882. DOI: 10.1080/10447318.2022.2041882.
[19] 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.
[20] Michael Xieyang Liu, Aniket Kittur, and Brad A. Myers. 2021. To reuse or not to reuse? A framework and system for evaluating summarized knowledge. *Proc. ACM Hum.-Comput. Interact.*, 5, CSCW1, (Apr. 2021). DOI: 10.1145/3449240.
[21] Helena M Reis et al. 2012. Towards reducing cognitive load and enhancing usability through a reduced graphical user interface for a dynamic geometry system: An experimental study. In *2012 IEEE International Symposium on Multimedia*. IEEE, 445–450.
[22] Benjamin Lafreniere, Andrea Bunt, and Michael Terry. 2014. Task-centric interfaces for feature-rich software. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design* (OzCHI '14). Association for Computing Machinery, New York, NY, USA, 49–58. DOI: 10.1145/2686612.2686620.
[23] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proc. ACM Hum.-Comput. Interact.*, 5, CSCW1, (Apr. 2021). DOI: 10.1145/3449161.
[24] Emin İbili. 2019. Effect of augmented reality environments on cognitive load: pedagogical effect, instructional design, motivation and interaction interfaces. *International Journal of Progressive Education*, 15, 5, 42–57.
[25] Amy X. Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proc. ACM Hum.-Comput. Interact.*, 2, CSCW, (Nov. 2018). DOI: 10.1145/3274465.
[26] Steven P Lalley, E Glen Weyl, et al. 2016. Quadratic voting. *Available at SSRN*.
[27] Eric A Posner and E Glen Weyl. 2018. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press.
[28] Ryan Naylor et al. 2017. First year student conceptions of success: What really matters? *Student Success*, 8, 2, 9–19.
[29] Charlotte Cavaille and Daniel L Chen. [n. d.] Who Cares? Measuring Preference Intensity in a Polarized Environment.
[30] Vitalik Buterin, Zoë Hitzig, and E. Glen Weyl. 2019. A Flexible Design for Funding Public Goods. *Management Science*, 65, 11, (Nov. 2019), 5171–5187. DOI: 10.1287/mnsc.2019.3337.
[31] Luis Mota Freitas and Wilfredo L. Maldonado. 2024. Quadratic funding with incomplete information. *Social Choice and Welfare*, (Feb. 2024). DOI: 10.1007/s00355-024-01512-7.
[32] Tobin South, Leon Erichsen, Shrey Jain, Petar Maymounkov, Scott Moore, and E. Glen Weyl. 2024. Plural Management. SSRN Scholarly Paper. Rochester, NY, (Jan. 2024). DOI: 10.2139/ssrn.4688040.
[33] 2023. Gov4git: A Decentralized Platform for Community Governance. (Mar. 2023). Retrieved June 13, 2024 from.
[34] 2024. RadicalxChange/quadratic-voting. RadicalxChange. (May 2024). Retrieved June 17, 2024 from.
[35] [n. d.] Read the Whitepaper | Gitcoin. https://www.gitcoin.co/whitepaper/read. (). Retrieved June 17, 2024 from.
[36] [n. d.] About RxC. https://www.radicalxchange.org/wiki/about/. (). Retrieved June 17, 2024 from.
[37] yehjxraymond. 2024. Yehjxraymond/qv-app. (Mar. 2024). Retrieved June 17, 2024 from.
[38] Charlotte Cavaille, Daniel L Chen, and Karine Van der Straeten. 2024. Who cares? Measuring differences in preference intensity.
[39] Syavash Nobarany, Louise Oram, Vasanth Kumar Rajendran, Chi-Hsiang Chen, Joanna McGrenere, and Tamara Munzner. 2012. The design space of opinion measurement interfaces: exploring recall support for rating and ranking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2035–2044.
[40] Paul Van Schaik and Jonathan Ling. 2007. Design parameters of rating scales for web sites. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 14, 1, 4–es.
[41] Jing Wei, Weiwei Jiang, Chaofan Wang, Difeng Yu, Jorge Goncalves, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding how to administer voice surveys through smart speakers. *Proc. ACM Hum.-Comput. Interact.*, 6, CSCW2, (Nov. 2022). DOI: 10.1145/3555606.
[42] Aman Khullar et al. 2021. Costs and benefits of conducting voice-based surveys versus keypress-based surveys on interactive voice response systems. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies* (Compass '21). Association for Computing Machinery, New York, NY, USA, 288–298. DOI: 10.1145/3460112.3471963.
[43] Martin Feick, Niko Kleer, Anthony Tang, and Antonio Krüger. 2020. The virtual reality questionnaire toolkit. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 68–69.
[44] Graham Cooper. 1998. Research into cognitive load theory and instructional design at UNSW. (1998).
[45] Stoo Sepp, Steven J. Howard, Sharon Tindall-Ford, Shirley Agostinho, and Fred Paas. 2019. Cognitive Load Theory and Human Movement: Towards an Integrated Model of Working Memory. *Educational Psychology Review*, 31, 2, (June 2019), 293–317. DOI: 10.1007/s10648-019-09461-9.
[46] Antonio Drommi, Gregory W Ulferts, and Dan Shoemaker. 2001. Interface design: A focus on cognitive science. In *The Proceedings of ISECON 2001*. Vol. 18.
[47] Kahneman Daniel. 2017. *Thinking, Fast and Slow*.
[48] 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.

[49] 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.

[50] N. T. Feather. 1973. The measurement of values: Effects of different assessment procedures. *Australian Journal of Psychology*, 25, 3, (Dec. 1973), 221–231. DOI: 10.1080/00049537308255849.

[51] Peter Coy. 2019. A New Way of Voting That Makes Zealotry Expensive - Bloomberg. *Bloomberg*, (May 2019). Retrieved Dec. 16, 2023 from.

[52] 2022. Quadratic Voting Frontend. Public Digital Innovation Space. (Jan. 2022). Retrieved Dec. 16, 2023 from.

[53] Henry Montgomery. 1983. Decision Rules and the Search for a Dominance Structure: Towards a Process Model of Decision Making. In *Advances in Psychology*. Vol. 14. Elsevier, 343–369. DOI: 10.1016/S0166-4115(08)62243-8.

[54] Ola Svenson. 1992. Differentiation and consolidation theory of human decision making: A frame of reference for the study of pre- and post-decision processes. *Acta Psychologica*, 80, 1-3, (Aug. 1992), 143–168. DOI: 10.1016/0001-6918(92)90044-E.

[55] Fritz Strack and Leonard L. Martin. 1987. Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. In *Social Information Processing and Survey Methodology*. Recent Research in Psychology. Hans-J. Hippler, Norbert Schwarz, and Seymour Sudman, editors. Springer, New York, NY, 123–148. DOI: 10.1007/978-1-4612-4798-2_7.

[56] John Sweller. 2011. Cognitive Load Theory. In *Psychology of Learning and Motivation*. Vol. 55. Elsevier, 37–76. DOI: 10.1016/B978-0-12-387691-1.00002-8.

[57] Robert Münscher, Max Vetter, and Thomas Scheuerle. 2016. A Review and Taxonomy of Choice Architecture Techniques. *Journal of Behavioral Decision Making*, 29, 5, 511–524. DOI: 10.1002/bdm.1897.

[58] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness. Nudge: Improving Decisions about Health, Wealth, and Happiness.* Yale University Press, New Haven, CT, US, x, 293.

[59] A Norman Donald. 2013. *The Design of Everyday Things*. MIT Press.

[60] Christopher D Wickens and Anthony D Andre. 1990. Proximity compatibility and information display: Effects of color, space, and objectness on information integration. *Human factors*, 32, 1, 61–77.

[61] Jon A Krosnick, Charles M Judd, and Bernd Wittenbrink. 2018. The measurement of attitudes. In *The Handbook of Attitudes*. Routledge, 45–105.

[62] Jerry P Timbrook. 2013. *A Comparison of a Traditional Ranking Format to a Drag-and-Drop Format with Stacking*. PhD thesis. University of Dayton.

[63] Duncan Rintoul. [n. d.] Visual and animated response formats in web surveys: Do they produce better data, or is it all just fun and games?, 126.

[64] Susan C. Herring and Ashley R. Dainas. 2020. Gender and Age Influences on Interpretation of Emoji Functions. *ACM Transactions on Social Computing*, 3, 2, (June 2020), 1–26. DOI: 10.1145/3375629.

[65] [n. d.] Center for Civic Design. https://civicdesign.org/. (). Retrieved June 17, 2024 from.

[66] Robert Ferber. 1952. Order Bias in a Mail Survey. *Journal of Marketing*, 17, 2, 171–178. JSTOR: 1248043. DOI: 10.2307/1248043.

[67] M. P. Couper. 2001. Web survey design and administration. *Public Opinion Quarterly*, 65, 2, 230–253. DOI: 10.1086/322199.

[68] 2023. Charity Navigator. https://www.charitynavigator.org/index.cfm?bay=search.categories. (May 2023). Retrieved Dec. 16, 2023 from.

[69] William F. Moroney and Joyce A. Cameron. 2019. *Questionnaire Design: How to Ask the Right Questions of the Right People at the Right Time to Get the Information You Need*. Human Factors and Ergonomics Society, (Feb. 2019).

[70] Thomas L. Saaty. 1987. Principles of the Analytic Hierarchy Process. In *Expert Judgment and Expert Systems*. Jeryl L. Mumpower, Ortwin Renn, Lawrence D. Phillips, and V. R. R. Uppuluri, editors. Springer, Berlin, Heidelberg, 27–73. DOI: 10.1007/978-3-642-86679-1_3.

[71] George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 2, 81–97. DOI: 10.1037/h0043158.

[72] Thomas L Saaty and Mujgan S Ozdemir. 2003. Why the magic number seven plus or minus two. *Mathematical and computer modelling*, 38, 3-4, 233–244.

[73] Alexander Chernev, Ulf Böckenholt, and Joseph Goodman. 2015. Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, 25, 2, (Apr. 2015), 333–358. DOI: 10.1016/j.jcps.2014.08.002.

[74] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183.

[75] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50, 9, (Oct. 2006), 904–908. DOI: 10.1177/154193120605000909.

[76] Brad Cain. 2007. A review of the mental workload literature. *DTIC Document*.

[77] Qin Gao, Yang Wang, Fei Song, Zhizhong Li, and Xiaolu Dong. 2013. Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, 56, 7, (July 2013), 1070–1085. DOI: 10.1080/00140139.2013.790483.

[78] Susana Rubio, Eva Díaz, Jesús Martín, and José M. Puente. 2004. Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology*, 53, 1, 61–86. DOI: 10.1111/j.1464-0597.2004.00161.x.

[79] Oskar Palinko, Andrew L. Kun, Alexander Shyrokov, and Peter Heeman. 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10*. ACM Press, Austin, Texas, 141. DOI: 10.1145/1743666.1743701.

[80] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. ACM, Copenhagen Denmark, (Sept. 2010), 301–310. DOI: 10.1145/1864349.1864395.

[81] Judith S. Olson and Wendy A. Kellogg, eds. 2014. *Ways of Knowing in HCI*. Springer, New York, NY. DOI: 10.1007/978-1-4939-0378-8.

[82] 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.

[83] Richard McElreath. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.

[84] Gerd Gigerenzer and Daniel G. Goldstein. 1996. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 4, 650–669. DOI: 10.1037/0033-295X.103.4.650.

[85] Thomas Kundinger, Celena Mayr, and Andreas Riener. 2020. Towards a Reliable Ground Truth for Drowsiness: A Complexity Analysis on the Example of Driver Fatigue. *Proceedings of the ACM on Human-Computer Interaction*, 4, EICS, (June 2020), 1–18. DOI: 10.1145/3394980.

[86] Enamul Karim, Hamza Reza Pavel, Sama Nikanfar, Aref Hebri, Ayon Roy, Harish Ram Nambiappan, Ashish Jaiswal, Glenn R Wylie, and Fillia Makedon. 2024. Examining the landscape of cognitive fatigue detection: a comprehensive survey. *Technologies*, 12, 3, 38.

[87] John W. Payne, James R. Bettman, and Eric J. Johnson. 1993. *The Adaptive Decision Maker*. Cambridge University Press, Cambridge. DOI: 10.1017/CBO9781139173933.

[88] Herbert A. Simon. 1996. *The Sciences of the Artificial*. (3rd ed ed.). MIT Press, Cambridge, Mass.

[89] Bert Weijters, Kobe Millet, and Elke Cabooter. 2021. Extremity in horizontal and vertical Likert scale format responses. Some evidence on how visual distance between response categories influences extreme responding. *International Journal of Research in Marketing*, 38, 1, (Mar. 2021), 85–103. DOI: 10.1016/j.ijresmar.2020.04.002.

[90] Vera Toepoel and Frederik Funke. 2018. Sliders, visual analogue scales, or buttons: Influence of formats and scales in mobile and desktop surveys. *Mathematical Population Studies*, 25, 2, (Apr. 2018), 112–122. DOI: 10.1080/08898480.2018.1439245.

[91] Dana Chisnell. 2016. Democracy Is a Design Problem. 11, 4.

[92] 2015. Designing usable ballots Center for civic design. https://civicdesign.org/fieldguides/designing-usable-ballots/. (June 2015). Retrieved June 17, 2024 from.

[93] Jonathan N. Wand, Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane, Michael C. Herron, and Henry E. Brady. 2001. The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida. *The American Political Science Review*, 95, 4, 793–810. Retrieved Dec. 16, 2023 from JSTOR: 3117714.

[94] Whitney Quesenbery. 2020. Opinion | Good Design Is the Secret to Better Democracy. *The New York Times*, (Oct. 2020). Retrieved June 17, 2024 from.

[95] Sarah P. Everett, Kristen K. Greene, Michael D. Byrne, Dan S. Wallach, Kyle Derr, Daniel Sandler, and Ted Torous. 2008. Electronic voting machines versus traditional methods: improved preference, similar performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08). Association for Computing Machinery, New York, NY, USA, (Apr. 2008), 883–892. DOI: 10.1145/1357054.1357195.

[96] Seunghyun "Tina" Lee, Yilin Elaine Liu, Ljilja Ruzic, and Jon Sanford. 2016. Universal Design Ballot Interfaces on Voting Performance and Satisfaction of Voters with and without Vision Loss. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16). Association for Computing Machinery, New York, NY, USA, (May 2016), 4861–4871. DOI: 10.1145/2858036.2858567.

[97] Kathryn Summers, Dana Chisnell, Drew Davies, Noel Alton, and Megan McKeever. 2014. Making voting accessible: designing digital ballot marking for people with low literacy and mild cognitive disabilities. In *2014 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE 14)*.

[98] Shaneé Dawkins, Tony Sullivan, Greg Rogers, E. Vincent Cross, Lauren Hamilton, and Juan E. Gilbert. 2009. Prime III: an innovative electronic voting interface. In *Proceedings of the 14th International Conference on Intelligent User Interfaces* (IUI '09). Association for Computing Machinery, New York, NY, USA, (Feb. 2009), 485–486. DOI: 10.1145/1502650.1502727.

[99] Juan E. Gilbert, Jerone Dunbar, Alvitta Ottley, and John Mark Smotherman. 2013. Anomaly detection in electronic voting systems. *Information Design Journal (IDJ)*, 20, 3, (Sept. 2013), 194–206. DOI: 10.1075/idj.20.3.01gil.

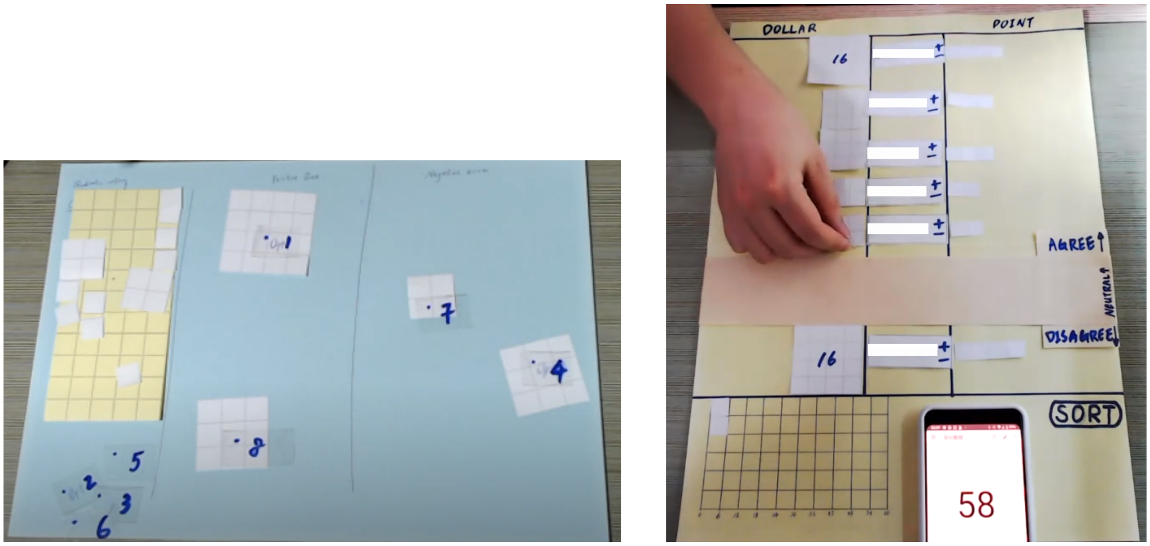
[100] Frederick G. Conrad, Benjamin B. Bederson, Brian Lewis, Emilia Peytcheva, Michael W. Traugott, Michael J. Hanmer, Paul S. Herrnson, and Richard G. Niemi. 2009. Electronic voting eliminates hanging chads but introduces new usability challenges. *International Journal of Human-Computer Studies*, 67, 1, (Jan. 2009), 111–124. DOI: 10.1016/j.ijhcs.2008.09.010.

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

[102] Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5–7, 2001 Proceedings 12*. Springer, 145–156.

A Interface design process

In this section, we outline the design process leading to our final interface. As mentioned in the paper, our design iteration began from existing QV applications in the wild.



(a) In this paper prototype, issues are denoted by different numbers that appear on mouseover. Pretest respondents can move options anywhere in the two sections of the interface, one denoting positive and one negative. The blocks represent the cost for each option, with no indication of the number of current votes. The credits are shown in the yellow box on the left.

(b) This paper prototype separates the positive and negative areas with a 'band' at the center. Undecided options are placed inside this band. The cost and the votes on both sides of the interface are denoted by small blocks. The budget is shown in the yellow box below the interface with a numerical counter.

Fig. 18. Initial paper prototypes designed for QS interface

A.1 Prototype 1: Ranking-Vote

Considering that relative preference is often through ranking items, we tested whether ranking options before voting would help establish an individual's relative preference in our prototype 1. This prototype allowed respondents to reposition options before voting. Pretests revealed that respondents rarely moved the options and questioned the necessity of full ranking, as it did not influence their QS submission. Additionally, many were unaware that options were draggable until shown. This insight indicates that full ranking is unnecessary for establishing relative preferences. Therefore, we decided to ask respondents to select a subset of options instead of requiring a full rank among all options.

A.2 Prototype 2: Select-then-Vote

Based on feedback from Prototype 1, instead of *allowing* individuals to rank options, Prototype 2 implemented a two-phase process that *intentionally* asks respondents to select options to express opinions before voting. As shown in Figure 20, survey respondents selected their preferred options (Figure 20a), and the interface positioned these options at the top of the list for voting (Figure 20b). We identified several issues during the prototype 2 pretest: many respondents marked most options as 'options they care about,' which undermined the design's purpose. Additionally, the lack of clear distinction between selected and unselected options confused respondents about the necessity of Step 1. Thus, we need a clearer distinction and connection between the two phases to effectively construct relative preferences.

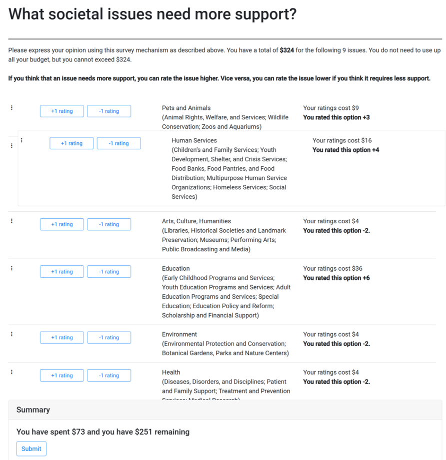
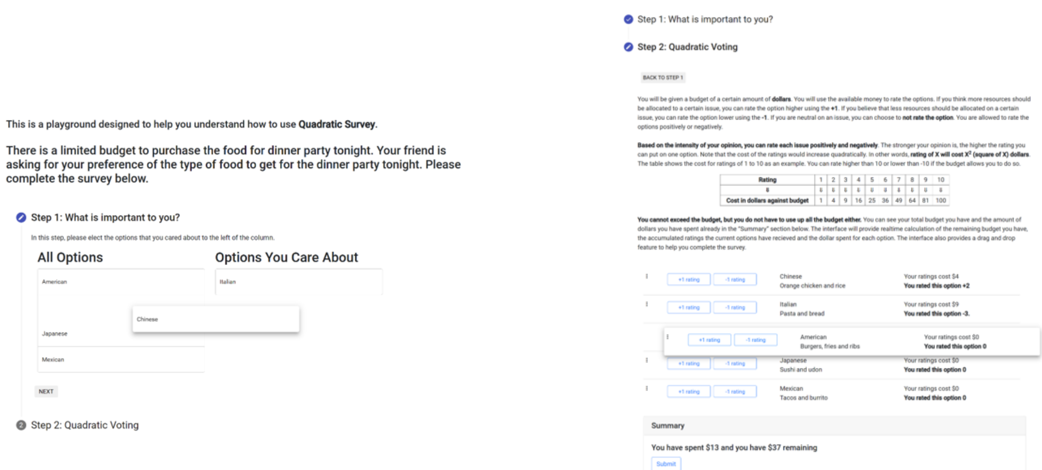


Fig. 19. A Ranking-Vote Prototype: The goal of this prototype is to test whether ranking options prior to voting help establish an individual's relative preferences and reduce effort when voting. Each option is draggable to position in a specific location amongst the full list of options. Votes can be updated using the buttons to the right of the interface with vote count and costs to the right of the interface. A summary box is placed sticky to the bottom of the screen.



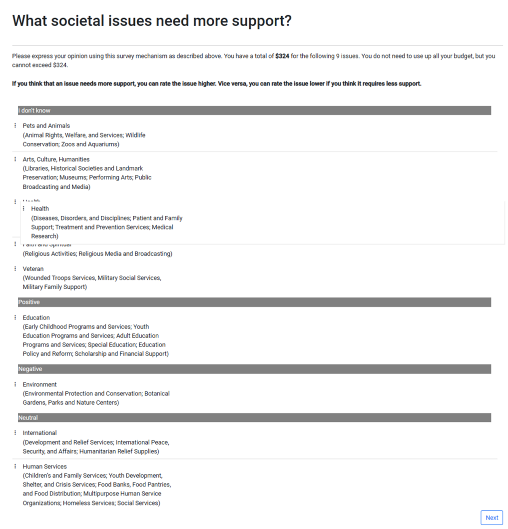
(a) Options are dragged and dropped to the 'Option You Care About' box.

(b) The previous step collapses showing all voting options.

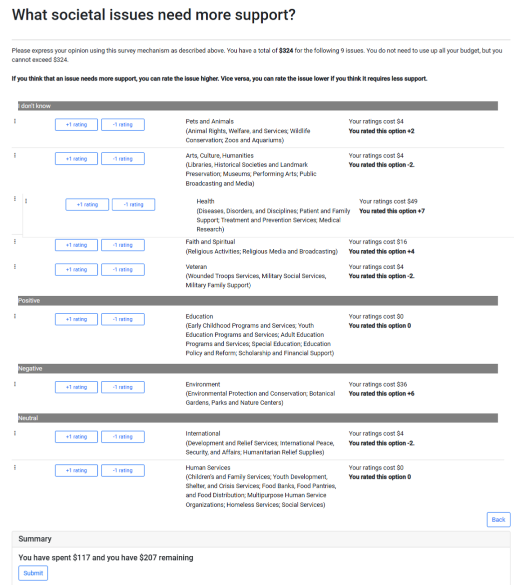
Fig. 20. A Select-then-Vote Prototype: The goal of this prototype is to nudge participants to focus on a subset of options to vote, rather than ranking all of them. This prototype introduces a two-step voting process. As shown in Fig. 20a, the first step involves selecting options for further consideration. Important options are placed at the top of the list for voting shown in Fig. 20b, but options can be placed anywhere on the list if desired. The rest of the controls remain the same as the previous prototype.

A.3 Prototype 3: Organize-then-Vote

Figure 21 shows the last prototype where we built on the previous takeaway by providing finer-grain groupings and creating a clear connection between option organization and voting position. Specifically, we provided three categories: Lean Positive, Lean Negative, and Lean Neutral. Initially, respondents see all options under the section labeled 'I don't



(a) The Organization Interface: Options are shown initially in the first bin labeled as 'I don't know.' Survey respondents can then drag and drop these options into the latter bins: Lean Positive, Lean Neutral, or Lean Negative. Only the details of each option are shown on this interface.



(b) The Voting Interface: Voting controls appear on the left side of each option, showing the current votes and associated costs on the right. A budget summary is stuck at the bottom of the screen.

Fig. 21. Organize-then-Vote Prototype: The goal of this prototype is to encourage participants to derive finer grain categories among options before voting. Survey respondents first organize their thoughts into categories and then vote on the options.

know,' which includes only the option descriptions. We ask respondents to move these options into the categories below. Voting controls and information appear on each option once respondents move to the subsequent page, forming a clear connection between option groups, positions, and voting controls.

Feedback indicated that survey respondents are comfortable with the two-phase organize-then-vote design, demonstrating it as a central strategy for our interface development. However, several areas for enhancement were identified: First, the dragging and dropping mechanism in the organization phase is cumbersome and may inadvertently suggest a ranking process, contrary to our intentions. Second, placing unorganized options at the top of the voting list is counterintuitive. Third, the voting controls are disconnected from the option summaries, dividing attention between the left and right sides of the screen. These insights guided refinements in the final two-phase interface, adhering to the two-phase organize-then-vote design framework.

B Voting Interface Breakdown

In this section, we outline additional literature that informed this study. There are two sets of literature that we surveyed: Survey response format and voting interfaces.

B.1 Survey response format

Research in the marketing and research communities focusing on survey and questionnaire design, usability, and interactions examines the influence of presentation styles and 'response format.' Weijters et al. [89] demonstrated that horizontal distances between options are more influential than vertical distances, with the latter recommended for reduced bias. Slider bars, which operate on a drag-and-drop principle, show lower mean scores and higher nonresponse rates compared to buttons, indicating they are more prone to bias and difficult to use. In contrast, visual analog scales that operate on a point-and-click principle perform better [90]. These studies show how even small design changes can have a large impact on usability, highlighting the importance of designing interfaces that prioritize human-centered interaction rather than focusing solely on functionality.

B.2 Voting Interfaces

Compared to digital survey interfaces, voting interfaces are a specialized type of survey interface can significantly influence democratic processes [13, 91, 92] and often have consequential impacts. Researchers believe that ill-designed voting interfaces We categorize these related works into three main categories detailed below:

Designs that shifted voter decisions: For example, states without straight-party ticket voting (where voters can select all candidates from one party through a single choice) exhibited higher rates of split-ticket voting [13]. Another example from the Australian ballot showing incumbency advantages is where candidates are listed by the office they are running for, with no party labels or boxes.

Designs that influenced errors: Butterfly ballots, an atypical design, may have influenced the outcome of the 2000 U.S. Presidential Election [93]. It increased voter errors because voters could not correctly identify the punch hole on the ballot. Splitting contestants across columns increases the chance for voters to overvote [94]. On the other hand, Everett et al. [95] showed the use of incorporating physical voting behaviors, like lever voting, into graphical user interfaces.

Designs that incorporated technologies: Other projects like the Caltech-MIT Voting Technology Project have sparked research to address accessibility challenges, resulting in innovations like EZ Ballot [96], Anywhere Ballot [97], and Prime III [98]. In addition, Gilbert et al. [99] investigated optimal touchpoints on voting interfaces, and Conrad et al. [100] examined zoomable voting interfaces.

Response format literature and voting interfaces informed how interfaces significantly influence respondent behavior, decision accuracy, and cognitive load. These burdens are especially problematic for complex systems like QS, where high cognitive demands may deter researchers and users alike. Developing effective, human-centered interfaces for QS could enhance usability, reduce cognitive overload, and increase adoption in both research and practical applications.

C Demographic Breakdown

We provide the table for a more detail demographic breakdown per group.

Table 1. Participant Age and Gender Distribution by Experimental Condition

Condition	Mean Age	SD	Range	25th	Median	75th	Male	Female	Non-binary
Short Text	31.6	13.7	18–67	23.8	29.5	32.8	4	6	0
Short 2 Phase	32.1	14.0	18–52	20.3	27.0	44.5	4	6	0
Long Text	36.0	14.8	21–61	24.0	33.0	42.8	2	7	1
Long 2 Phase	38.8	19.6	19–71	25.0	28.5	53.0	2	8	0

D List of Options

We provide the full list of options presented on the survey.

- **Animal Rights, Welfare, and Services:** Protect animals from cruelty, exploitation and other abuses, provide veterinary services and train guide dogs.
- **Wildlife Conservation:** Protect wildlife habitats, including fish, wildlife, and bird refuges and sanctuaries.
- **Zoos and Aquariums:** Support and invest in zoos, aquariums and zoological societies in communities throughout the country.
- **Libraries, Historical Societies and Landmark Preservation:** Support and invest public and specialized libraries, historical societies, historical preservation programs, and historical estates.
- **Museums:** Support and invest in maintaining collections and provide training to practitioners in traditional arts, science, technology, and natural history.
- **Performing Arts:** Support symphonies, orchestras, and other musical groups; ballets and operas; theater groups; arts festivals; and performance halls and cultural centers.
- **Public Broadcasting and Media:** Support public television and radio stations and networks, as well as providing other independent media and communications services to the public.
- **Community Foundations:** Promote giving by managing long-term donor-advised charitable funds for individual givers and distributing those funds to community-based charities over time.
- **Housing and Neighborhood Development:** Lead and finance development projects that invest in and improve communities by providing utility assistance, small business support programs, and other revitalization projects.
- **Jewish Federations:** Focus on a specific geographic region and primarily support Jewish-oriented programs, organizations and activities through grantmaking efforts
- **United Ways:** Identify and resolve community issues through partnerships with schools, government agencies, businesses, and others, with a focus on education, income and health.
- **Adult Education Programs and Services:** Provide opportunities for adults to expand their knowledge in a particular field or discipline, learn English as a second language, or complete their high school education.
- **Early Childhood Programs and Services:** Provide foundation-level learning and literacy for children prior to entering the formal school setting.
- **Education Policy and Reform:** Promote and provide research, policy, and reform of the management of educational institutions, educational systems, and education policy.

- **Scholarship and Financial Support:** Support and enable students to obtain the financial assistance they require to meet their educational and living expenses while in school.
- **Special Education:** Provide services, including placement, programming, instruction, and support for gifted children and youth or those with disabilities requiring modified curricula, teaching methods, or materials.
- **Youth Education Programs and Services:** Provide programming, classroom instruction, and support for school-aged students in various disciplines such as art education, STEM, outward bound learning experiences, and other programs that enhance formal education.
- **Botanical Gardens, Parks, and Nature Centers:** Promote preservation and appreciation of the environment, as well as leading anti-litter, tree planting and other environmental beautification campaigns.
- **Environmental Protection and Conservation:** Develop strategies to combat pollution, promote conservation and sustainable management of land, water, and energy resources, protect land, and improve the efficiency of energy and waste material usage.
- **Diseases, Disorders, and Disciplines:** Seek cures for diseases and disorders or promote specific medical disciplines by providing direct services, advocating for public support and understanding, and supporting targeted medical research.
- **Medical Research:** Devote and invest in efforts on researching causes and cures of disease and developing new treatments.
- **Patient and Family Support:** Support programs and services for family members and patients that are diagnosed with a serious illness, including wish granting programs, camping programs, housing or travel assistance.
- **Treatment and Prevention Services:** Provide direct medical services and educate the public on ways to prevent diseases and reduce health risks.
- **Advocacy and Education:** Support social justice through legal advocacy, social action, and supporting laws and measures that promote reform and protect civil rights, including election reform and tolerance among diverse groups.
- **Development and Relief Services:** Provide medical care and other human services as well as economic, educational, and agricultural development services to people around the world.
- **Humanitarian Relief Supplies:** Specialize in collecting donated medical, food, agriculture, and other supplies and distributing them overseas to those in need.
- **International Peace, Security, and Affairs:** Promote peace and security, cultural and student exchange programs, improve relations between particular countries, provide foreign policy research and advocacy, and United Nations-related organizations.
- **Religious Activities:** Support and promote various faiths.
- **Religious Media and Broadcasting:** Support organizations of all faiths that produce and distribute religious programming, literature, and other communications.
- **Non-Medical Science & Technology Research:** Support research and services in a variety of scientific disciplines, advancing knowledge and understanding of areas such as energy efficiency, environmental and trade policies, and agricultural sustainability.
- **Social and Public Policy Research:** Support economic and social issues impacting our country today, educate the public, and influence policy regarding healthcare, employment rights, taxation, and other civic ventures.

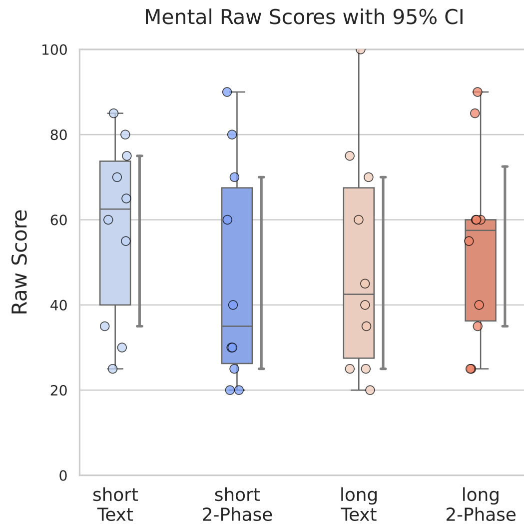


Fig. 22. Mental Demand Raw Score: Across all four experiment groups, participants' reported mental demand is spread across a wide range with many participants experiencing high mental demand.

E Detailed Qualitative Cognitive Load Breakdown

In addition to the discussion on cognitive load sources presented in the main text, we provide additional details on the six cognitive dimensions. Among all dimensions, we also provide the codes representing different types of demand in a table form. The shaded cells represent the percentage of participants citing each source of mental demand, allowing for comparison within columns. The abbreviations in the columns: ST (Short Text Interface), S2P (Short Two-phase Interface), LT (Long Text Interface), and L2P (Long Two-phase Interface). Short and Long refer to the sum across both interfaces; Text and Inter refer to the sum across both survey lengths. We include Sparklines for comparisons across these experiment groups. Future studies can use these as initial codebooks to conduct interface studiues on preference construction.

F Sources of Mental Demand

Mental demand refers to the amount of mental and perceptual activity required to complete a task. Table F lists all the mental demand codes. Figure 22 shows the boxplot of participant's subscale response.

F.1 Sources of Physical Demand

Physical demand refers to the physical effort required to complete a task, such as physical exertion or movement. Most participants reported minimal physical demand ($N = 32$), reflected in the low NASA-TLX physical demand scores (Figure 23). Notably, 11 out of 20 participants who used the two-phase interface mentioned physical demand from using the mouse, reflecting their increased interaction with the interface. This is further supported by the raw NASA-TLX physical demand scores (Figure 23), which show a significant visual difference between short and long two-phase interfaces as well as between text and two-phase interfaces in long surveys. Table 3 presents all the relevant codes across experiment groups.

Table 2. This table lists all the causes participants mentioned as contributing to their Mental Demand.

[Mental Demand]	Total	Version ST	SI	LT	LI	Experiment Conditions Short	Long	Text	Inter
Budget Management	14	3	3	5	3	6	8	8	6
Budget within limited credit	5	2	2	1	0	4	1	3	2
Track remaining credits	10	2	2	3	3	4	6	5	5
Maximize credit usage	8	2	3	2	1	5	3	4	4
Operational	12	3	2	4	3	5	7	7	5
Strategic	7	2	4	1	0	6	1	3	4
Preference Construction	39	10	9	10	10	19	20	20	19
Determining relative preference	16	4	4	5	3	8	8	9	7
Option prioritization	17	6	4	3	4	10	7	9	8
Precise resource allocation	30	9	6	9	6	15	15	18	12
Narrow - Consider a few options/personal causes	23	6	6	8	3	12	11	14	9
Broad - Considering all options or higher order values	23	5	5	4	9	10	13	9	14
Demand from Experiment Setup	24	6	6	6	6	12	12	12	12
Many options on the survey	6	0	0	3	3	0	6	3	3
QS Mechanism	4	2	0	2	0	2	2	4	0
Recalling experience or understanding options	20	5	6	4	5	11	9	9	11
Justification or Reflection on response	8	2	2	1	3	4	4	3	5
External Factors	12	3	1	4	4	4	8	7	5
Demand due to Interface	8	2	2	0	4	4	4	2	6
Increase	4	1	1	0	2	2	2	1	3
Decrease	4	1	1	0	2	2	2	1	3

Table 3. Physical Demand Causes: Most participants expressed little or no physical demand. Results reflected that participants in the long two-phase interface required more actions, hence the higher mention of mouse usage as a source.

[Physical]	Total	Version ST	SI	LT	LI	Experiment Conditions Short	Long	Text	Inter
Reading	4	0	2	1	1	2	2	1	3
Mouse	16	3	5	2	6	8	8	5	11
Vertical Screen	4	1	0	1	2	1	3	2	2
None/Little	32	8	9	8	7	17	15	16	16

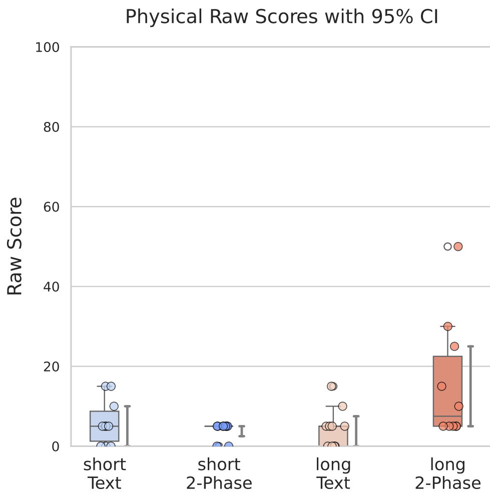


Fig. 23. Physical Demand Raw Score: Participants other than the long two-phase interface reported minimal physical demand. The long two-phase interface had the highest physical demand, likely due to increased mouse clicks and extended time spent looking at the vertical screen.

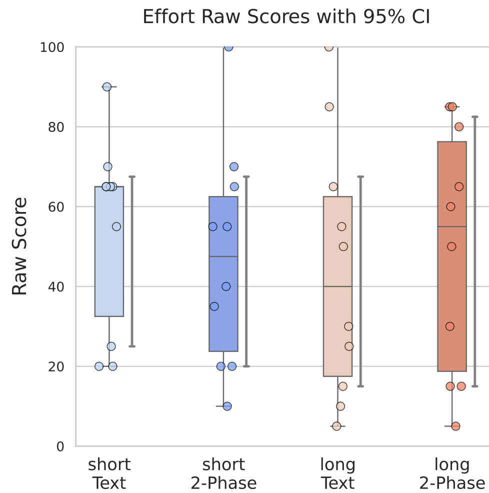


Fig. 24. Effort Raw Score: Effort scores shows indifference across groups.

Table 4. Effort Sources: Participants using the text interface focused more on operational tasks, while those using the two-phase interface focused more on strategic planning.

[Effort]	Total	Version				Experiment Conditions			
		ST	SI	LT	LI	Short	Long	Text	Inter
Operational	21	6	5	8	2	11	10	14	7
Strategic	28	6	8	5	9	14	14	11	17
Personal	22	4	7	5	6	11	11	9	13
Global	11	2	3	2	4	5	6	4	7
None/Little/a bit	9	2	1	3	3	3	6	5	4

F.2 Source of Effort

Effort refers to how hard participants felt they worked to achieve the level of performance they did. Since effort includes both mental and physical resource intensity, refer to ?? and Appendix F.1 for definitions. Raw NASA-TLX effort scores (Figure 24) showed a similar spread across experiment groups, the qualitative analysis showed more distinction that participants using the two-phase interface considered options more comprehensively and felt less effort on completing operational tasks, similar to what we found on mental demands (Section ??). Table 4 contains codes.

F.2.1 Effort Source #1: Operational Tasks. 14 of the 20 participants using the text interface mentioned Operational Tasks as effort sources, compared to 7 using the two-phase interface, with the lowest mention by the long two-phase interface group (N = 2). Quotes below illustrated participants putting in effort to manipulate the interface.

> *I wanted to bump up (an option) maybe to 4 or <option> to 5 and realize I couldn't. [...] that would be effort came in of how do I want to really rearrange this to make it (the budget spending) maximize?*
>
> *– S029, short text interface*

> *So it was like it was very ... I have to put a lot of effort in terms of you know ... think about each dimension that if I give one credit to <option name> whether it will affect my credits on <another option name>.* *– S005, long text interface*

F.2.2 Effort Source #2: Strategic Planning. Different from Operational Tasks, 11 participants in the text interface compared to 17 participants described strategic planning as sources of effort, with almost all participants ($N = 9$) from the long two-phase interface. We further categorize strategic planning into *narrow* and *broad* scopes as we did for mental demand ??. Participants using the two-phase interface ($N = 7$) had nearly mentioned double ($N = 4$) times regarding global strategies. For example:

> *And really the bulk of the effort was how to rank order these (options) and allocate the resources behind the upvotes so that I can accurately depict what I want ... say, a committee to focus on and allocate actual fungible resources, too.* *– S019, long two-phase interface*

Table 5. Performance Causes: Most causes are shared across experiment conditions. We provided qualitative interpretations of their own performance assessments.

[Performance]	Total	Version				Experiment Conditions			
		ST	SI	LT	LI	Short	Long	Text	Inter
Operational Action	13	2	3	3	5	5	8	5	8
Budget Control	6	1	1	2	2	2	4	3	3
Preference Reflection	6	1	1	2	2	2	4	3	3
Limited Resources	5	1	2	1	1	3	2	2	3
Social Responsibility	8	2	2	2	2	4	4	4	4
Decision maker	7	1	2	2	2	3	4	3	4
Outcome Uncertainty	7	1	2	2	2	3	4	3	4
Performance Assessment									
Did their best	8	2	1	3	2	3	5	5	3
Feel Good	17	3	5	3	6	8	9	6	11
Good Enough	10	2	2	3	3	4	6	5	5

F.3 Source from Performance

Performance refers to a person's perception of their success in completing a task. Lower values mean good perceived performance; higher values mean poor perceived performance. We found minimal qualitative differences between experiment groups regarding factors influencing perceived performance. Two influencing factors emerged: *Operational Actions* and *Social Responsibility*. Despite most participants reporting positively on their performance, nuances exist in how different groups interpret their performance.

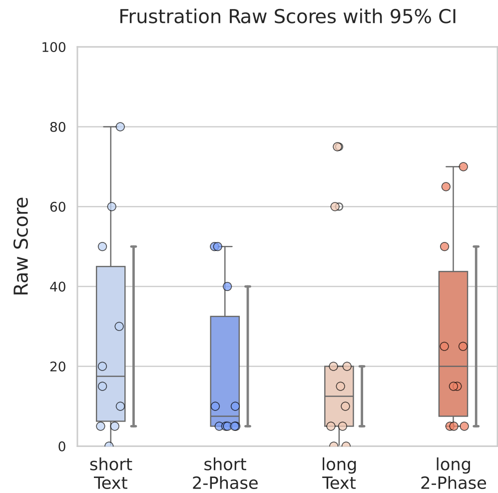


Fig. 25. Frustration Raw Score: Participants other than the long text interface highlighted several operational tasks that led to frustration. All groups share causes from strategic planning.

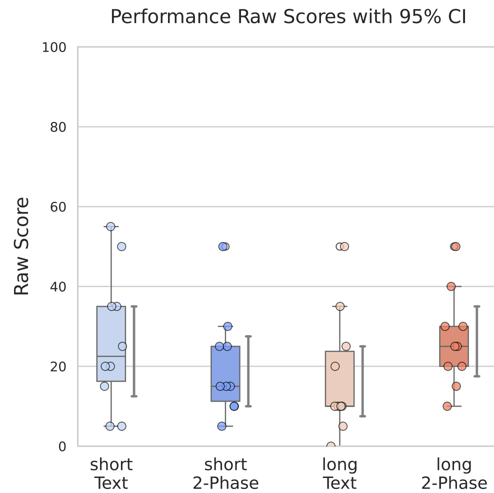


Fig. 26. Performance Demand Raw Score: Participants showed indifferent performance raw scores across experiment conditions, all trending toward satisfactory.

F.3.1 Operational Actions. Operational actions, like the theme presented in temporal demand, refer to specific, executable procedures participants perform in the survey. This could involve: pressure to spend all credits or stay within budget ($N = 6$), fears that final vote choices did not reflect true preferences($N = 5$), or concerns that they had finished the task inefficiently ($N = 6$).

F.3.2 Social Responsibility. Social responsibility-based concerns around performance came up when participants reflected on how their final vote counts would be perceived by others (S041 *I don't want people to think that I just like don't care about <ethnicity> people at all*) or influence real-world decision-making (S027 *Some of these things might . . . have outcomes that I didn't foresee*).

All groups cited social responsibility as source to evaluate effort. Raw NASA-TLX scores (Figure 26) show participants had indistinguishable performance scores. This aligns with the interview results where most participants felt positive about their final submission.

To dig deeper, we also analyzed participants' language when they described their performance. Expressions like "good enough" may be indicative of satisficing behaviors – our results suggest participants are satisfied at similar rates regardless of the interface. 1/4 of the participants in the text interface expressed "done their best," referring to exhausting their effort. Participants who used a two-phase interface were generally more positive about their final outcome – they were twice as likely to report "feeling good" about their final results ($N = 11$ v.s. $N = 6$).

F.4 Temporal Demand

Table F.4 lists all the mental demand codes.

F.5 Frustration

Table F.5 lists all the mental demand codes.

Table 6. Temporal Demand Sources: Decision-making and Operational Tasks are the main causes. Participants framed their decision-making sources differently.

[Temporal]	Total	Version				Experiment Conditions			
		ST	SI	LT	LI	Short	Long	Text	Inter
Budget Management	4	0	1	1	2	1	3	1	3
Decision Making	15	5	2	3	5	7	8	8	7
Affirmative	9	0	2	2	5	2	7	2	7
Negative	8	5	1	2	0	6	2	7	1
Operational	16	5	6	3	2	11	5	8	8
Task completion	8	2	2	3	1	4	4	5	3
Being efficient	8	3	4	0	1	7	1	3	5

Table 7. Frustration Sources: Frustration comes from different levels of strategic operations or operational tasks.

[Fustration]	Total	Version				Experiment Conditions			
		ST	SI	LT	LI	Short	Long	Text	Inter
Strategic	17	4	4	5	4	8	9	9	8
Higher-level	11	3	2	3	3	5	6	6	5
x Conflict between personal preference and broader society and common values	6	1	1	2	2	2	4	3	3
x Trade-offs among all options	8	3	1	2	2	4	4	5	3
Lower-Level	10	3	3	2	2	6	4	5	5
x Conflict between personal preference and	4	1	2	0	1	3	1	1	3
x Trade-offs among a few options	8	2	2	2	2	4	4	4	4
Operational	15	4	5	2	4	9	6	6	9
Credit management	6	2	3	1	0	5	1	3	3
Adhering to the Quadratic Mechanism	5	2	1	1	1	3	2	3	2
Deciding number of votes for an option	4	2	0	0	2	2	2	2	2
Making multiple decisions	3	2	0	0	1	2	1	2	1
Understanding Option	4	0	3	0	1	3	1	0	4
None/Little	16	4	5	5	2	9	7	9	7

G Additional voting behavior data

In this section, we describe the additional voting behavior that we observed. The reason why we decided to focus on the percentage of remaining credits comes from prior literature 'scarcity frames value' [101], a driver that makes researchers believe makes quadratic voting more accurate [4]. We did not follow Quarfoot et al. [6] in counting accumulated votes over time due to varying total times across individuals.

We observed the number of vote adjustments given a remaining vote credit percentage. Figure 27 showed all the voting actions over the remaining credit for the four experiment conditions. Here we see two distinct patterns between the short survey and the long survey in terms of participant behaviors. In long surveys, participants exhibited more actions both when the budget was abundant and when it began to run out. This pattern was more pronounced with the long two-phase interface. This difference is why we further focused on the long QS group.

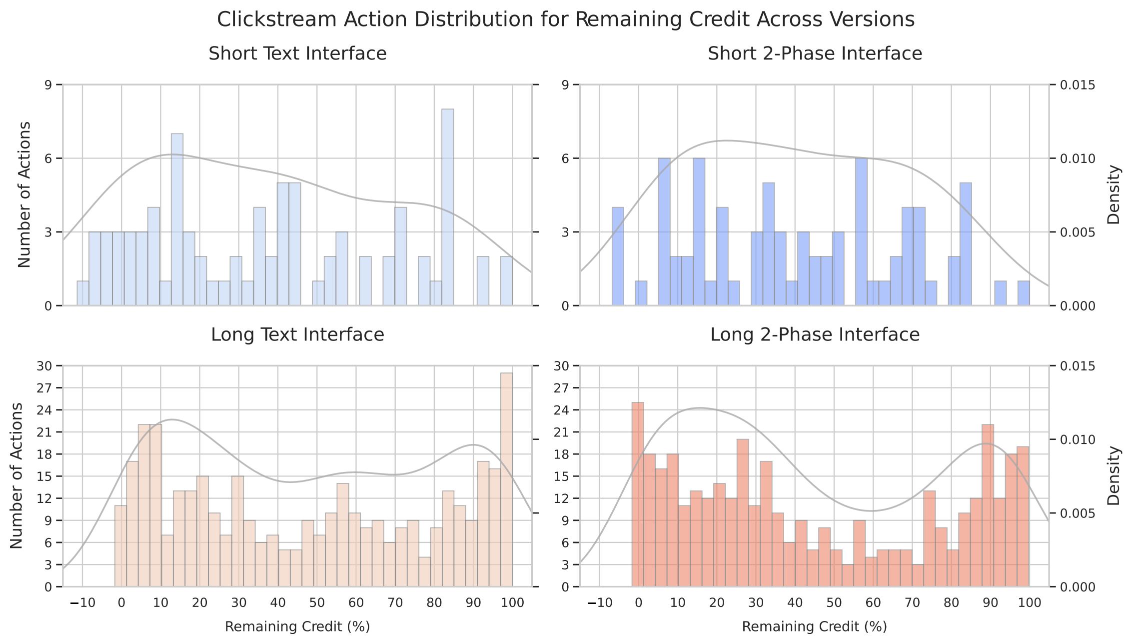


Fig. 27. This plot counts the number of voting actions when there are x percentages of credits remaining. A KDE plot is provided to help better understand the action distribution.

Figure 28 presents the comparison between when participants make small or large vote adjustments at different budget levels. Revisiting the KDE curve in the second row in Figure 27 and the curve of the second row in Figure 28 show a stronger bimodal distribution for small vote adjustments across interfaces. In fact, the bimodal distribution is more pronounced in the two-phase interface. This suggests that participants make small adjustments both at the beginning and toward the end of the QS. However, the two-phase interface shows more frequent and faster edits towards the end. In comparison, participants also made more large vote adjustments early on that spread more equally compared to the text interface. This indicates that participants had a clearer idea of how to distribute their credits across the options.

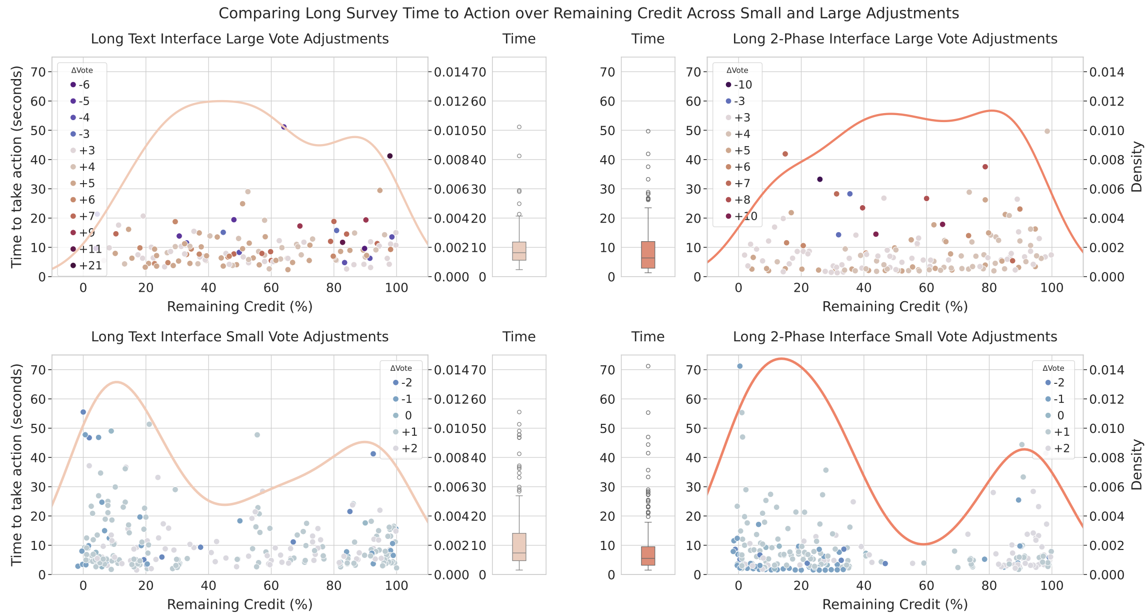


Fig. 28. This plot further separates participants' interaction behavior based on the number of votes participants adjusted. We observed a bimodal interaction pattern across long QS when small vote adjustments are made.

H Modeling NASA-TLX Weighted Scores and Subscales

In this section, we first describe the modeling approach for the NASA-TLX weighted scores and subscales, and then present all subscale results.

H.1 Modeling Approach

We modeled the NASA-TLX weighted scores and subscales using a hierarchical Bayesian ordinal regression model.

H.1.1 Dependent variables.

NASA-TLX weighted scores. are transformed from a continuous 0–100 scale to cognitive levels: low, medium, somewhat high, high, and very high, as described by Hart and Staveland [74]. This transformation helps the model adapt to sparse data. In our study, there were no participants who expressed "low" or "very high"; thus, we modeled the predictive variables as "medium," "somewhat high," and "high."

NASA-TLX subscale ratings. are transformed into ordinal groups using using minimum frequency binning [102]. Minimum frequency binning involves grouping adjacent response categories until each bin meets a predefined minimum number of observations. The subscale uses a 21-point Likert scale, with 40 participants, it makes the ordinal data very sparse. Minimum frequency binning mitigates this allowing similiar number of participants in each bin. We applied weighted bins across all participants within the same subscale, ensuring that each bin contained at least 10 participants.

H.1.2 Independent Variables. For this model, we used three independent variables: length (γ_i), interface type (β_I), and the interaction between the two (ϕ_{ij}). Length, categorized as "low" and "short," was modeled as an ordinal variable, as shown in Equation 4. Since there are only two categories, this approach allowed us to model the baseline length effect and the added effect of the longer length. Interface types were set up with hyperpriors, from which the interfaces were drawn. The interaction effect used a non-centered parameterization constrained by an LKJ prior to account for correlations, as described in Equation 5. Weakly informed priors were used for all parameters, as shown in Equations 6, 7, and 8.

H.1.3 Overall Model. We modeled the dependent variables using an Ordered Logistic (Equation 1). The Ordered Logistic model is particularly suited for ordinal outcome variables, where the categories have a natural order but the intervals between them are not necessarily equal. This model has two input parameters: η_i and τ. η_i is the latent predictor derived from a regression equation that incorporates the independent variables, demonstrated as Equation 2. The purpose of it, intuitively, is to model how specific independent variables pushes this latent value towards a higher or lower category. τ as modeded by Equation 3 are the cutpoints that demarcate the boundaries between the ordinal categories. This cutpoint draws from a normal distribution and being transformed to ensure that the thresholds are ordered. The Ordered Logistic model then compares η_i to τ to determine the probability of the observed outcome y_i falling into a specific ordinal category.

$$y_i \sim \text{OrderedLogistic}(\eta_i, \boldsymbol{\tau}) \tag{1}$$

$$\eta_i = \alpha + \gamma_i + \beta_I[I_i] + \phi_{ij} \tag{2}$$

$$\boldsymbol{\tau} \sim \text{OrderedTransform}(\mathcal{N}(0,1)^{K-1}) \tag{3}$$

$$\gamma_i = \mu_L + \beta_L \cdot L_i \tag{4}$$

$$\phi_{ij} = L_\Omega \cdot (\sigma_\phi \odot z_\phi) \tag{5}$$

Priors. We specify priors for all model parameters. The priors are defined as follows:

$$\mu_L, \mu_{\beta_L}, \mu_{\beta_I} \sim \mathcal{N}(0,1), \quad \sigma_{\beta_L}, \sigma_{\beta_I} \sim \text{Exponential}(1) \tag{6}$$

$$\beta_L \sim \mathcal{N}(\mu_{\beta_L}, \sigma_{\beta_L}), \quad \beta_I \sim \mathcal{N}(\mu_{\beta_I}, \sigma_{\beta_I}) \tag{7}$$

$$L_\Omega \sim \text{LKJ}(2), \quad \sigma_\phi \sim \text{Exponential}(1), \quad z_\phi \sim \mathcal{N}(0,1) \tag{8}$$

In Equation 6 and 7 we present the hyperpriors reflecting our belief that the mean effects of length and interface are centered around zero with a standard deviation of one. Hyperpriors were used to enable partial pooling where information is shared across different levels of the interface type, improving estimation accuracy especially in cases with limited data per group. Equation 8 describes the correlation metrix used for the interaction effect. The LKJ prior of 2 refers to a moderate correlation without being too restrictive allowing the model to learn appropriate levels of interaction terms. σ_ϕ ensuring that the variability of the interaction effects remains positive and allowing the model to flexibly adapt to different levels of interaction strength and z_ϕ were assigned to serves as a standardized component that, when scaled by σ_ϕ with the correlation matrix L_Ω captures the magnitude and the dependencies of the interaction terms effectively.

H.1.4 Posterier predictive plots. Our Bayesian model converged successfully, as evidenced by an $\hat{R}$ value of 1 for each subscale and the overall weighted tlx scores. We plotted the posterior predictive distribution of the model to compare the observed data with the model's predictions. Figure 29 shows the posterier predictions vs. observed data for the six subscales.

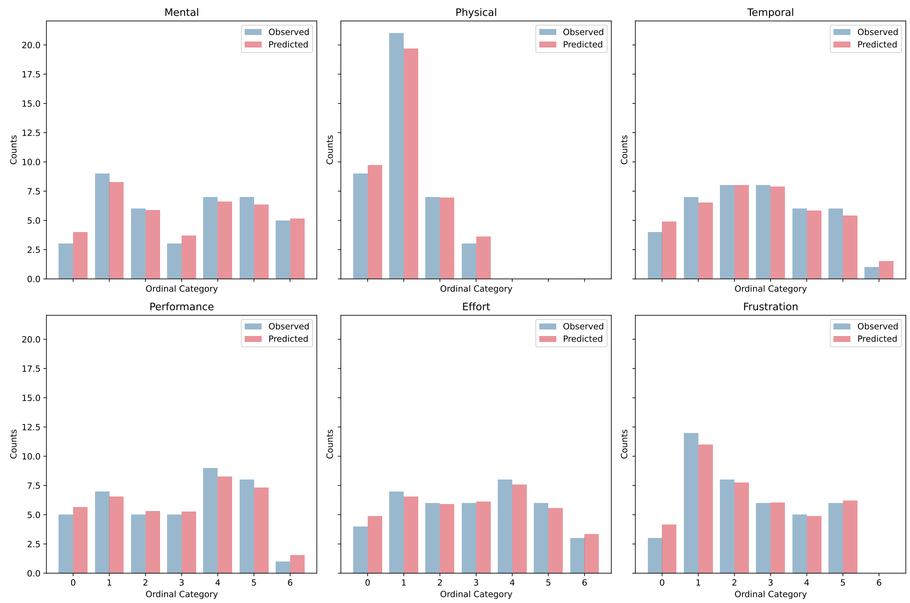


Fig. 29. Posterier Predictions vs. observed data for NASA-TLX subscales. The plot showed observed number of participants in each bin compared to the posterier predictions from the model. **Takeaway of the plot**: We believe that the model is reasonable at capturing the distribution of the subscales given the sparcity of the data.

H.1.5 Model Results. We conducted the Bayesian analysis using NumPyro, a widely used framework for Bayesian inference. We used Markov Chain Monte Carlo (MCMC) sampling, a method commonly applied in Bayesian inference. All the models showed that the Gelman-Rubin statistic ($\hat{R}$) parameters were equal to 1 across two chains, indicating that the multiple sampling chains converged. We present each subscale result and provide a short description of these results.

H.1.6 Mental Subscale. Figure 30 shows pairwise bayesian results from mental demand highlighted 70.4% of posterier probability that participants in the long two-phase condition had a higher mental demand compared to the short two-phase condition. On the other hand, the short text condition had a 74.5% posterior probability of having a higher mental demand compared to the short two-phase condition. This is additional evidence that prompted us to believe that the participants in the short two-phase participants benifited from the organization phase. The sheer number of added options in the long two-phase condition may have added additional demand to participants, leading to higher mental demand.

H.1.7 Physical Subscale. Figure 31 shows the pairwise comparison of the physical subscale. Noteable results shows that there is a 86.1% posterior probability that the long text condition had a lesser physical demand compared to the short text condition. This is counter intuitive as the long text participants actually traversed much higher edit distances. We are not clear what prompted their self reported value and requires future research.

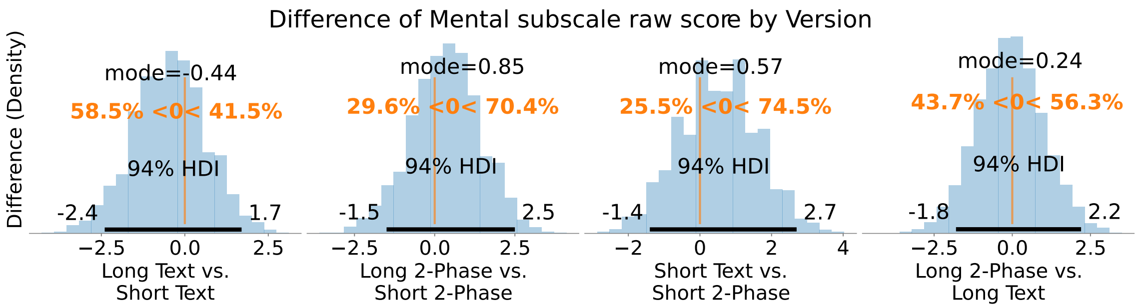


Fig. 30. Differences in the mental subscale scores by version. **Main Takeaway:** Participants in the long two-phase condition shows trends to increase mental demand compared to the short two-phase. Within the short text condition, participants in the short two-phase condition shows a trend to reduce mental demand.

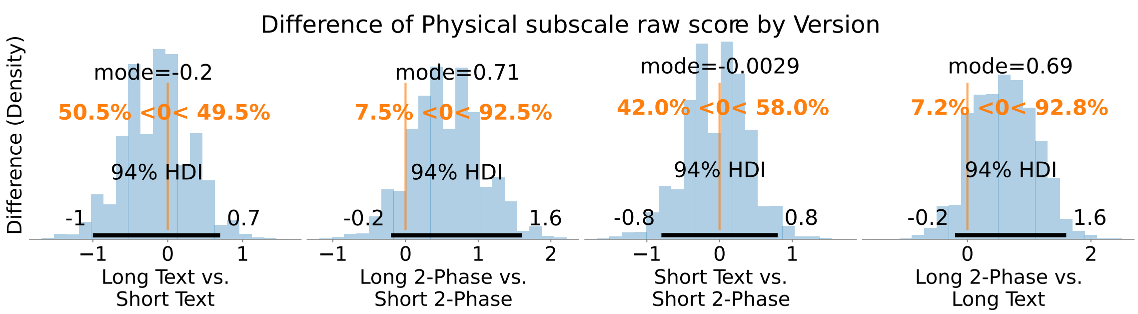


Fig. 31. Differences in the physical subscale scores by version. **Main Takeaway:** Participants in the long two-phase condition shows trends to increase physical demand compared to short two-phase and long text dispite the long text participants traversing higher edit distances.

H.1.8 Temporal Subscale. Figure 32 shows the pairwise comparison of the temporal subscale. The results show that the long two-phase condition once again had a 74.6% posterior probability of having a lower temporal demand compared to the short text condition. Conversely, participants in the long two-phase condition had a 71.1% posterior probability of having a higher temporal demand compared to the short two phase condition, reflecting the longer time they took to complete the survey questions. We believe that the lower temporal demand in the long two-phase condition are potential indicators of participant's satisficing behavior.

H.1.9 Performance Subscale. We omit the pairwise comparison of the performance subscale due to the mixed signals. We focused on the qualitative results analyzed in the main text.

H.1.10 Effort Subscale. We omit the pairwise comparison of the effort subscale due to its similarity to the mental demand subscale.

H.1.11 Frustration Subscale. Figure 33 shows the pairwise comparison of the frustration subscale. The results show that the long two-phase condition had a 68.3% posterior probability of having a higher frustration compared to the short two-phase condition, likey due to the added number of options to assess.

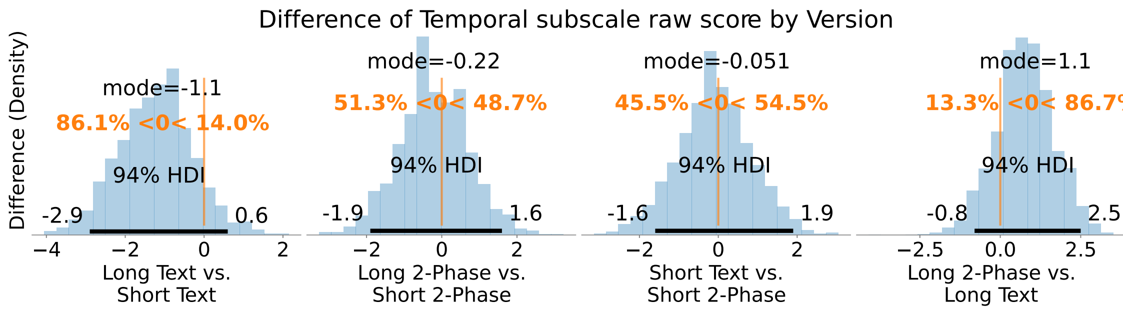


Fig. 32. Differences in the temporal subscale scores by version. **Main Takeaway:** Participants in the long text condition shows a trend that it reduces temporal demand compared to the short text condition and the long two-phase condition.

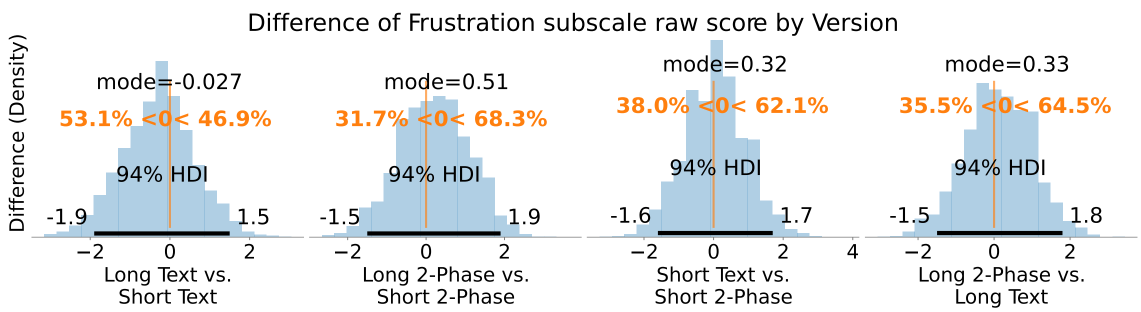


Fig. 33. Differences in the frustration subscale scores by version. **Main Takeaway:** The model does not see a significant difference in the frustration subscale between experiment groups other than a trend for participants in the long two-phase condition to have higher frustration than the short two-phase participants.

I Modeling Total Time

In this section, we discuss how we modeled the total time per option for each experimental condition.

I.0.1 Dependent Variables. Total time (T_i) refers to the time participants spent on each option, including the time allocated to the organization phase, where participants categorized or reordered options before voting.

I.0.2 Experimental Conditions. We categorize the data into four experimental conditions: Short Text, Short Two-Phase, Long Text, and Long Two-Phase. These conditions are indexed by k, and separate submodels are fit for each condition.

I.1 Modeling Approach

We modeled the total time for each experimental condition using separate Gamma likelihood models. The Gamma distribution is well-suited for modeling positive continuous data, such as time measurements, which are often skewed and strictly positive. Equation 9 shows the model for the total time. The shape parameter α_k and rate parameter β_k were each assigned priors drawn from their own Gamma distributions, as described in Equations 10 and 11.

$$T_i \sim \text{Gamma}(\alpha_k, \beta_k) \tag{9}$$

$$\alpha_k \sim \text{Gamma}(2.0, 0.5) \tag{10}$$

$$\beta_k \sim \text{Gamma}(1.0, 1.0) \tag{11}$$

J Modeling edit distance

In this section, we describe the details for the three models we used to analyze the edit distance data.

J.1 Model 1: Edit Distance per Option

J.1.1 Dependent variables. The dependent variable for this model is the edit total distance accumulated for an option D_i. Distance is a positive continuous variable.

J.1.2 Independent variables. The independent variables for this model are the length of the option L_i, modeled as a ordinal variable (Equation 15); interface type I_i, modeled as a categorical variable; user effect U_i as categorical variables. The ordinal variable L_i consists of a intercept μ_L and added effect β_L, given the interface ordinal value. Since we only have two interfaces, we do not have to worry about the interval between two or more interfaces. Priors are weakly informed in Equation 18. We reparamtereized U_i given the sparser sample from each participant. This is written in Equations 17. Both reparameterization contains an intercept and scaling of the effect due to this user. This will imporve sampling efficiency and help the model converge. Relavent priors are written in Equations 18 and 20. We added an interaction effect between length and interface type ϕ_{ij} described in Equation 16. Similiar to cognitive load model, the interaction effect used a non-centered parameterization constrained by an LKJ prior to account for correlations. Priors for the interaction effect is listed in Equations 19 and 21. Detailed description can be found in Appendix H.

J.1.3 Overall model and Likelihood function. We modeled the dependent variable using an Exponential distribution (Equation 12). Since Exponential distribution takes in a positive value, we transformed it as Equation 13. The observed outcome variable D_i represents the response for the i-th observation parameterized by the latent predictor η_i. η_i is described in Equation 14 as the regression with length, interface, the interaction effect and the interface.

$$D_i \sim \text{Exponential}(\lambda_i) \tag{12}$$

$$\lambda_i = \exp(\eta_i) \tag{13}$$

$$\eta_i = \gamma_i + \beta_I[I_i] + \phi_{ij} + U_i \tag{14}$$

$$\gamma_i = \mu_L + \beta_L \cdot L_i \tag{15}$$

$$\phi_{ij} = L_\Omega \cdot (\sigma_\phi \odot z_\phi) \tag{16}$$

$$U_i = \mu_U + \sigma_U \cdot z_U \tag{17}$$

Priors are defined as:

$$\mu_L, \mu_I, \mu_U, \beta_L, \beta_I, z_\phi, z_U \sim \mathcal{N}(0, 1) \tag{18}$$

$$\sigma_\phi \sim \text{HalfNormal}(0.5) \tag{19}$$

$$\sigma_U \sim \text{Exponential}(0.5) \tag{20}$$

$$L_\Omega \sim \text{LKJ}(3) \tag{21}$$

J.1.4 Posterier predictive plots. Our Bayesian model converged successfully, as evidenced by an $\hat{R}$ value of 1 in the model summary. We plotted the posterior predictive distribution for the edit distance per option in Figure 34. This figure compares the models posterier predictive distribution with the observed data.

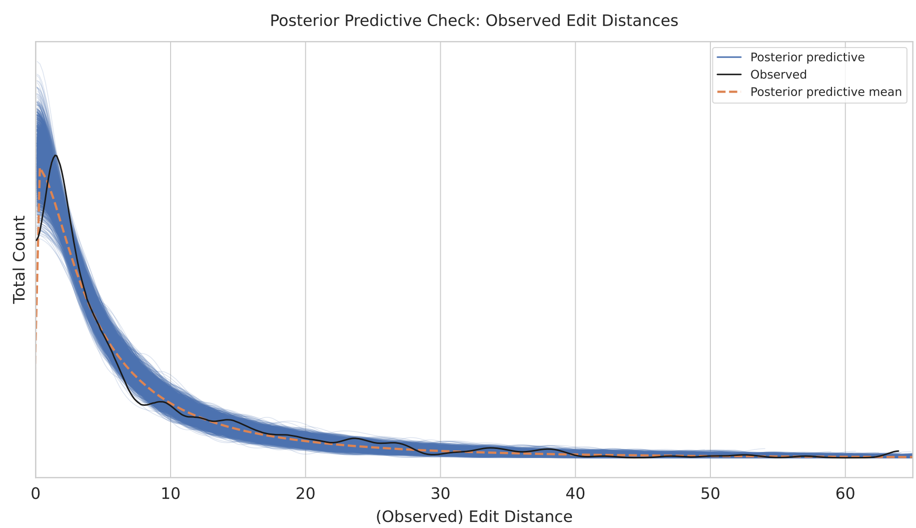


Fig. 34. Posterier Predictions vs. observed data for edit distance per option. Each blue line represents a draw from the posterior distribution, while the black line represents the observed data. Dotted line represents the mean of the posterier data. **Takeaway of the plot**: We believe that the model is reasonable at capturing the distribution.

J.2 Model 2: Edit Distance with Separate Mean and Variance Predictors

J.2.1 Dependent Variables. The dependent variable for this model is the edit distance (with directions) D_i, a positive edit distance refers to participants moving downward. A negtaitve edit distrance refers to a upward movement.

J.2.2 Overall Model. We modeled the dependent variable D_i using a Normal distribution (Equation 22). Since the goal of this model, unlike some, aims to model the variance since we believe participants in two-phase interface would exhibit less oscillation then the text interface. Hence, we model independent variables effecting both μ and σ independently for this analysis to exaime this hypothesis.

J.2.3 Independent Variables. The independent variables for this model are:

- **Length of the option** L_i: Modeled as an ordinal variable. Since we will be modeling both μ_i and σ of a Normal distribution, Equation 24 and 29 reflects the ordinal variable. Both formula consists of a intercept $\mu_{L,\mu}$, $\mu_{L,\sigma}$ and added effect $\beta_{L,\mu}, \beta_{L,\sigma}$, given the interface ordinal value. Since we only have two interfaces, we do not have to worry about the interval between two or more interfaces. Priors of both ordinal relationship are weakly informed in Equation 33 and 34

- **Interface type** I_i: Modeled as a categorical variable. Following the previous discussion, they are drawn from a hyperprior. We reparamtereized this independent variable given the added complexity of this model. This is written in Equations 25 and 30. Both reparameterization contains an intercept and scaling of the effect due to this interface. Relavent priors are written in Equations 33, 34, and 35.
- **User effect** U_i: Users are modeled as categorical variables. Following the interface, it is also reparamtereized as Equations 27 and 32. Priors are defined in Equations 33, 34, and 37
- **Interaction between length and interface type** ϕ_{ij}: Similiar to the interaction effect for cognitive load, we used a non-centered parameterization constrained by an LKJ prior to account for correlations. This is described by Equation 26 and 31. Refer to Appendix H for a more detailed explaination. Relevent priors are described in Equation 38 and 36. We relaxed the LKJ priors comapred to the cognitive load model given the complexity of the model allowing a lesser belief in correlation among the two variables.

J.2.4 Likelihood Function. Given these independent variables, we model both μ and σ as linear regressions. While we can directly model *mu* (Equation 23), we need to make sure *sigma* is strictly positive, we applied a transformation described in 28. Hence, both μ_i and $\log(\sigma_{\text{obs},i})$ now regresses on the linear combination of length, interface, interaction effect, and user effect.

$$D_i \sim \text{Normal}(\mu_i, \sigma_{\text{obs},i}) \tag{22}$$

$$\mu_i = \gamma_{\mu,i} + \beta_{I,\mu}[I_i] + \phi_{\mu,ij} + U_{\mu,i} \tag{23}$$

$$\gamma_{\mu,i} = \mu_{L,\mu} + \beta_{L,\mu} \cdot L_i \tag{24}$$

$$\beta_{I,\mu}[I_i] = \mu_{I,\mu} + \sigma_{I,\mu} \cdot \mathrm{I}_{\mu,I_i} \tag{25}$$

$$\phi_{\mu,ij} = L_{\Omega,\mu} \cdot (\sigma_{\phi,\mu} \odot z_{\phi,\mu}) \tag{26}$$

$$U_{\mu,i} = \mu_{U,\mu} + \sigma_{U,\mu} \cdot z_{U,\mu,i} \tag{27}$$

$$\log(\sigma_{\text{obs},i}) = \gamma_{\sigma,i} + \beta_{I,\sigma}[I_i] + \phi_{\sigma,ij} + U_{\sigma,i} \tag{28}$$

$$\gamma_{\sigma,i} = \mu_{L,\sigma} + \beta_{L,\sigma} \cdot L_i \tag{29}$$

$$\beta_{I,\sigma}[I_i] = \mu_{I,\sigma} + \sigma_{I,\sigma} \cdot \mathrm{I}_{\sigma,I_i} \tag{30}$$

$$\phi_{\sigma,ij} = L_{\Omega,\sigma} \cdot (\sigma_{\phi,\sigma} \odot z_{\phi,\sigma}) \tag{31}$$

$$U_{\sigma,i} = \mu_{U,\sigma} + \sigma_{U,\sigma} \cdot z_{U,\sigma,i} \tag{32}$$

J.2.5 Priors. Priors are defined as:

$$\mu_{L,\mu}, \mu_{I,\mu}, \mu_{U,\mu}, \beta_{L,\mu}, \beta_{I,\mu}, z_{\phi,\mu}, z_{U,\mu,i} \sim \mathcal{N}(0,1) \tag{33}$$

$$\mu_{L,\sigma}, \mu_{I,\sigma}, \mu_{U,\sigma}, \beta_{L,\sigma}, \beta_{I,\sigma}, z_{\phi,\sigma}, z_{U,\sigma,i} \sim \mathcal{N}(0,1) \tag{34}$$

$$\sigma_{I,\mu}, \sigma_{I,\sigma} \sim \mathrm{HalfNormal}(0.5) \tag{35}$$

$$\sigma_{\phi,\mu}, \sigma_{\phi,\sigma} \sim \mathrm{HalfNormal}(0.5) \tag{36}$$

$$\sigma_{U,\mu}, \sigma_{U,\sigma} \sim \mathrm{Exponential}(0.5) \tag{37}$$

$$L_{\Omega,\mu}, L_{\Omega,\sigma} \sim \mathrm{LKJ}(3) \tag{38}$$

J.2.6 Posterier predictive plots. Our Bayesian model converged successfully, as evidenced by an $\hat{R}$ value of 1 in the model summary. We plotted the posterior predictive distribution for the edit distance per option in Figure 35. This figure compares the models posterier predictive distribution with the observed data.

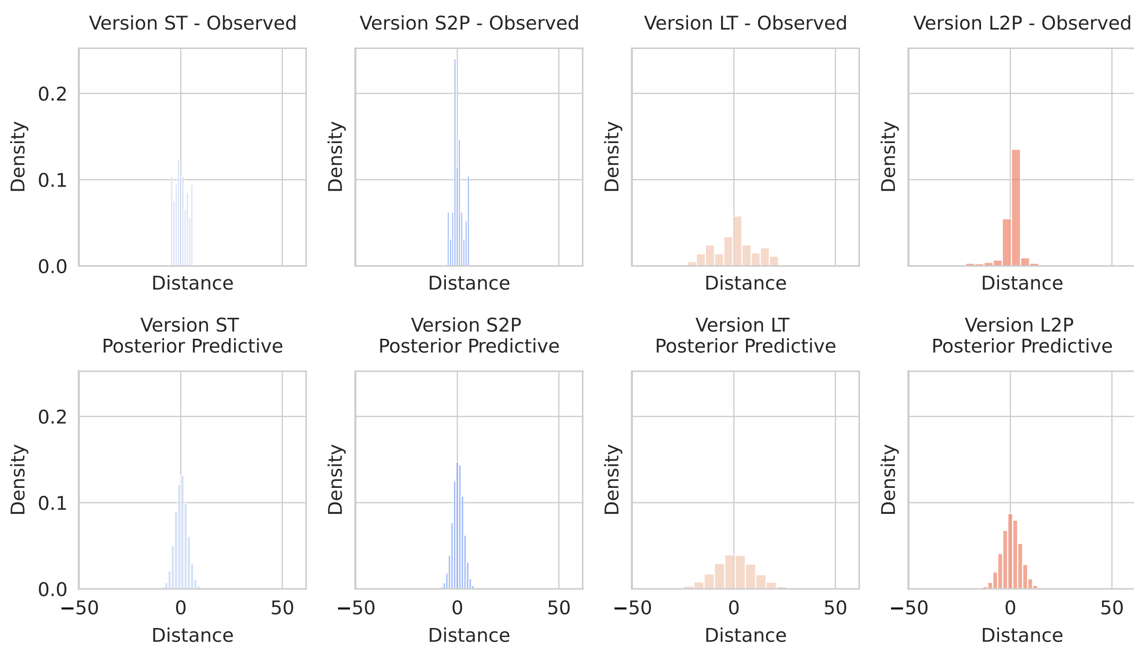


Fig. 35. Posterior Predictions vs. Observed Data for Edit Distance per Option. The first row represents the distribution of edit distance per version. The second row shows the posterier predictions after multiple draws **Takeaway of the plot**: We believe that the model is reasonable at capturing the shape of the distributions though being slightly conservative for extreme values at the center. Future model enhancements could re-modle them with a student-t distribution.

J.2.7 Model Results. Here we provide all pairwise comparisons for the variance which the main text only provided the comparison within the same survey length. Figure 36 shows the pairwise comparison of the variance of edit distance in the first row followed by the effect size in the second row. An notable result that we omit from the main text is that if we compare the variance between the long and short text, and the variance between the long and short two-phase, we see that the text group had three times the standard deviation compared to the two-phase group. This indicates that the organization phase minimize the added length of the survey.

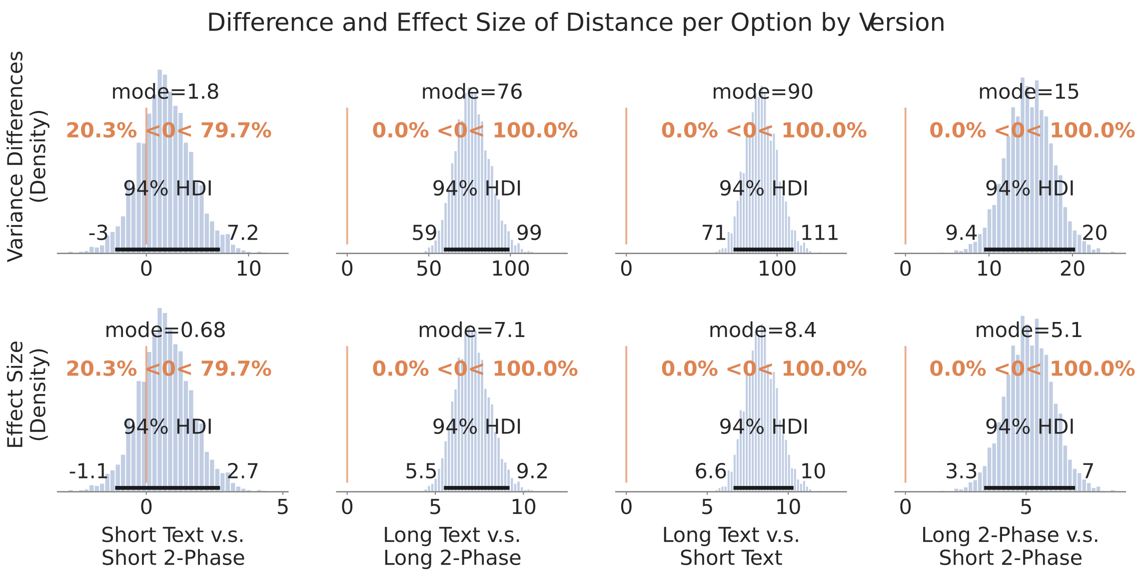


Fig. 36. Differences in the variance of edit distance by version. **The main takeaway:** In addition to the takaway from the main text, this plot shows that with two-phase interface, there is a reduction is edit distance when the number of option grows.

J.3 Model 3: Cumulative Edit Distance for long QS

J.3.1 Dependent Variables. The dependent variable for this model is the cumulative edit distance D_i. Cumulative edit distance is a positive continuous variable measured at each step within a version for each user.

J.3.2 Independent Variables. The independent variables for this model involve the following. Steps refers to the n-th step when completing QS (S_i), and interface version refers to the type of interface used (V_i). User-specific effects are included as (U_i). Both interface versions and user-specific effects are modeled with their own hyperpriors to capture variability across these groups.

Equation 46, refers to interface versions, $\beta_v[V_i]$ are drawn from a Normal distribution with hyperparameters defined in Equations 47 and 48 corresponding to the mean and variance of this distribution.

Instead of directly sampling U_i from a hyper distribution, we reparameterize it to account for limited data for each user. This reparameterization is presented in Equation 41. μ_U models the overall mean user effect from users, with σ_U used to capture variability in user effects (Equation 44). A standard normal random variable, Equation 45 introduced individual randomness for each user.

J.3.3 Overall Model and Likelihood Function. We modeled the dependent variable D_i using a Truncated Normal distribution (Equation 39). The observation-specific standard deviation, drawn from a Half-Normal distribution as described in Equation 42. The latent predictors μ_i is modeled as a regression equation (Equation 40). This equation reflects our intuition that the effects from versions and user differences are amplified by steps as the participants complete the survey. The intercept α_{shared} is assigned a prior described in Equation 43. The effect of users σ_U and version $\beta_v[V_i]$ are amplified by the step number S_i.

$$D_i \sim \text{TruncatedNormal}(\mu_i, \sigma_{\text{obs},i}, \text{lower} = 0) \tag{39}$$

$$\mu_i = \alpha_{\text{shared}} + \beta_v[V_i] \cdot S_i + U_i \cdot S_i \tag{40}$$

$$U_i = \mu_U + \sigma_U \cdot z_{U,i} \tag{41}$$

Priors used in this model are listed.

$$\sigma_{\text{obs},i} \sim \text{HalfNormal}(0.3) \tag{42}$$

$$\alpha_{\text{shared}} \sim \mathcal{N}(2.0, 0.5) \tag{43}$$

$$\mu_U, \sigma_U \sim \mathcal{N}(0, 1),\ \text{HalfNormal}(0.1) \tag{44}$$

$$z_{U,i} \sim \mathcal{N}(0, 1) \tag{45}$$

$$\beta_v[V_i] \sim \mathcal{N}(\mu_\beta, \sigma_\beta) \tag{46}$$

$$\mu_\beta \sim \mathcal{N}(0.05, 0.05) \tag{47}$$

$$\sigma_\beta \sim \text{HalfNormal}(0.1) \tag{48}$$

J.3.4 Posterier predictive plots. Our Bayesian model converged successfully, as evidenced by an $\hat{R}$ value of 1 in the model summary. We plotted the posterior predictive distribution for the cumulative edit distance in Figure 37. This figure compares the models posterier predictive distribution with the observed data.

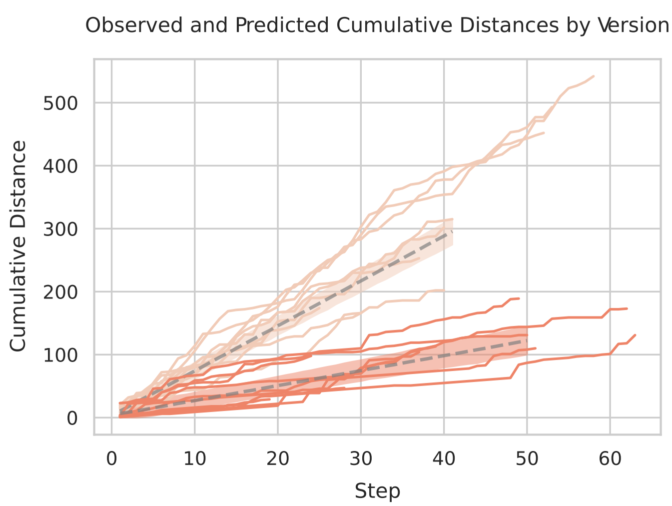


Fig. 37. Posterior Predictions vs. observed data for cumulative edit distance. The plot showed each observed user's cumulative edit distance in different shades for the two groups of participants. Dotted line represent the posterier predictive mean. **Takeaway of the plot**: We believe that the model is reasonable at capturing slop of the cumulative trends.