Budget, Cost, or Both? An Empirical Exploration of Mechanisms in Quadratic Surveys

Ti-Chung Cheng*
University of Illinois Urbana-Champaign
Urbana, USA
tcheng10@illinois.edu

Tiffany Wenting Li*
University of Illinois Urbana-Champaign
Urbana, USA
Stevens Institute of Technology
Hoboken, USA
wenting7@illinois.edu

Karrie Karahalios
University of Illinois Urbana-Champaign
Urbana, USA
kkarahal@illinois.edu

Hari Sundaram
University of Illinois Urbana-Champaign
Urbana, USA
hs1@illinois.edu

Abstract

The Quadratic Survey (QS) is an emerging preference elicitation method designed for collective intelligence contexts, where effective decision-making depends on capturing rich individual preferences, including both what people support and how strongly they care about them. QS sets itself apart from traditional tools through two core mechanisms: a fixed credit budget and a quadratic cost function. This study empirically examines the role of the two components in QS's effectiveness in isolation, by comparing the performance of QS and the Likert scale survey to two variants of QS: Unlimited QS, which removes the budget constraint, and Linear Survey, which replaces the quadratic cost with a linear function. In a controlled experiment with MTurk participants, survey responses from Unlimited QS and Linear Survey were evaluated alongside the existing QS and Likert scale responses reported in prior work, and all responses were compared against an incentive-compatible donation task. Hierarchical Bayesian analyses reveal that QS more effectively aligns expressed preferences with individuals' donation behavior, while omitting either component degrades performance. The results also confirm that both pairwise rankings and interval intensity differences between options captured by QS closely reflect individual behavior, outperforming both the Likert scale and Constant Sum-like surveys. These findings advance our understanding of QS and provide practitioners with an alternative tool to collect reflective individual preferences for collective decision-making contexts such as democratic engagement or public resource allocation.

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**.

Keywords

Quadratic Survey; Quadratic Voting; Preference Construction; Constant Sum Survey; Likert scale survey; Collective Decision Making

ACM Reference Format:
Ti-Chung Cheng, Tiffany Wenting Li, Karrie Karahalios, and Hari Sundaram. 2025. Budget, Cost, or Both? An Empirical Exploration of Mechanisms in Quadratic Surveys. In *Collective Intelligence Conference (CI 2025), August 04–06, 2025, San Diego, CA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3715928.3737474

1 Introduction

[T]he many, who are not as individuals excellent men, nevertheless can, when they have come together, be better than the few best people, not individually but collectively, just as feasts to which many contribute are better than feasts provided at one person's expense.

– Aristotle, Politics III

From renewable energy planning [16] and ride-sharing regulation [27] to corporate forecasting [11] and government budgeting [12], both public initiatives and academic studies [5, 33, 59, 60] show that effective collective intelligence (CI) depends on integrating truthful, diverse, and rich preference signals. Conventional tools such as Likert scale surveys, public polls, and one-person-one-vote schemes reduce individual choices to surface-level tallies, thereby obscuring the intensity and trade-offs that lead to better outcomes [33, 44, 47]. Quadratic Survey (QS) addresses this limitation by assigning respondents a fixed credit budget and applying a quadratic cost to each vote, prompting survey respondents to express not only which options they support but also how strongly they care about them [4, 7, 47]. QS can identify intense minority preferences when they outweigh mild majority support in resource-constrained scenarios such as public policy [7, 47], product design [7], or co-housing communities [31]. However, the higher cognitive demands of completing QS [8] have spurred researchers [4, 7] to propose plausible simplifications, such as replacing quadratic costs with linear ones or removing budgets entirely, in order to reduce participant frustration. Yet, there remains limited empirical understanding of whether such changes preserve the features that make QS effective and of why QS works.

*Both authors contributed equally to this research.

CI 2025, San Diego, CA, USA

ACM ISBN 979-8-4007-1489-4/25/08
https://doi.org/10.1145/3715928.3737474

Researchers [4, 7] attribute QS's effectiveness primarily to two components: a fixed budget constraint and a quadratic cost function. Prior research has predominantly compared QS with Likert scale surveys, highlighting its effectiveness in capturing realistic participant behaviors and clearly distinguishing preference intensity across many options when participants must prioritize under constraints [4, 7]. A closely related but simpler forced-choice approach, constant sum survey (CSS) [1] [42], has long utilized a linear constraint to require explicit prioritization among options. Despite the long-standing use of CSS and its linear budget constraint, prior research has not clearly isolated whether QS's effectiveness derives uniquely from its quadratic cost structure or if merely enforcing a fixed budget alone sufficiently generates the perceived trade-offs. Clarifying this distinction presents an opportunity to streamline QS, potentially reducing complexity without diminishing its effectiveness.

We introduce two additional survey variants designed to isolate the core components of QS and evaluate them alongside QS and Likert scale survey responses reported in prior work [7]. The first, which we call Unlimited QS (UQS), removes the budget constraint but retains the quadratic cost function. The second, Linear Survey (LS), retains the fixed budget but replaces the quadratic cost with a linear one. These two variants allow us to disentangle the individual effects of QS's budget constraint and cost function. In addition, to our knowledge, no prior study has evaluated QS using pairwise comparisons of rankings and preference intensities, providing a more rigorous empirical lens [9]. Formally, we ask:

RQ1. How effectively does QS capture participant preferences in pairwise rankings and preference intensities compared to Likert scale survey?

RQ2. How do the budget constraint and quadratic cost, as core components of QS, individually and jointly affect how well elicited preferences reflect participants' behavior?

To investigate these questions, we recruited 202 MTurk participants using stratified sampling to approximate U.S. census demographics, using a modified version of open-source QS software described in prior work [7]. Participants completed either UQS or one of three LS versions with credit budgets of 18, 54, or 162. Each survey asked how the local government should allocate resources across a set of societal issues. Participants then completed an incentive-compatible donation task. Together with open data from prior research, we developed two Bayesian models to assess how well survey responses align with participants' actual behavior. The first assesses whether each method captures the same pairwise ranking of preferences as the donation task. The second examines whether larger differences in reported survey preferences correspond to greater behavioral intensity, offering an interval-based perspective.

Our findings show that, in terms of pairwise ranking, QS outperforms the Likert scale survey, while both UQS and LS underperform relative to the Likert scale survey. For pairwise intensity differences, all methods perform similarly when the preference gap between two options is small. However, as the gap increases, QS, both its vote and credit-based measures, more reliably reflects behavioral intensity compared to other approaches. UQS performs similarly to the Likert scale survey whereas LS trails behind the Likert scale survey under these conditions. Their performance deteriorates further as preference differences grow. These results highlight the importance of both the credit budget and the quadratic cost function in effective preference elicitation. Our findings reaffirm QS's ability to represent individual preferences in resource-constrained contexts and surface the limitations of linear or unconstrained alternatives.

This paper makes two contributions: an empirical analysis of QS's core mechanisms and a modeling approach for evaluating survey-behavior alignment.

Empirical Contribution: This paper provides a more detailed empirical understanding of the Quadratic Survey mechanism by isolating and evaluating its two core components: fixed budgets and quadratic voting costs. Prior works [7, 15, 19, 47] established the theoretical and empirical grounds for QS's advantages but did not clarify whether the budget constraint, the quadratic cost, or both are necessary to achieve these advantages. We addressed this gap through controlled experiments framed around public resource allocation, using Bayesian modeling to examine how QS's budget and cost structures influence the alignment between stated survey preferences and participant donation behavior. Our findings reveal that removing either the quadratic cost or the credit budget weakens QS's ability to capture pairwise ranking and differences in preference intensity, suggesting that rather than simplifying QS through a linear cost function, future work should design interfaces to support survey respondents in expressing their preferences when answering QS.

Methodological Contribution: This paper introduces two Bayesian models to evaluate how different survey methods capture participants' preferences as reflected in their behavior. Prior evaluations have relied on submission-level or single-point behavioral comparisons, missing finer distinctions in preference structure. Our models evaluate both pairwise ranking and intensity by comparing stated pairwise preferences to donation behavior, drawing on common CSS pairwise comparison approaches [9, 25]. This method enables more precise empirical evaluation of preference elicitation tools and informs future studies on QS design and validation.

2 Related Work

In this section, we describe related work regarding QS and the quadratic mechanism embedded within. We then discuss forced choice surveying techniques that follow a linear constraint.

2.1 Quadratic Survey and the Quadratic Mechanism

QS uses a quadratic mechanism in which participants 'purchase' approval or disapproval votes to express their preference within a fixed budget. Because the vote cost increases quadratically, participants are discouraged from extreme responses and encouraged to allocate votes based on relative preference strength. Participants may assign varying numbers of positive and negative votes to reflect relative preferences. Survey designers compute collective preferences by summing votes for each option across participants.

Formally, a participant receives a QS with K options and a budget B, and may allocate n_k votes to each option k, with vote cost defined

[1] CSS is also referred to as chip-allocation survey, point-allocation survey, fixed-sum survey, and the budget pie method in the literature [24, 41, 49, 55, 56, 62].

by a quadratic function: $c_k = n_k^2$, where $n_k \in \mathbb{Z}$. Votes may be positive or negative to express support or opposition. Respondents must ensure that their total expenditure does not exceed their budget: $\sum_{k=1}^{K} c_k \leq B$. The collective preference for each option is then determined by summing the votes across all participants: $\sum_{i=1}^{S} n_{i,k}$, where S is the number of respondents and $n_{i,k}$ represents the votes allocated by participant i to option k.

The quadratic mechanism originates from economic theory, particularly for public goods allocation [23]. It gained prominence through **Quadratic Voting (QV)**, which addresses the "tyranny of the majority" by allowing individuals to express preference intensity rather than cast a binary vote [45]. Unlike QV, which produces binding decisions, QS gathers opinions to inform decision-makers or the public [4, 8].

Empirical studies have evaluated QS in settings ranging from lab experiments [7, 47] to policy polling [4, 26], education research [43] and community decision-making [31]. They show that QS elicits both rankings and ratings—an advantage over traditional survey methods [7]. QS also reduces extreme response biases, even on polarized topics, and captures richer preference data than Likert scale surveys [4, 7, 43, 47]. Recent studies have reported stronger alignment between QS-based stated preferences and observed behavior compared to Likert scale surveys [4, 7].

QS imposes higher cognitive demands to complete [8], with empirical studies showing that participants report medium to high cognitive load, especially when evaluating longer lists of options [4]. While heightened cognitive load can lead to deeper engagement with survey options, prior survey research literature suggests it also drives down participation rates and increases dropout [3, 17]. In response, researchers [4] have proposed simplifying QS by replacing the quadratic cost with a linear one. Yet no empirical study has systematically examined the trade-offs between quadratic and linear cost structures.

2.2 Linear Constraint-Based Collective Decision-Making Mechanisms

While QS's quadratic cost structure is novel, the practice of imposing fixed budgets in surveys has a long history in marketing, psychology, and political science. These methods require participants to allocate a limited number of points, tokens, or money across options, forcing trade-offs. Among these, *Constant Sum Survey* and *Knapsack Voting* (KV) are the most relevant comparisons to QS. Unlike other forced-choice techniques, such as MaxDiff [50, 57], Best-Worst Scaling [37], or conjoint analysis [1], CSS and KV impose explicit linear resource constraints, making them conceptually closer to the QS and LS examined in this study.

2.2.1 Constant Sum Survey. CSS has existed since the 1950s [10, 39, 53], originally designed as 100-point splits between two options [42] and later extended to multiple-option settings [24, 62]. In CSS, participants receive a fixed point budget (often 100) to distribute across K options, reflecting their relative perceived importance. Although survey platforms vary in their mechanism's implementation [36, 46, 54], the core constraint remains: respondents must stay within their allotted budget.

Studies show CSS elicits both ranking and rating information, making it useful in domains such as marketing and political science [9]. Validation against behavioral measures is mixed: CSS often aligns with pairwise comparisons [14], but can diverge from revealed preferences, as reflected in measures like willingness to pay (WTP) [38]. Despite these differences, CSS remains popular for capturing preference intensities within a linear budget constraint.

LS closely resembles CSS but differs in three minor ways. First, CSS does not typically allow negative point allocations. Second, many CSS implementations require participants to exhaust the full budget. Last, CSS is typically not framed as a vote allocation process, unlike QS, which emphasizes 'vote buying' as part of its interface metaphor. While LS can be reformulated to match CSS mathematically, for example, by interpreting the negative votes as additional disagreement options on the survey, or residual budgets as a dummy option, differences in framing may lead to distinct participant behaviors [30, 52]. Accordingly, we conservatively treat LS as distinct from CSS, though their structural similarities support methodological comparisons.

2.2.2 Knapsack Voting and participatory budgeting. KV is another forced-choice surveying technique developed for participatory budgeting, a process where community members express preferences for how public resources should be allocated [20, 21]. In KV, participants receive a fixed budget and select from options with predefined costs. Participants may choose any combination of options, as long as the total cost remains within budget. This approach requires participants to contribute predefined 'chunks' of budget following a linear relationship, which we do not explore in this study, as QS options do not necessarily come with defined costs.

3 Experimental Setup

This section adopts an experimental design consistent with prior research protocols [7], and was reviewed and approved by an IRB.

3.1 Study Design Details

To ensure comparability with prior work [7] for subsequent analysis, we retained the original between-subjects design and extended their survey software to implement two additional instruments, yielding four new experimental conditions. The study's survey context and donation task remain unchanged, while the procedural modifications are described in the following subsections. Figure 1 illustrates how this study fits within the broader experimental flow established by prior work.

Additional Experimental Conditions. We introduced four new experimental conditions, grouped into two categories: UQS and LS. In the UQS condition, we removed the budget constraint to isolate the effect of the quadratic cost function. In the LS conditions, we replaced the quadratic cost function with a linear one and subdivided the design into three budget levels:

- LS18: A small-budget LS with 18 credits
- LS54: A medium-budget LS with 54 credits
- LS162: A large-budget LS with 162 credits

Following prior work, we allocated two credits per option, allowing participants to assign up to ±2, which mirrors the expressive

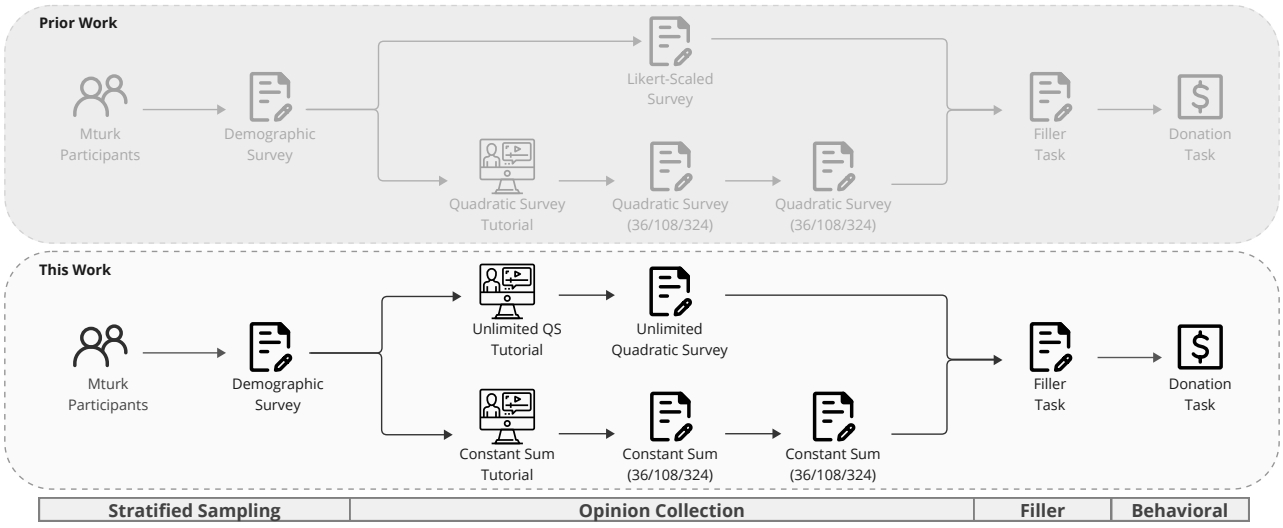


Figure 1: Experimental design. Our study mirrors the structure of prior work (top, [7]), differing only in the opinion collection methods (bottom). This study introduced UQS and LS conditions. The timeline below segments the procedure into four stages, highlighting how our approach mirrors prior work: stratified sampling, opinion collection, filler task, and behavioral task.

range of a 5-point Likert scale and enables them to express moderate intensity in either direction. With nine options, this corresponds to 18 credits for the LS18 condition. We then scaled the budgets using $O(K)$, $O(K^{1.5})$, and $O(K^2)$ to derive LS18, LS54, and LS162, respectively. For example, $2 \times 9^{1.5} = 54$ is LS54.

3.1.1 Survey content. The study frames the survey as a collective decision-making task, where participants express preferences across 9 societal issues such as education, environment, or health. Participants expressed their degree of preference by assigning positive or negative votes under the UQS or LS mechanism.

3.1.2 Surveying process and interface. Participants in both groups were first introduced to the survey and how to use it via a video tutorial. To ensure their understanding of the survey mechanism, participants were asked to complete a quiz with 5 multiple-choice questions. A minimum of three correct answers was required to proceed. We altered the questions based on the survey mechanisms. The interfaces for UQS and LS are shown in Figure 2.

3.1.3 Filler task and donation. After the survey, participants completed a filler task to reduce direct association between survey options and the charities in the donation task. Participants then donated to a set of charities, each representing a distinct cause.

3.1.4 Debrief and Compensation. After the study, a debriefing page informed participants of the study's purpose. Participants were compensated with $2.50 for their time.

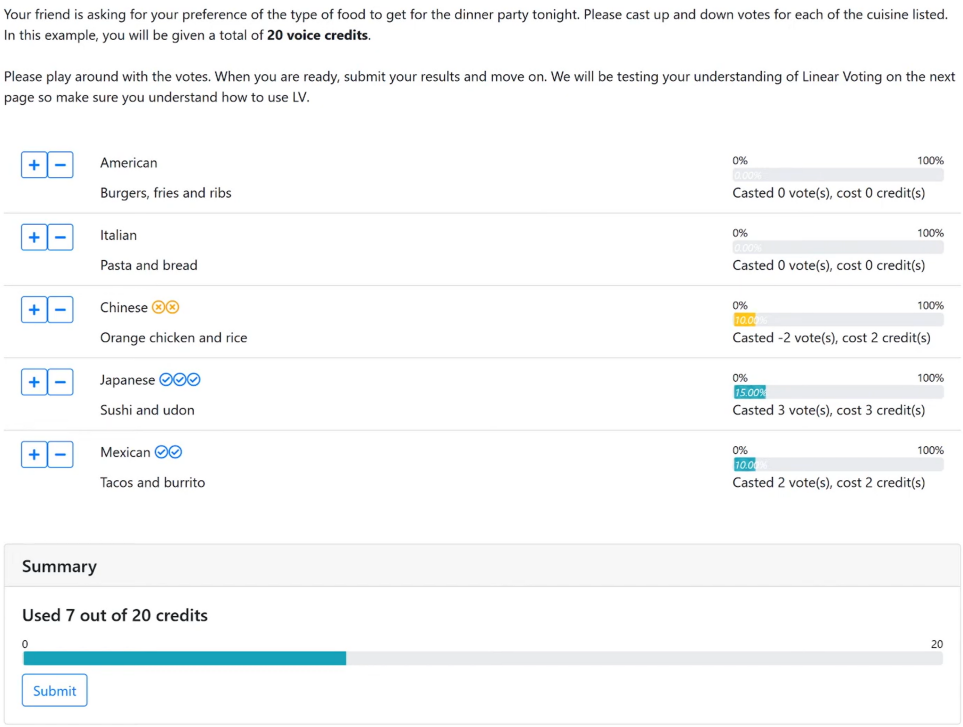
3.2 Participant Recruitment and Integrating Prior Data

This study includes both newly collected and publicly available data. We recruited 202 Amazon Mechanical Turk (MTurk) participants using stratified sampling to approximate U.S. census distributions across age, gender, income, and education. Participants were randomly assigned to one of the newly introduced experimental conditions (LS and UQS). This sampling strategy aimed to mitigate demographic imbalances in participation [48].

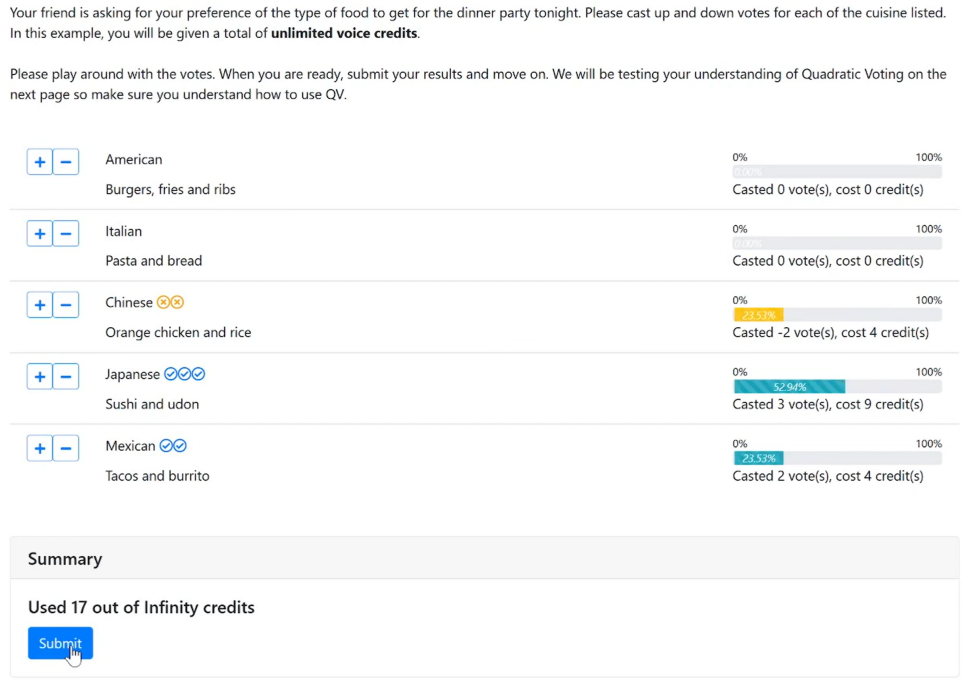
We obtained data from previous QS and Likert scale conditions published in [7] and included it in this study for comparative evaluation. It covered 219 MTurk participants across two survey types and four experimental conditions: a Likert scale survey and QS with three credit budgets (36, 108, and 324).

To support comparisons across methods, we distinguish between the number of votes assigned to each option (used in the original QS analysis) and the total credits spent per option, which better reflects the cost-based intensity embedded in the quadratic mechanism. We refer to these cost-based representations as QSC36, QSC108, and QSC324.

Altogether, our study evaluates four types of survey instruments: QS, Likert scale survey, LS, and UQS. As summarized in Table 1, we cover 11 experimental conditions, including three LS budget levels and three QS conditions evaluated using both vote- and cost-based measures. All survey responses were assessed against participant behavior in an incentive-compatible donation task, which serves as a shared behavioral benchmark.



(a) LS Interface: each additional vote is 1 credit



(b) UQS Interface: each additional vote is n^2 credits but does not have a budget constraint

Figure 2: Survey interfaces for the two additional conditions. Each screenshot shows an interactive sandbox that allows participants to practice the survey mechanism before completing the main task.

Condition	Budget	Cost Function	Description
Likert	–	–	A 5-point traditional ordinal-scale survey.
Donation	–	–	Incentive-compatible donation task used as behavioral benchmark for validating expressed preferences.
QS36	36	Quadratic	QS conditions with three different budgets ($O(K)$, $O(K^{1.5})$, $O(K^2)$). Participants expressed preferences by allocating votes, where the cost of each vote increased quadratically, deducted from the budget.
QS108	108	Quadratic	
QS324	324	Quadratic	
QSC36	36	Quadratic	Credits that participants contributed per option. Using the results from QS, QSC reflects the actual cost incurred per option to reflect perceived intensity of preference and explore alignment with donation outcomes.
QSC108	108	Quadratic	
QSC324	324	Quadratic	
LS18	18	Linear	Linear-cost versions of QS with budgets scaled as $O(K)$, $O(K^{1.5})$, and $O(K^2)$. Participants expressed preferences by allocating votes, where the cost of each vote increased linearly, deducted from the budget.
LS54	54	Linear	
LS162	162	Linear	
UQS	Unlimited	Quadratic	QS without a budget where participants expressed preferences by allocating votes, where the cost of each vote increased quadratically, but no budget constraint was enforced.
UQS Credits	Unlimited	Quadratic	Credits participants contributed per option. Using the results from UQS, credits reflect the actual cost incurred per option to reflect perceived intensity of preference and explore alignment with donation outcomes.

Table 1: Overview of survey conditions evaluated in the study, including budget levels, cost structures, and their modeling roles.

4 Modeling for Pairwise Ranking and Preference Intensity Analyses

Two recent empirical studies have evaluated whether elicited survey responses align with participant behavior, using outcomes such as charitable donations [4, 7] or letter-writing effort [4] as behavioral proxies for underlying preferences. One approach used Bayesian cosine similarity to compare high-dimensional response vectors with behavioral outcomes [7]; another applied linear regression to estimate the gap between stated and revealed preferences [4].

However, cosine similarity poses interpretability challenges: vectors with identical pairwise rankings can still be judged dissimilar, while near-aligned vectors may reflect contradictory preferences. Moreover, distinct behavioral measures (e.g., donations vs. letter writing) complicate comparisons of pairwise preferences across options within the same participant.

To address these limitations, we evaluate survey instruments based on their ability to recover (1) *pairwise preference rankings* and (2) *preference intensity differences* between options, within participants. This dual evaluation draws from methods used in point-allocation and forced-choice survey studies [9], and allows us to separately assess ordinal and interval-level performance.

We employ Bayesian modeling in both cases to support transparent assumptions, handle uncertainty, and enable interpretation beyond binary significance thresholds [32, 40]. The two models are described below.

4.1 Pairwise Ranking Model

Our first analysis evaluates how well different survey instruments capture pairwise rankings that align with those inferred from actual donation amounts. We model the binary observation (y_i) of whether participant i's pairwise ranking expressed via the survey instrument matches that from the donation results for a given societal issue pair as a Bernoulli distribution in Equation (1):

$$y_i \sim \text{Bernoulli}(\theta_i) \tag{1}$$

The alignment probability, θ_i, is defined via a logit link function (Eq. 2) and modeled as a function of several experimental variables:

$$\text{logit}(\theta_i) = \alpha + \beta_c[C_i] + \beta_o[O_i] + \beta_p[P_i] + \beta_t[T_{1i}] + \beta_t[T_{2i}] \tag{2}$$

The variables represent experimental conditions and relevant controls. Specifically, C_i denotes the survey instrument, spanning eight conditions (see Table 1[2]): three QS variants with different budgets, three LS variants with different budgets, a Likert scale survey, and a UQS condition. Since some participants worked on multiple survey instruments, O_i captures the order in which the participant completed the survey C_i to account for ordering effects. In addition, P_i represents whether a participant's pairwise ranking in an earlier survey aligned with the donation-based ranking, accounting for carryover effects. Lastly, T_{1i} and T_{2i} account for the topic-level effects of the two issues in comparison.

Given the complexity and nested structure of the data, we used a hierarchical Bayesian logistic regression model with non-centered parameterization [40]. Hierarchical modeling enables partial pooling across different experimental conditions or topic pairings, which improves estimate robustness [40].

We model the coefficients of each experimental variable (β_c, β_o, β_p, and β_t) using a hierarchical structure, drawing them from a normal distribution centered at a group-level mean μ_β and scaled by

[2]Since this model only considers pairwise rankings, UQS and QS votes and credits yield the same result.

a group-level standard deviation σ_β. For example, the hierarchical structure of the coefficient β_c for the survey condition variable C_i is defined as:

$$\beta_c[C_i] = \mu_{\beta_c} + \sigma_{\beta_c} \cdot \eta[C_i], \quad \eta[c_i] \sim \mathcal{N}(0,1) \tag{3}$$

$$\mu_{\beta_c} \sim \mathcal{N}(0, 0.5), \quad \sigma_c \sim \text{Uniform}(0,1) \tag{4}$$

Other coefficients follow the same structure, but some have different hyper-priors. Specifically, topic coefficients β_t use a narrower hyper-prior $\mu_{\beta_t} \sim \mathcal{N}(0, 0.25)$ to reflect a smaller expected effect.

4.2 Pairwise Preference Intensity Model

The pairwise intensity model evaluates how effectively each survey instrument captures the magnitude of preference differences between options. We seek to model how the response difference between two options on a survey Δ_{Survey} *predicts* the donation difference Δ_{Donation}. Besides the eight survey conditions in the pairwise ranking model, we additionally analyze the number of credits spent on an option in QS (three budgets) and UQS (Table 1).

Comparing preference differences elicited via various survey instruments (Δ_{Survey}) to donations (Δ_{Donation}) is not trivial since some yielded continuous data while others were ordinal. Following the convention, we model Δ_{Likert} as ordinal data. Given the uncertainty about how participants accounted for the varying costs associated with QS votes, we treat $\Delta_{\text{QS Vote}}$ as ordinal as well. In contrast, LS votes increment with a consistent cost on a scale and are therefore modeled as a continuous variable. UQS has no upper limit; hence, it is not ordinal. Finally, QS credits and monetary donations are continuous by nature.

Another challenge of this comparison is that raw differences from various instruments ($\Delta_{\text{Survey Raw}}$) and $\Delta_{\text{Donation Raw}}$ fall into varying data ranges. Thus, we apply the following data normalizations.

Normalize Continuous Survey Difference. We apply a variation of min-max scaling to project continuous $\Delta_{\text{Survey Raw}}$ onto the $[-1, 1]$ interval[3]. For QS credits and LS, we use the predefined bounds as the *min* and *max* in scaling. Since UQS lacks fixed bounds, we calculate the *min* and *max* for each participant based on the total votes and credits they used.

Normalize Ordinal Survey Difference. To enable direct comparison with continuous data, we project ordinal difference categories onto a latent continuous scale between 0 and 1, using cutpoints drawn from a Dirichlet-based model. Specifically, for each instrument, we derive K discrete ordinal *difference* categories. For example, vote difference in QS36 has 17 possible difference categories ($\Delta_{\text{QS36 Vote Raw}} = [-8, -7, ..., 7, 8]$)[4]. We sample the second to Kth elements of cutpoints $\boldsymbol{\alpha}$ from a Dirichlet$(\mathbf{1} \cdot \delta)$ with $\delta = 2$ as a weakly informative prior so that they sum to 1:

$$\boldsymbol{\alpha}_{[2,...,K]} \sim \text{Dirichlet}(\mathbf{1} \cdot \delta), \text{where } \delta = 2 \tag{5}$$

The first cutpoint is $\alpha_1 = 0$. We then map ordinal $\Delta_{\text{Survey Raw}} = k$ to a latent continuous value between $[0, 1]$ as Δ_{Survey}:

$$\text{For } \Delta_{\text{Survey Raw}} = k, \Delta_{\text{Survey}} = \sum_{j=1}^{k} \alpha_j. \tag{6}$$

Normalize Donation Difference. Finally, we apply the same variation of min-max scaling to the donation differences of each participant based on their total donation amount. Δ_{Donation} ranges from $[-1, 1]$.

Model Specification: We model Δ_{Donation} as a Normal distribution:

$$\Delta_{\text{Donation}_i} \sim \mathcal{N}(\mu_{D_i}, \sigma_{D_i}). \tag{7}$$

Since it is reasonable to expect that the variance in donation differs across experimental conditions, we make σ_{D_i} condition-dependent: $\sigma_i = \beta_\sigma[C_i]$, where $\beta_\sigma[C_i]$ is drawn from the prior Exponential(1). μ_{D_i} is predicted by a linear regression of survey response difference Δ_{Survey_i}, survey instrument C_i, survey order O_i, and topics T_{1i}, T_{2i}.

$$\mu_i = \beta_S[C_i] \cdot \Delta_{\text{Survey}_i} + \beta_c[C_i] + \beta_o[O_i] + \beta_t[T_{1i}] + \beta_t[T_{2i}]. \tag{8}$$

We model the slope of survey response differences β_S for each survey condition with partial pooling and non-centered parameterization.

$$\beta_S[C_i] = \mu_{\beta_{\text{vote}}} + \sigma_{\beta_{\text{vote}}} \cdot \eta_{\beta_{\text{vote}}}[C_i], \quad \eta[c_i] \sim \mathcal{N}(0,1) \tag{9}$$

$$\mu_{\beta_{\text{vote}}} \sim \mathcal{N}(0,1), \quad \sigma_{\beta_{\text{vote}}} \sim \text{Uniform}(0,1). \tag{10}$$

We model intercepts β_o and β_t in a similar way but with a hyper-prior of $\mu_\beta \sim \mathcal{N}(0, 0.1)$. Finally, we sample the condition-based intercept β_c from the prior $\mathcal{N}(0, 0.2)$ without pooling.

5 Results

We present findings on pairwise rankings and preference intensity in Section 5.1 and Section 5.2, respectively[5].

5.1 Pairwise Preference Ranking Results

Results interpretation: To evaluate how well a survey tool reflects a participant's preference ranking between two causes, we calculate the posterior distribution of the probability that the pairwise preference ranking reflected through a survey tool aligns with that reflected in donation amounts (Figure 3). Furthermore, we compare survey tools' abilities to elicit accurate pairwise preference rankings using the odds ratio of the predicted odds of alignment between survey and donation preference rankings. For instance, an odds ratio of 2 between survey tools A and B means that the odds of participants expressing the same preference rankings in survey tool A and donations is twice those of participants using survey tool B. We say that two survey tools differed significantly when the 94% Highest Posterior Density Interval (HPDI) of the odds ratio's posterior distribution does not include the reference value of 1 (odds ratio = 1 means having the same odds).

QS outperformed the Likert scale survey in eliciting preference rankings consistent with donations, with a small effect

[3] $\frac{x - min(x)}{(max(x) - min(x))} \times 2 - 1$

[4] Here, the largest vote difference possible with 36 credits occurs with 5 votes on option A and −3 votes on option B.

[5] The analysis notebook and experimental data are available at https://github.com/CrowdDynamicsLab/ci-2025-analysis.

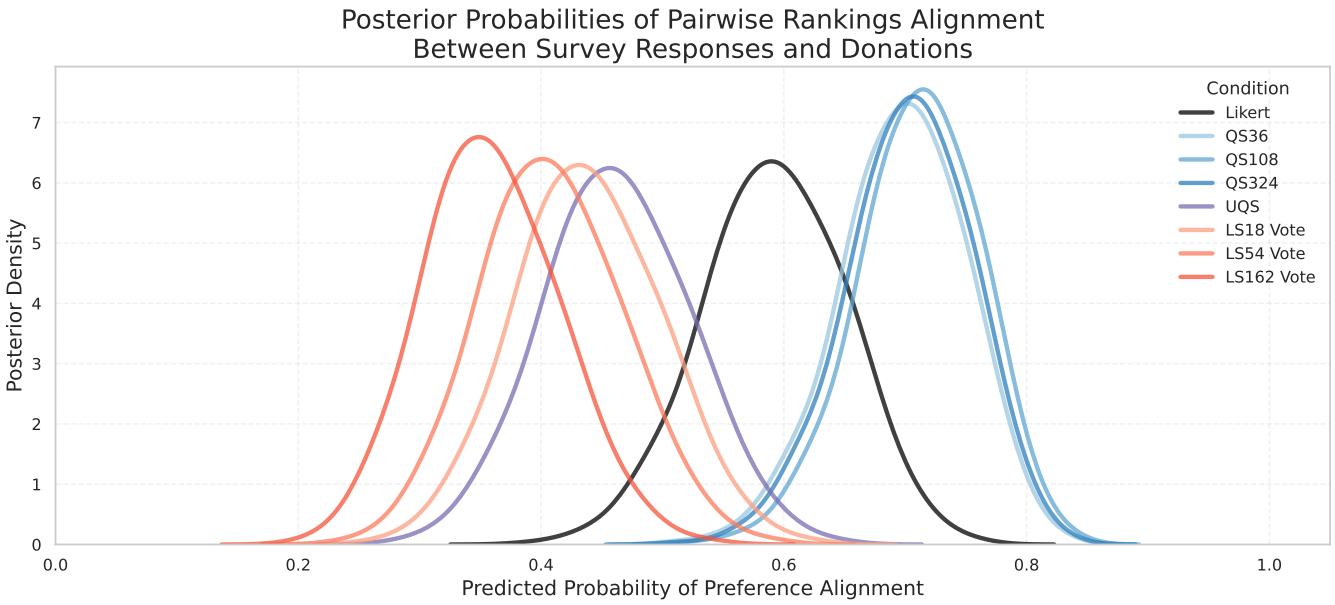


Figure 3: This figure presents the posterior density distributions of the probability that pairwise rankings from various survey tools align with participants' actual donation behavior. The x-axis shows the predicted probability of correct pairwise ranking alignment, where 1 indicates perfect alignment. The y-axis represents the posterior density across the sampled distribution. Each curve corresponds to a different survey condition. QS with varying budgets cluster together, exhibiting higher alignment probabilities as reflected by their distributions peaking further to the right. In contrast, UQS and LS show lower alignment, with LS performance declining as its budget increases. Main takeaway: Budget-constrained QS elicit pairwise rankings that align more closely with participants' donation behaviors, highlighting their effectiveness in capturing directional preferences.

size (odds ratio mean = 1.65, 94% HPDI = [1.55, 1.76])[6]. The model predicted that a participant's preference ranking in QS aligned with that in donations with a 70% chance on average, higher than the 59% average probability for the Likert scale survey.

When the budget from QS was removed, UQS performed worse than the Likert scale with a small effect size (odds ratio mean = 0.59, 94% HPDI = [0.56, 0.62]). Participants expressed consistent pairwise preference rankings with Unlimited QS and donations 46.2% of the time on average (94% HPDI = [35.0%, 57.1%]).

LS, a variation of QS with a linear instead of quadratic cost, was also less effective than the Likert scale with a small effect size (odds ratio mean = 0.46, 94% HPDI = [0.37, 0.55]). In addition, LS's performance worsened as its budget increased. The average predicted probability of consistent pairwise preference rankings between LS and donations was 43.9%, 40.8%, and 35.9% for LS with a small, medium, and large budgets.

5.2 Pairwise Preference Intensity Results

Results interpretation: With the fitted intensity model, we calculate the posterior predictive distribution of the mean of donation differences ($\mu_{\Delta_{\text{Predicted Donation}}}$), given a preference difference intensity between any two options elicited by a survey tool (Δ_{Survey}). Three such distributions are constructed for each survey tool, one each for small, medium, and large preference differences elicited by the tool respectively (i.e., Δ_{Survey} = 0.19, 0.38, 0.57, corresponding to $median(\Delta_{\text{Survey}}) + k \times std(\Delta_{\text{Survey}})$, where $k = 1, 2, 3$). We then perform two comparison tasks using these posterior distributions of $\mu_{\Delta_{\text{Predicted Donation}}}$.

First, we evaluate if predicted normalized donation differences $\Delta_{\text{Predicted Donation}}$ significantly differ from the "perfect" predicted donation difference ($\Delta_{\text{Donation Ref}}$). A predicted normalized donation difference between two options is "perfect" when it equals the normalized difference between the preferences elicited by the survey tool ($\Delta_{\text{Donation Ref}} = \Delta_{\text{Survey}}$). When $\Delta_{\text{Predicted Donation}} < \Delta_{\text{Donation Ref}}$, it means that our participants donated less to their preferred option than they said they would on the survey (relative to the less-preferred option), and vice-versa. We conclude that a survey tool failed to reflect a given preference difference intensity well when the 94% Highest Posterior Density Interval (HPDI) of the distribution of $\mu_{\Delta_{\text{Predicted Donation}}}$ does not include $\Delta_{\text{Donation Ref}}$.

Second, we compare the posterior distributions of $\mu_{\Delta_{\text{Predicted Donation}}}$ between survey conditions for the same Δ_{Survey}. Such comparisons provide insights into how a survey tool performs relative to another tool. We construct the posterior distribution of Cohen's d to quantify the difference between the $\mu_{\Delta_{\text{Predicted Donation}}}$ of a pair of survey conditions. We report that a survey tool's ability to reflect a preference intensity differs from another when the 94% HPDI of the Cohen's d distribution excludes zero.

[6]Odds ratio = 1.68, 3.47, 6.71 corresponds to a small, medium, and large effect size, respectively [6]

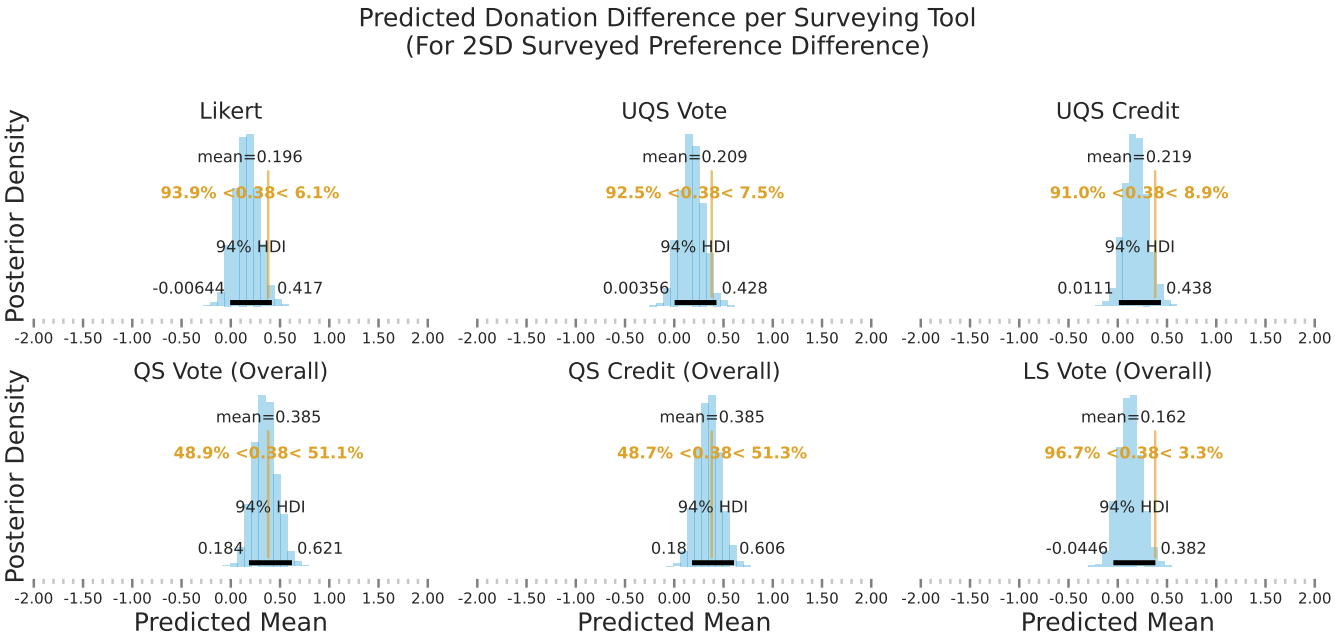


Figure 4: The posterior predictive distributions of donation differences for various surveying tools, assuming that the surveyed preference difference between two options is 2 standard deviations (SD) across all pairwise differences in our dataset. The x-axis displays predicted mean donation differences, while the y-axis represents posterior density, indicating the probability of various predicted mean donation differences occurring. The region indicated by the bolded black line represents the 94% Highest Density Interval (HDI), providing the most credible range for the predicted donation difference. A vertical gold line at 0.38 reflects the "perfect" donation value corresponding to a 2 SD-surveyed preference difference. In this plot, aside from QS (votes and credit), other surveying tools produced donation predictions that fell short of the 0.38 threshold, suggesting that these tools overly expressed preference intensity rather than accurately capturing it. Main takeaway: QS is capable of capturing the intensity of medium-sized (2 SD) preference differences between options accurately.

Small differences in survey responses (Δ_{Survey}) reliably predicted differences in donation behavior ($\Delta_{\text{Predicted Donation}}$) across all conditions. Among them, Likert, UQS vote, UQS credit, and LS results aligned best with donation differences ($\mu_{\Delta_{\text{Predicted Donation}}}$ = 0.15, 0.17, 0.18, 0.15, respectively; $\Delta_{\text{Donation Ref}}$ = 0.19). When participants expressed a small difference in QS vote and credit between two options, they expressed larger differences in donations (mean of $\mu_{\Delta_{\text{Predicted Donation}}}$ = 0.29, 0.26, respectively). But their donation differences did not differ significantly from the "perfect" difference ($\Delta_{\text{Donation Ref}}$ = 0.19).

As Δ_{Survey} increased to medium and large sizes, only those elicited by QS (both vote and credit, regardless of the budget size) were well-reflected in the corresponding donation differences. For instance, Figure 4 shows that when QS vote and credit Δ_{Survey} = 0.38 (medium difference), the mean of $\mu_{\Delta_{\text{Predicted Donation}}}$ = 0.39 (94% HPDI for QS vote = [0.18, 0.62], for QS credit = [0.18, 0.61]). **Results from QS aligned significantly better with donation results than those from the Likert scale with a medium to large effect size**[7]. Moreover, QS's advantage over the Likert scale increased with Δ_{Survey}, as shown in Figure 5. Using the donation prediction accuracy of QS credit vs. Likert scale as an example, the mean Cohen's d increases from 0.71 to 0.99 when Δ_{Survey} changes from medium (Cohen's d 94% HPDI = [0.62, 0.81]) to large (Cohen's d 94% HPDI = [0.89, 1.09]).

On the other hand, for medium and large Δ_{Survey}, UQS (i.e., QS without a budget) predicted donation difference similarly to the Likert scale, hence significantly worse than QS. Furthermore, LS with various budget sizes (i.e., QS without the quadratic cost) performed worse than Likert and UQS with a small effect size. When participants conveyed a medium-sized or larger preference difference between two options in Likert, UQS, or LS, they expressed a weaker difference in donations. A large Δ_{Survey} in Likert, UQS, and LS, for instance, predicted a mean $\mu_{\Delta_{\text{Predicted Donation}}}$ of 0.25, 0.25, and 0.17 respectively, far lower than the ideal difference ($\Delta_{\text{Donation Ref}}$ = 0.57).

6 Discussion

This section addresses the research questions, interprets findings, and offers practical recommendations.

6.1 QS's effectiveness and its dual components

We evaluated the effectiveness of QS in capturing pairwise preferences (RQ1) and whether both the quadratic cost function and budget constraint are necessary (RQ2). Results indicate that QS,

[7]Cohen's d = 0.2, 0.5, 0.8 corresponds to a small, medium, and large effect size, respectively

whether analyzed through votes or costs, consistently outperforms Likert scale surveys in recovering ordinal rankings and preference intervals—especially when resources are constrained. Moreover, QS shows increasing advantages as preference gaps widen, capturing intensities with greater accuracy over other methods.

Results suggest that both the quadratic cost function and the fixed budget constraint are essential to QS's effectiveness. Performance drops significantly when either is removed, as observed in LS and UQS (see Section 5.2). This gap between LS and QS requires further investigation, as discussed in the following subsections.

6.2 Mechanisms underlying QS's effectiveness

This subsection explores the plausible mechanisms underlying QS's effectiveness. While our results replicate and extend the observed advantage of QS over Likert scale surveys [7], we focus this subsection on two mechanisms: (1) how the quadratic cost function aligns preferences with behavior, and (2) how budget constraints shape expression. We examine these mechanisms using LS and UQS results.

6.2.1 Quadratic cost function corrects perception distortion and bias in response strengths. As discussed in Section 4.2, QS credits and votes remain aligned with participants' revealed behaviors across varying levels of preference strength. In contrast, our model shows increased exaggeration in LS and Likert scale survey results as the pairwise donation difference widens (see Figure 5). We identify two plausible explanations.

Unequal perceived 'preference units'. We define a 'preference unit' as the incremental amount a participant uses to express additional preference (e.g., an extra vote, extra credits spent, or an extra level on a Likert scale). Our results indicate that LS participants perceive successive preference units as representing smaller incremental differences in strength. This pattern aligns with the Law of Diminishing Marginal Utility in economics [22, 30], which states that each additional unit of consumption yields less utility than the one before. Thus, survey respondents allocate more preference units to express their intended preferences.

Even if participants do not explicitly interpret preference units in monetary terms, psychophysics offers similar concepts. According to the Weber-Fechner law [13, 34], the just noticeable difference between stimuli is proportional to the baseline stimulus intensity. As the marginal difference between options appears to shrink, participants may feel their previous input was insufficient and overcorrect as a result. Early CSS validation by Dudek and Baker [14] cautioned that participants may misjudge how well their numerical input reflects their subjective attitudes. Interestingly, Fechner's law [34] describes perception as following a logarithmic curve, which conceptually aligns with the quadratic cost structure in QS. In QS, the rising cost of additional votes may have corrected participants' diminishing perceptual increments, helping to mitigate exaggerations in expressed preferences stemming from participants' perceptual biases.

Extreme response bias. Another plausible explanation involves the large decision space and the ease of expressing extreme opinions offered by LS. With the same budget size, participants face a wider array of allocation choices when votes incur a linear cost than a quadratic cost (e.g., 324 choices in LS162 vs. 25 choices in QS162 on an option). The psychology literature suggests that cognitive overload leads to satisficing [28, 51], where individuals settle for sufficient rather than optimal choices by relying on heuristics. As the allocated budget increased, participants increasingly underutilized their budgets, a pattern consistent with satisficing behavior (Figure 6). With a satisficing mindset, rather than carefully weighing and quantifying the differences between options, participants may resort to exaggerating responses to signal distinctions between choices. While this exaggeration strategy is possible with both LS and QS, it was discouraged by the quadratic cost structure in QS. The quadratic cost function imposes increasing costs for each additional vote, which makes people think twice before expressing strong opinions. In LS, on the contrary, it does not cost much to exaggerate preferences since each vote carries equal weight. In summary, the quadratic cost structure in QS may have reduced the occurrences of exaggerated responses by alleviating cognitive load and imposing a higher cost to expressing extreme opinions.

6.2.2 The quadratic cost function reduces the cognitive burden influencing pairwise rankings. This expanded decision space may have also undermined LS's ability to preserve pairwise rankings that are consistent with participants' opinions. The linear cost in LS increases the number of possible allocation outcomes compared to QS with the same budget size, raising the cognitive demand required to maintain consistent relative preferences. In contrast, QS's quadratic cost structure progressively narrows this decision space, easing the cognitive burden and facilitating more consistent preference rankings. This cognitive burden likely contributes to noisier and more inconsistent pairwise rankings.

6.2.3 The role of budget: Anchor and a sense of scarcity. UQS performs similarly to LS, likely due to two key factors. First, without a fixed budget, participants face unlimited allocation options, eliminating clear opportunity costs and tradeoffs. This can result in more extreme allocations or exaggerated expressions (e.g., one participant allocated 105 votes to a single option). Second, without a budget constraint, participants lack a stable reference for a 'preference unit,' since the meaning or weight of each additional vote shifted with each allocation. This undermines anchoring [29, 58], often necessary for initiating preference construction [35], and further expands the decision space. We observe that less than half of the participants spend more than 38 credits, followed by a long tail of credit usage (Figure 7), possibly due to cognitive overload or because they felt further input was unnecessary. This behavior constrained the expansive decision space, suggesting that budget limits may help regulate expressive intensity.

In summary, QS's quadratic cost function helps mitigate distortions caused by perceptual biases and choice overload, particularly when participants express large differences between options. The fixed budget constraint anchors participants' interpretations of preference units and limits the decision space, thereby reducing cognitive load and supporting more accurate expression.

Together, these mechanisms likely explain why QS aligns more closely with participants' behavioral outcomes than the alternatives. While each mechanism is grounded in prior literature and supported by our findings, their interaction remains unclear. Perceptual distortion, overload, and anchoring may be interdependent

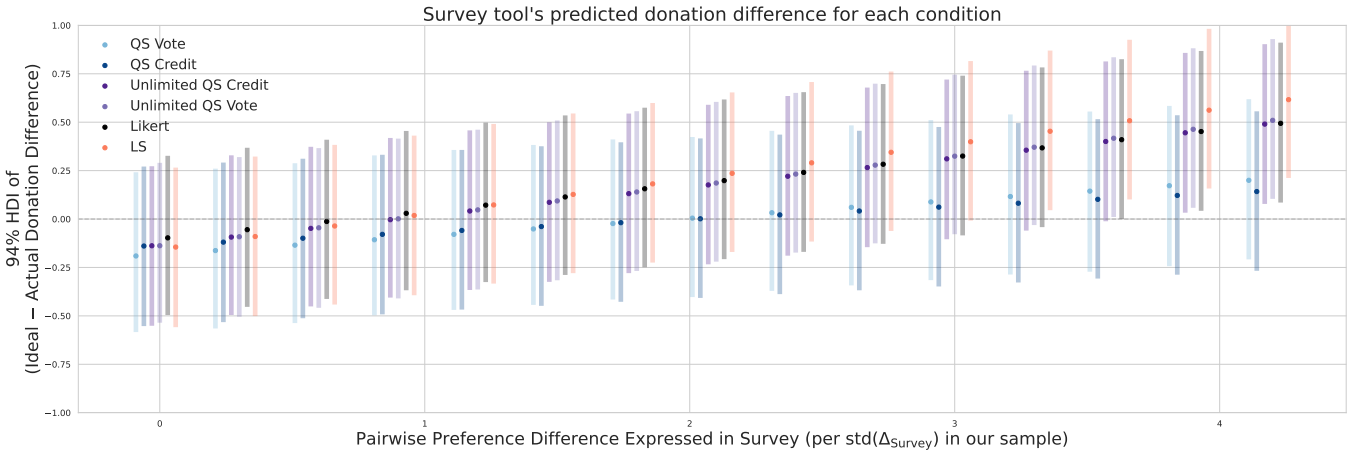


Figure 5: Differences between survey-reported preferences and actual donation behaviors across survey tools. Dots represent mean normalized differences, with bars showing the 94% Highest Density Interval (HDI). The horizontal line at 0 indicates perfect alignment; values above represent overstated preferences. QS Vote and QS Credit consistently show close alignment even as actual behavioral differences increase, while Likert, LS, and Unlimited QS increasingly deviate at larger differences. Takeaway: QS methods (Vote and Credit) better capture participants' actual preference intervals, especially as intensity differences grow.

rather than isolated effects. Future research should investigate how participants' internally valued preferences map onto their expressed responses in QS. Cognitive interviews aimed at constructing mental models could reveal how participants interpret budget constraints, cost structures, and manage preference tradeoffs during preference construction.

6.3 Takeaways for QS practitioners

Our findings offer practical implications for practitioners using QS to elicit preferences in collective intelligence settings:

Balancing simplicity and accuracy for basic ranking tasks. Likert scale surveys may suffice for simple rankings that do not require aggregation, given their simplicity. While QS still yields slightly better alignment with behavioral outcomes, the added complexity may not be justified in low-stakes or ordinal-only contexts.

QS excels in capturing preference intensity and enabling aggregation. QS is particularly effective for capturing nuanced differences in preference strength across competing options and aggregating those preferences across individuals. Our results demonstrate that QS significantly outperforms traditional methods in capturing preference intensities and enabling aggregation.

Do not compromise the QS mechanism. Both the quadratic cost function and the budget constraint are essential for QS to function effectively. Removing either component (as in LS or UQS) substantially reduces performance. Since these features introduce cognitive load, it is essential to design usable interfaces that support participants' preference construction process.

Credit budget matters less, but a medium range is recommended. Although we observed no statistically significant differences among QS36, QS108, and QS324, Bayesian analysis consistently favored QS108 for stable, balanced outcomes. Thus, practitioners are advised to scale the credit budget to the power of 1.5 relative to the number of options ($O(k^{1.5})$ as defined by prior research [7].

7 Limitations and Future work

Limitations of donation-based preference elicitation. Behavioral donation tasks are widely used in prior research [2, 7, 18, 61] to approximate "true preferences" through real-stakes decisions for survey validation. While the donation task in this study was adapted from prior designs to ensure incentive compatibility and behavioral realism, not all preferences are naturally expressed through monetary contributions. Participants' mental models, shaped by donation amounts, prior giving, or personal motivations, may differ from those guiding their responses to survey instruments. Moreover, although charitable donations offer incentive-aligned behavior, the domain involves distinct social and motivational factors that may not generalize to other allocation contexts. Governmental resource allocation, for instance, includes political accountability, public transparency, and long-term planning considerations. Future field studies applying QS in public decision-making settings could further assess its generalizability.

Future work: comparing QS with other forced-choice methods. The limited performance of LS highlights the need for future comparative studies between QS and other forced-choice preference elicitation tools, such as KS and conjoint analysis. These methods rely on different constraints and mechanisms to reveal preference intensity. Comparing them across decision contexts and preference distributions would help practitioners identify the most appropriate tools. Future evaluations should consider elicitation accuracy, cognitive load, and user experience.

8 Conclusion

This work deepens our understanding of QS used to elicit richer signals in collective intelligence systems. By experimentally isolating

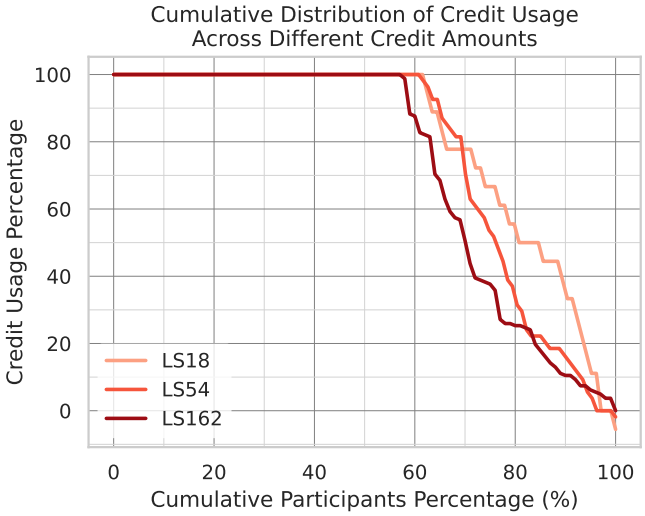


Figure 6: Percentage of participants who fully utilized their credits across budget levels. As budgets increase, cumulative usage drops off more steeply. Takeaway: Higher budgets lead to earlier and more widespread underutilization of credits.

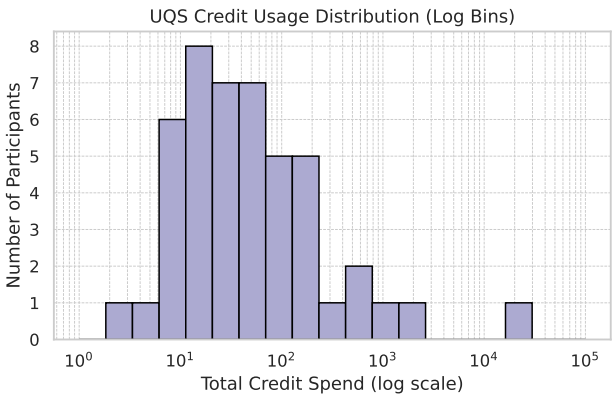


Figure 7: The credit usage distribution for UQS participants. The credit usage follows a long tail. Without an anchor, less than half of participants spend more than 38 credits across 9 options. Takeaway: Participants did not put in enough effort to use as many credits as needed to express nuanced differences between option preferences.

the effects of the quadratic cost function and budget constraint, we compared QS with LS, UQS, and Likert scale surveys. Results show that removing either component weakens QS's ability to capture accurate rankings and intensities, particularly as preference gaps grow.

These findings position QS as a promising instrument for collective decision-making in a resource-constrained context where it is critical to capture not just what people prefer, but how strongly they prefer it. Future CI research should explore how people mentally model credit constraints, costs, and trade-offs when responding with QS, thereby guiding the design of more intuitive and effective tools for collective choice.

Acknowledgments

We thank the study participants for their time and contributions. We are also grateful to Dr. Michel Regenwetter, Vinay Koshy, James Eschrich, Andrew Chen, and Aditya Karan for their valuable feedback. This work would not have been possible without the support of peers from the Social Spaces and Crowd Dynamics Lab. Additional thanks to Hsin-Ni Yu and Chi-Chun Huang. This research was partially supported by the Center for Just Infrastructures.

References

[1] Richard Bagozzi. 1994. *Advanced Marketing Research.* John Wiley & Sons.
[2] Matthias Benz and Stephan Meier. 2008. Do People Behave in Experiments as in the Field?—Evidence from Donations. *Experimental economics* 11, 3 (2008), 268–281.
[3] Kylie Brosnan, Bettina Grün, and Sara Dolnicar. 2021. Cognitive Load Reduction Strategies in Questionnaire Design. *International Journal of Market Research* 63, 2 (March 2021), 125–133. doi:10.1177/1470785320986797
[4] Charlotte Cavaillé, Daniel L. Chen, and Karine Van der Straeten. 2024. Who Cares? Measuring Differences in Preference Intensity. *Political Science Research and Methods* (2024), 1–17. doi:10.1017/psrm.2024.27
[5] Damon Centola. 2022. The Network Science of Collective Intelligence. *Trends in Cognitive Sciences* 26, 11 (2022), 923–941.
[6] Henian Chen, Patricia Cohen, and Sophie Chen. 2010. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—simulation and Computation®* 39, 4 (2010), 860–864.
[7] 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 (April 2021), 1–43. doi:10.1145/3449281
[8] Ti-Chung Cheng, Yutong Zhang, Yi-Hung Chou, Vinay Koshy, Tiffany Wenting Li, Karrie Karahalios, and Hari Sundaram. 2025. Organize, Then Vote: Exploring Cognitive Load in Quadratic Survey Interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 475. doi:10.1145/3706598.3714193
[9] Marion Collewet and Paul Koster. 2023. Preference Estimation from Point Allocation Experiments. *Journal of choice modelling* 48 (2023), 100430.
[10] Donald R. Cooper and Pamela S. Schindler. 2013. *Business Research Methods, 12th Edition* (hardcover ed.). McGraw-Hill Education. 692 pages.
[11] Bo Cowgill and Eric Zitzewitz. 2015. Corporate Prediction Markets: Evidence from Google, Ford, and Firm x. *The Review of Economic Studies* 82, 4 (2015), 1309–1341.
[12] BOAVENTURA de SOUSA SANTOS. 1998. Participatory Budgeting in Porto Alegre: Toward a Redistributive Democracy. *Politics & Society* 26, 4 (Dec. 1998), 461–510. doi:10.1177/0032329298026004003
[13] Stanislas Dehaene. 2003. The Neural Basis of the Weber–Fechner Law: A Logarithmic Mental Number Line. *Trends in Cognitive Sciences* 7, 4 (April 2003), 145–147. doi:10.1016/S1364-6613(03)00055-X
[14] Frank J. Dudek and Katherine E. Baker. 1957. On the Validity of the Point-Assignment Procedure in the Constant-Sum Method. *The American Journal of Psychology* 70, 2 (1957), 268–271. doi:10.2307/1419333 jstor:1419333
[15] Jon X Eguia, Nicole Immorlica, Katrina Ligett, E Glen Weyl, and Dimitrios Xefteris. 2019. Quadratic Voting with Multiple Alternatives. *Available at SSRN 3319508* (2019).
[16] James S. Fishkin. 2003. Consulting the Public through Deliberative Polling. *Journal of Policy Analysis and Management* 22, 1 (2003), 128–133. jstor:3325851
[17] Mirta Galesic. 2006. Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey. *Journal of Official Statistics* 22, 2 (June 2006), 313.
[18] Philip Gendall and Benjamin Healey. 2010. Effect of a Promised Donation to Charity on Survey Response. *International Journal of Market Research* 52, 5 (2010), 565–577.
[19] Laura Georgescu, James Fox, Anna Gautier, and Michael Wooldridge. 2024. Fixed-Budget and Multiple-issue Quadratic Voting. arXiv:2409.06614
[20] Ashish Goel, Anilesh K Krishnaswamy, and Sukolsak Sakshuwong. 2016. Budget Aggregation via Knapsack Voting: Welfare-maximization and Strategy-Proofness. *Collective Intelligence* (2016), 783–809.
[21] Ashish Goel, Anilesh K. Krishnaswamy, Sukolsak Sakshuwong, and Tanja Aitamurto. 2019. Knapsack Voting for Participatory Budgeting. *ACM Transactions on*

Economics and Computation 7, 2, Article 8 (July 2019). doi:10.1145/3340230
[22] Hermann Heinrich Gossen. 1983. *The Laws of Human Relations and the Rules of Human Action Derived Therefrom.* The MIT Press, Cambridge, MA.
[23] Theodore Groves and John Ledyard. 1977. Optimal Allocation of Public Goods: A Solution to the "Free Rider" Problem. *Econometrica* 45, 4 (1977), 783–809. doi:10.2307/1912672 jstor:1912672
[24] Will S. Harwood and MaryAnne Drake. 2019. Understanding Implicit and Explicit Consumer Desires for Protein Bars, Powders, and Beverages. *Journal of Sensory Studies* 34, 3 (2019), e12493. doi:10.1111/joss.12493
[25] John R. Hauser and Steven M. Shugan. 1980. Intensity Measures of Consumer Preference. *Operations Research* 28, 2 (1980), 278–320. jstor:170448
[26] Alishia C Holland. 2022. Distributive Impacts and Support for Mass Transportation Projects: An Experimental Evaluation in Bogotá, Colombia. (2022).
[27] Yu-Tang Hsiao, Shu-Yang Lin, Audrey Tang, Darshana Narayanan, and Claudina Sarahe. 2018. vTaiwan: An Empirical Study of Open Consultation Process in Taiwan. *Taiwan: Center for Open Science* (2018).
[28] 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 (2000), 995.
[29] Daniel Kahneman. 2017. *Thinking, Fast and Slow.* Farrar, Straus and Giroux.
[30] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–291. doi:10.2307/1914185 jstor:1914185
[31] Tara Karpinski, Michel van Dartel, and Martijn de Waal. 2025. The Potential of Quadratic Voting for Cohousing Communities: A Situated Design Case Study. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (Chi Ea '25).* Association for Computing Machinery, New York, NY, USA, Article 693. doi:10.1145/3706599.3706671
[32] 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.
[33] Jon A Krosnick. 1999. Survey Research. *Annual review of psychology* 50, 1 (1999), 537–567.
[34] Lester E. Krueger. 1989. Reconciling Fechner and Stevens: Toward a Unified Psychophysical Law. *Behavioral and Brain Sciences* 12, 2 (June 1989), 251–267. doi:10.1017/S0140525X0004855X
[35] Sarah Lichtenstein and Paul Slovic (Eds.). 2006. *The Construction of Preference* (1. publ ed.). Cambridge University Press, Cambridge.
[36] lorraine. 2022. Constant Sum Question: A New Matrix Question Type to Customize Your Survey.
[37] Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications.* Cambridge University Press.
[38] Jordan J. Louviere and Towhidul Islam. 2008. A Comparison of Importance Weights and Willingness-to-Pay Measures Derived from Choice-Based Conjoint, Constant Sum Scales and Best–Worst Scaling. *Journal of Business Research* 61, 9 (Sept. 2008), 903–911. doi:10.1016/j.jbusres.2006.11.010
[39] Naresh K. Malhotra, David F. Birks, and Peter Wills. 2012. *Marketing Research: An Applied Approach* (paperback ed.). Pearson Education. 1037 pages.
[40] Richard McElreath. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* Chapman and Hall/CRC.
[41] John P. McIver and Elinor Ostrom. 1976. Using Budget Pies to Reveal Preferences: Validation of a Survey Instrument. (June 1976). doi:10.1332/030557376783015695
[42] Milton Metfessel. 1947. A Proposal for Quantitative Reporting of Comparative Judgments. *The Journal of Psychology* 24, 2 (Oct. 1947), 229–235. doi:10.1080/00223980.1947.9917350
[43] Ryan Naylor et al. 2017. First Year Student Conceptions of Success: What Really Matters? *Student Success* 8, 2 (2017), 9–19.
[44] Eric A Posner and E Glen Weyl. 2017. Quadratic Voting and the Public Good: Introduction. *Public Choice* 172, 1-2 (2017), 1–22.
[45] Eric A Posner and E Glen Weyl. 2018. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society.* Princeton University Press.
[46] Qualtrics. 2025. Constant Sum Question. https://www.qualtrics.com/support/survey-platform/survey-module/editing-questions/question-types-guide/specialty-questions/constant-sum/.
[47] David Quarfoot, Douglas von Kohorn, Kevin Slavin, Rory Sutherland, David Goldstein, and Ellen Konar. 2017. Quadratic Voting in the Wild: Real People, Real Votes. *Public Choice* 172, 1-2 (2017), 283–303.
[48] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. 2019. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *2019 IEEE Symposium on Security and Privacy (SP).* 1326–1343. doi:10.1109/SP.2019.00014
[49] Ron Roberts and Paul Goodwin. 2002. Weight Approximations in Multi-Attribute Decision Models. *Journal of Multi-Criteria Decision Analysis* 11, 6 (2002), 291–303. doi:10.1002/mcda.320
[50] Joshua Schramm and Marcel Lichters. 2024. Incentive Alignment in Anchored MaxDiff Yields Superior Predictive Validity. *Marketing Letters* (Jan. 2024). doi:10.1007/s11002-023-09714-2
[51] Barry Schwartz, Andrew Ward, John Monterosso, Sonja Lyubomirsky, Katherine White, and Darrin R. Lehman. 2002. Maximizing versus Satisficing: Happiness Is a Matter of Choice. *Journal of Personality and Social Psychology* 83, 5 (2002), 1178–1197. doi:10.1037/0022-3514.83.5.1178
[52] Anuj K. Shah, Eldar Shafir, and Sendhil Mullainathan. 2015. Scarcity Frames Value. *Psychological Science* 26, 4 (April 2015), 402–412. doi:10.1177/0956797614563958
[53] Scott M. Smith and Gerald S. Albaum. 2013. *Basic Marketing Research: Official Training Guide from Qualtrics.* Qualtrics Labs, Incorporated.
[54] SurveySparrow. 2025. What Is Constant Sum | Understanding Constant Sum. https://surveysparrow.com/what-is-constant-sum/.
[55] Susan J Thomas. 2004. *Using Web and Paper Questionnaires for Data-Based Decision Making: From Design to Interpretation of the Results.* Corwin Press.
[56] 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 (April 2019), 57–74. doi:10.1177/0759106319834665
[57] Stelios Tsafarakis, Panagiotis Gkorezis, Dimitrios Nalmpantis, Evangelos Genitsaris, Andreas Andronikidis, and Efthymios Altsitsiadis. 2019. Investigating the Preferences of Individuals on Public Transport Innovations Using the Maximum Difference Scaling Method European Transport Research Review. *European Transport Research Review* 11 (Jan. 2019). doi:10.1186/s12544-018-0340-6
[58] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. jstor:1738360
[59] E. Glen Weyl, Audrey Tang, and Community. 2024. *Plurality: The Future of Collaborative Technology and Democracy.* Independently published.
[60] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *science* 330, 6004 (2010), 686–688.
[61] Ziang Xiao, Po-Shiun Ho, Xinran Wang, Karrie Karahalios, and Hari Sundaram. 2019. Should We Use an Abstract Comic Form to Persuade? Experiments with Online Charitable Donation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.
[62] Shu-Hong Zhu and Norman H Anderson. 1991. Self-Estimation of Weight Parameter in Multiattribute Analysis. *Organizational Behavior and Human Decision Processes* 48, 1 (Feb. 1991), 36–54. doi:10.1016/0749-5978(91)90004-D