Efficiency and Usability of Participatory Budgeting Methods

Gerdus Benadè
Tepper School of Business
Carnegie Mellon University

Nevo Itzhak
Dept. of Information Systems Engineering
Ben-Gurion University

Nisarg Shah
Dept. of Computer Science
University of Toronto

Ariel D. Procaccia
Computer Science Dept.
Carnegie Mellon University

Ya'akov (Kobi) Gal
Dept. of Information Systems Engineering
Ben-Gurion University

Abstract

Participatory budgeting is an influential paradigm that engages residents in the process of allocating a city's budget. Different implementations vary in the input format they use to elicit participants' preferences on potential projects. Our goal is to compare input formats on two dimensions: efficiency, which is measured in terms of the social welfare of the resulting outcomes, and usability, which is evaluated through a combination of objective and subjective indicators. To this end, we conduct an extensive empirical study, in which more than 1200 voters used different methods in a controlled setting. Our results suggest that a popular input format known as k-approval imposes low cognitive burden and is strikingly efficient, but is not necessarily perceived as such.

1 Introduction

Participatory budgeting (Cabannes 2004) is an emerging democratic paradigm that allows members of a community (typically residents of a city) to play a role in the process of allocating a public budget. Since its invention in the Brazilian city of Porto Alegre in 1988, it has spread dramatically, and today is used by thousands of municipalities around the world (Röcke 2014). In fact, one nonprofit organization, the *Participatory Budgeting Project*,[1] has helped allocate more than $239M to more than 1630 projects across North America, as of August 24, 2018. And some of the world's major cities are allocating large sums of money by holding participatory budgeting elections: €100M in Paris in 2016, €24M in Madrid in 2016, $47M in Mexico City in 2016, and $40M in New York City in 2017, just to name a few.

From the viewpoint of computational social choice, there are two key decisions in the design of participatory budgeting systems: first, the choice of *input format* — the way in which voters express their preferences over potential projects (hereinafter, alternatives) through votes; second, the choice of *aggregation method* — the way in which votes are aggregated into a feasible allocation of the budget among alternatives. Much of the classical computational social choice literature fixes an input format — typically, a ranking over the alternatives — and studies different aggregation methods (Brandt et al. 2016, Chapter 2). By contrast, real-world participatory budgeting elections invariably use some sort of greedy aggregation method but differ significantly in the input formats they adopt.

The most popular input format, by far, is k-*approval*, whereby each voter designates her k favorite alternatives. For example, elections administered through the Stanford Participatory Budgeting Platform[2] used 4-approval in Boston, MA, in 2015, and 5-approval in Greensboro, NC, in 2016. A fundamentally different input format is known as *knapsack vote* (Goel et al. 2016) or *shopping cart vote*: each voter selects her favorite *set* of alternatives under a budget constraint. This input format has been used in participatory budgeting elections in Reykjavík since 2012,[3] as well as in the aforementioned 2016 election in Madrid.[4] Other input formats that have been discussed — and, in some cases, deployed — include *ranking by value* (ranking by perceived benefit), *ranking by value for money* (rank, or compare alternatives by their perceived benefit per unit of cost, i.e., 'bang for the buck'), and *threshold approval* (select all alternatives whose perceived benefit is above some given threshold).

In order to choose among these different input formats, one must first specify what one is trying to improve through participatory budgeting. Goel et al. (2016), who have administered participatory budgeting elections in North America (including some of those mentioned earlier), formulate the objective as economic *efficiency*: maximize *(utilitarian) social welfare* — the sum of utilities, with respect to voters' utility functions over alternatives — subject to a budget constraint, which takes into account a given cost for each alternative. The catch is that the participatory budgeting system does not have direct access to voters' utilities. Benadè et al. (2017) posit, following a recent line of work (Boutilier et al. 2015), that votes cast in a given input format serve as a proxy for voters' (unknown) utility functions. In other words, different input formats can be seen as lossy representations for voters' underlying utilities. Therefore, they can be compared in terms of the degree to which they allow an aggregation method — even the best one — to achieve efficiency, i.e., maximize the sum of the underlying utilities.

But if efficiency was the only criterion for selecting input formats, we would simply elicit the voters' full utility

[1] https://www.participatorybudgeting.org

[2] https://pbstanford.org

[3] https://www.citizens.is

[4] https://decide.madrid.es

functions. The obstacle is that reporting numerical utilities for alternatives is known to be difficult for people (Camerer 2011). Indeed, the second dimension on which input formats should be compared is *usability*, that is, how easy is it to learn about and use the input format, especially in terms of the cognitive burden imposed on voters. For example, to cast a knapsack vote, a voter must — in theory — solve an instance of the eponymous, NP-hard knapsack problem, which presumably makes this input format cognitively demanding. Or does it?

In this paper we empirically compare different input formats by measuring various indicators of their efficiency and usability. Our goal is to identify specific input formats that stand out in both dimensions.

1.1 Our Results

We conduct experiments based on data collected from more than 1200 voters on Amazon Mechanical Turk. Voters were asked to vote over items to take to a desert island. In the first of our two studies, voters were asked to cast a vote in a single input format, or report utilities. In the second, they were asked to both cast a vote and report utilities, as well as to answer several questions about their subjective experience.

Recall that we adopt the viewpoint that the goal of participatory budgeting is to find efficient outcomes (in terms of social welfare). To evaluate whether the different input formats lead to efficient outcomes, we aggregate a sample of votes and evaluate the outcome on a random sample of submitted utility profiles. The aggregation happens by finding the *distortion-minimizing* (and budget-feasible) subset of alternatives. This aggregation method explicitly considers the inherent uncertainty that exists when voters' utility functions can only be accessed via proxies, and returns the outcome that provides the best approximation of the optimal social welfare over all utility profiles that could have induced the observed votes. A key insight behind the experimental design is that we can measure the social welfare of an outcome selected by one group of voters using the utilities submitted by a different group, because the *average* utility of each item would be consistent across the groups by the law of large numbers (whose effect is present at the scale of our experiments).

We find that for most input formats distortion-minimizing aggregation leads to outcomes that are quite close to the welfare-maximizing outcome, even without access to the underlying utility profile. Moreover, we can see significant differences between different input formats, and some really shine. Most impressively, our results indicate that the k-approval and ranking by value for money input formats lead to outcomes that are essentially optimal.

Turning to usability, we consider two types of indicators. Objective indicators, which are computed from data, include consistency and response time. Subjective indicators, which are based on ratings reported by voters, include ease of use, likability, and expressiveness.

Consistency refers to the relation between a voter's utility function and her vote. For example, if we use k-approval as the input format, we expect a voter to approve the k alternatives for which she has the highest utility. If other alternatives are approved, it means that the voter may have misunderstood the instructions, or the cognitive burden imposed by the task was too high to perform it accurately. We find that k-approval by far leads to the highest degree of consistency, followed by threshold approval and knapsack. For response time, we find that k-approval again excels in terms of both time to learn and time to vote. By contrast, ranking by value for money does badly in both objective measures.

Finally, the subjective usability indicators generally favor ranking by value for money, especially in terms of how expressive it is perceived to be by voters. By contrast, k-approval is seen as the least expressive input format.

Nevertheless, overall we view k-approval as the clearest 'winner' based on our results. We discuss our conclusions in detail in Section 6.

1.2 Related Work

Our work is most closely related to two of the papers mentioned earlier. Building on their impressive practical work, Goel et al. (2016) establish theoretical results about knapsack voting, and provide an empirical analysis of data from participatory budgeting elections; here we focus on the latter set of results. One experiment provides timing data, indicating that "the knapsack interface is not much more time consuming than the k-approval interface." However, the authors note that "the knapsack interface follows the official k-approval interface, and so the voters were familiar with the projects when they attempted the knapsack vote." In our experiments, we observe that asking voters to cast votes in multiple input formats, one after another, affects not only time, but also other metrics, such as the accuracy with which they cast their votes (see Section 6). This led to the critical design choice of having each voter cast a single vote, using a single input format, in most of our experiments.

Goel et al. also observe, in two additional sets of experiments, that knapsack leads to funding lower-cost alternatives than k-approval, and that knapsack leads to a higher degree of agreement with pairwise comparisons than k-approval. These experiments arguably measure indirect indicators of societal benefit, but the authors note that they "stop short of claiming that knapsack voting leads to outcomes that are more beneficial to society as a whole." By contrast, we measure efficiency — the social welfare objective formulated by Goel et al. — directly by using the reported utilities of the voters themselves.

The second paper is the one by Benadè et al. (2017), who advocate the comparison of different input formats according to the degree to which they allow maximizing social welfare. They introduce the threshold approval input format, and prove that it has a significant advantage over other input formats *in the worst case*, by leveraging the *implicit utilitarian voting* paradigm (Procaccia and Rosenschein 2006; Boutilier et al. 2015; Anshelevich, Bhardwaj, and Postl 2015; Anshelevich and Postl 2016; Anshelevich and Sekar 2016; Caragiannis et al. 2017). They also conduct simulations (based on real data), which suggest that, in the average case, threshold approval has a small, but statistically significant, advantage over other input formats. Our extensive efficiency experiments, conducted with human subjects, are de-

signed to be much closer to practice; in these experiments, the efficiency advantage of threshold approval does not bear out.

2 Preliminaries

In the basic participatory budgeting model, a set of *voters* $N = \{1, \dots, n\}$ express their preferences over a set of m *alternatives* A. Each alternative a has a cost $c(a)$, and the total cost of the selected alternatives may not exceed a budget B. For $S \subseteq A$, let $c(S) = \sum_{a \in S} c(a)$.

Voter i has an additive utility function $v_i : A \to \mathbb{R}_+ \cup \{0\}$. We make a standard normalization assumption that voter utilities sum to the same amount, i.e., $v_i(A) = v_j(A)$ for all voters $i, j \in N$ (Benadè et al. 2017). If we were to elicit voter utilities directly, this would correspond to asking all voters to divide a fixed number of points between the alternatives in proportion to how much they like the alternatives; this, in a sense, enforces the *one person, one vote* principle. The vector $\vec{v} = (v_1, \dots, v_n)$ is called a *utility profile*.

Given a utility profile $\vec{v}$, the *social welfare* of an alternative a is $\mathrm{sw}(a, \vec{v}) = \sum_{i \in N} v_i(a)$. For $S \subseteq A$, let $\mathrm{sw}(S, \vec{v}) = \sum_{a \in S} \mathrm{sw}(a, \vec{v})$. Goel et al. (2016) advocate a *utilitarian* solution to participatory budgeting, that is, selecting

$$S^* \in \arg\max\{\mathrm{sw}(S, \vec{v}) : c(S) \leqslant B, S \subseteq A\}.$$

We subscribe to this utilitarian point of view. However, as noted earlier, eliciting the numerical utility functions (hereinafter, *utilities*) is too taxing for voters. Hence, real-world participatory budgeting systems collect partial information about voter preferences using a less taxing input format. We denote by ρ_i the vote cast by voter i in a given input format, and call $\vec{\rho} = (\rho_i, \dots, \rho_n)$ a *vote profile*. Let $v_i \rhd \rho_i$ denote that vote ρ_i is *consistent* with utility function v_i; $\vec{v} \rhd \vec{\rho}$ is defined analogously. We consider five input formats.

- A *ranking by value* of voter i is a strict total order over the alternatives, denoted σ_i. Let $\sigma_i(a)$ denote the position of alternative a in σ_i. We say that σ_i is consistent with utility function v_i if $\sigma_i(a) \leqslant \sigma_i(a')$ whenever $v_i(a) \geqslant v_i(a')$ for $a, a' \in A$.
- A *threshold approval vote* (Benadè et al. 2017) with threshold t of voter i is a binary vector $\tau_i \in \{0, 1\}^m$. This represents the alternatives for which the voter has utility at least as high as t. We say that τ_i is consistent with utility function v_i if for all $a \in A$, $\tau_i(a) = 1$ (i.e., a is *approved*) if and only if $v_i(a) \geqslant t$.
- A k-*approval* vote of voter i is a binary vector $\alpha_i \in \{0, 1\}^m$ with $\sum_{a \in A} \alpha_i(a) \leqslant k$. This represents the voter's (at most[5]) k most preferred alternatives. We say that α_i is consistent with utility function v_i if, for all $a, a' \in A$, $\alpha_i(a) > \alpha_i(a')$ implies $v_i(a) \geqslant v_i(a')$.

The remaining two input formats we study are proposed by Goel et al. (2016). In these formats, voters consider the cost of alternatives when expressing their preferences.

[5]In our experiments all voters except two chose exactly k.

Item	Cost	Utility
Mirror	10	5.8
Top coat	20	2.3
Water	3	29.3
Map	8	9.5
Pocket knife	5	14.8
Compass	5	9.4
Raincoat	10	5.4
First aid kit	10	14.9
Pistol	30	6.7
Sunglasses	25	1.9

Table 1: The 10 items used in the experiment, their costs, and voters' reported mean utilities. The budget is \$65.

- A *ranking by value for money* of voter i is a ranking σ_i of the alternatives by their 'bang for the buck'. Formally, we say that σ_i is consistent with utility function v_i if $\sigma_i(a) \leqslant \sigma_i(a')$ whenever $v_i(a)/c(a) \geqslant v_i(a')/c(a')$ for $a, a' \in A$.
- A *knapsack vote* of size B of voter i is a binary vector $\kappa_i \in \{0, 1\}^m$ with $\sum_{a \in A : \kappa_i(a) = 1} c(a) \leqslant B$. This represents the set of alternatives with total cost at most B that voter i has the highest total utility for. We say that κ_i is consistent with utility function v_i if $\{a \in A : \kappa_i(a) = 1\}$ is in $\arg\max\{\mathrm{sw}(S, v_i) : c(S) \leqslant B, S \subseteq A\}$.

Note that the utility function v_i is the most expressive; if this is known, one can induce voter i's vote in each of the five formats, up to ties.

3 Experimental Setup

We recruited more than 1200 voters on Amazon Mechanical Turk for our experiments, and asked them to evaluate a hypothetical scenario. Voters were told that they are stranded on a desert island, there is a set of items which may increase their chances of survival, each item has a cost, and there is a budget of \$65. The list of items is shown in Table 1, along with voters' reported mean utility for each item.

This abstract task is inspired by studies of group decision making (Hall and Watson 1970), and asks for a choice from a set of items, as in participatory budgeting. It was selected to eliminate biases based on voters' locations. For example, if we were to confront voters with a more traditional participatory budgeting setting in which one potential project involves upgrading a park, one may expect voters' utilities to vary drastically based on the health of their city's existing park system. This effect would be missing in real-world participatory budgeting elections, in which voters are typically residents of the same city.

In our experiments, voters are asked to report their preferences over the items in one of the five input formats described in Section 2 and/or report their numerical utilities for the different items. Votes in each input format (and utilities) are elicited using a dedicated user interface. Figure 1 shows the user interface for knapsack, in which voters use checkboxes to select items. Below, we describe how votes are elicited through each interface.

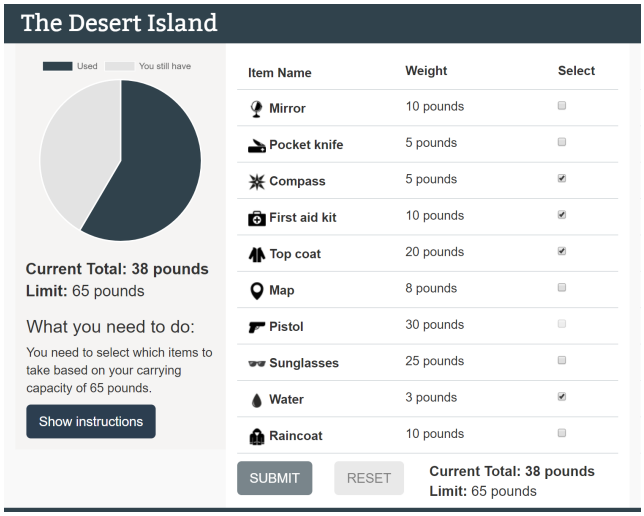

Figure 1: Screenshot of the knapsack interface.

- *Knapsack vote*: Voters are shown the (relatively elaborate) interface of Figure 1. The task is: "You need to select which items to take based on your carrying capacity of 65 pounds."
- *Ranking by value*: Voters are shown the list of items in a drag-and-drop interface. The task is: "Rank the items from the most important to the least important according to your best judgment."
- *Ranking by value for money*: Voters are shown the list of items and their weights in a drag-and-drop interface. The task is: "If you had to divide 100 points among the items based on how much you like them, rank the items in the decreasing order of the number of points they would receive divided by the cost."
- *Threshold approval*: Voters are shown a list of items with checkboxes. The task is: "If you had to divide 100 points among the items based on how much you like them, select all the items that would receive at least 10 points."
- 5*-approval*: Voters are shown a list of items with checkboxes. The task is: "You need to select up to 5 items from a list of 10 items according to your best judgment."
- *Utilities*: Voters are shown a list of items and sliders that control the number of points given to each. The task is: "You need to distribute 100 points among 10 items. The more points you assign to an item, the more important you think the item is to your survival."

We conducted two studies, which we refer to as A and B. In study A, 720 voters were recruited; each voter was randomly assigned one of the above input formats and cast a single vote in this format. This yields 120 votes in each format. The dataset from this study is used in the experiments detailed in Sections 4 and 5.1.

In study B, an additional 500 voters were recruited, and engaged in a two-stage process. In the first stage, half of the voters were asked to vote using one of the five input formats (randomly assigned). In the second stage, these voters were asked to specify their utility for each item. After each step, the voters were asked to rate how easy they found the activity, and how much they liked the user interface. To control for ordering effects, the other half of the voters were asked to perform the two stages in the reverse order (i.e., specify utilities in the first stage, and vote in a given input format in the second stage). The dataset from this study is used for the experiments detailed in Sections 5.1 and 5.2.

In both studies, participation is contingent on voters reading a short tutorial, passing a pre-task quiz which verifies voters' comprehension of the interface, and passing a post-task quiz which asks voters questions about their votes to ensure that the votes received at least some consideration. For example, the post-task quizzes in the ranking by value and ranking by value for money formats ask voters whether top coat was positioned higher than water in their ranking. Voters were paid 20 cents for completing the tutorial and the pre-task quiz, and a bonus of 10 cents for completing the post-task quiz.

4 Efficiency

Given the underlying utility functions $\vec{v}$ of the voters, our goal is to choose an optimal (welfare-maximizing) budget-feasible set of alternatives:

$$S^* \in \arg\max\{\mathrm{sw}(S, \vec{v}) : S \in \mathcal{F}_c\},$$

where

$$\mathcal{F}_c = \{S \subseteq A : c(S) \leqslant B\}$$

is the collection of all budget-feasible sets of alternatives. When choosing a suboptimal set $S \in \mathcal{F}_c$, we face an efficiency loss defined as

$$\mathrm{EL}(S, \vec{v}) = 1 - \frac{\mathrm{sw}(S, \vec{v})}{\max_{T \in \mathcal{F}_c} \mathrm{sw}(T, \vec{v})}.$$

In words, an efficiency loss of 0.05 (5%) means that the set chosen achieves 95% of the optimal welfare.

When votes are cast in an input format, we have access only to the votes $\vec{\rho}$, and not to the utility profile $\vec{v}$. While $\vec{\rho}$ provides partial information regarding $\vec{v}$ (specifically, that $\vec{v} \rhd \vec{\rho}$), some efficiency loss is inevitable. How do we minimize efficiency loss, despite the uncertainty about $\vec{v}$?

Benadè et al. (2017) advocate choosing the *distortion-minimizing* set, which minimizes the worst-case efficiency loss, where the worst case is taken over the utility profile $\vec{v}$ subject to the condition that $\vec{v} \rhd \vec{\rho}$. Formally, let f^* denote the deterministic aggregation method that returns a distortion-minimizing set. Then,

$$f^*(\vec{\rho}) \in \arg\min_{S \in \mathcal{F}_c} \sup_{\vec{v}: \vec{v} \rhd \vec{\rho}} \mathrm{EL}(S, \vec{v})$$

The benefits of the distortion-minimization approach is that it provides a consistent aggregation method across different input formats. See Section 6 for a discussion of the potential implications of using other aggregation methods.

In our efficiency experiment, we want to evaluate and compare the efficiency loss of the distortion-minimizing set chosen based on votes in each input format. Instead of evaluating the efficiency loss in the worst case, we want to evaluate it using the underlying utility profile. Specifically, we

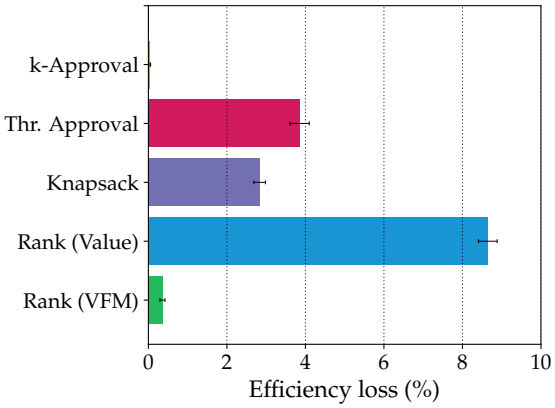

Figure 2: The average efficiency loss for each input format. Lower is better.

take the dataset from study A, sample 60 voters for each input format, compute the distortion-minimizing set for the corresponding vote profile, and evaluate its efficiency loss using the utility profile of another sample of 60 voters who were asked to submit their utility functions.

A crucial insight behind this experiment, which is necessary for its validity, is that the *average* utility of an item, according to the utility profile of the second set of voters, closely approximates its average utility according to the first set of voters. This is intuitively true by the law of large numbers, and is confirmed by our experiments in Section 5.1. For this reason, we can accurately estimate the social welfare of a subset of items with respect to the first set of voters, *without* asking these voters to report both utilities and votes.

Figure 2 reports the average efficiency loss (in percent) across 1000 repetitions of this experiment. The Mann-Whitney U test found a statistically significant difference in performance (at the $p = 0.05$ level) between every pair of input formats except between k-approval and ranking by value for money. Both k-approval and ranking by value for money perform incredibly well and achieve social welfare within 0.5% of optimal, suggesting that they capture sufficient information about voter preferences to allow computation of near-efficient outcomes. The worst performance is demonstrated by ranking by value, which incurs an 8% efficiency loss on average.

5 Usability

For an input format to be viable for deployment in participatory budgeting elections, we expect it to allow voters to accurately and quickly express their preferences, while also being easy to understand and use. To that end, we measure the usability of an input format through both *objective* and *subjective* indicators. While the objective indicators of usability are computed from data, the subjective indicators are self-reported by the voters.

We focus on two objective indicators. First, we want to ensure that the votes cast by voters in an input format are consistent with the utility functions expressed by the (same or different) voters. We call this *consistency*. Second, we record the amount of time it takes for voters to complete the tutorial and cast their vote, which is an indicator of the cognitive burden. We call this *response time*.

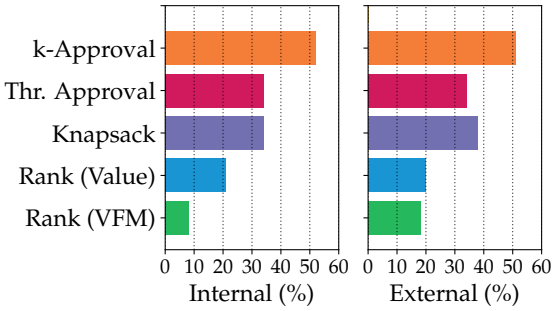

Figure 3: Internal and external consistency of different input formats. Higher is better.

We additionally ask voters about their experience of casting a vote in their assigned input format, and record three subjective indicators of usability: how easy it is to cast a vote, how much they like the user interface, and how well the input format allows them to express their preferences.

5.1 Objective Indicators

As noted earlier, we measure two objective indicators of usability: consistency between votes and utility functions, and time taken by voters.

Consistency. Intuitively, consistency measures whether voters' reported utility functions induce their votes cast in a given input format. If an input format allows voters to accurately express their preferences, we may expect a high level of consistency. We measure consistency in two forms.

For *internal consistency*, we call a voter consistent if the voter's reported utility function is consistent with that *same* voter's vote in the assigned input format (i.e., the utility function induces the vote, up to any ties). For each input format, we report the percentage of consistent voters.

Recall that we chose the desert island setting with the assumption that it minimizes the effect of voters' contextual background. If this assumption holds, we should expect consistency even between the votes and the utility functions reported by different sets of voters. We refer to this as *external consistency*. For this, we use data from study A, and from the first stage of study B. For each input format, the submitted votes form a vote profile $\vec{\rho}$, and the submitted utilities functions form a utility profile $\vec{v}$. We measure the fraction of votes induced by $\vec{v}$ that match with votes in $\vec{\rho}$. Formally, to account for ties, we create a bipartite graph with votes from $\vec{\rho}$ on one side and utility functions from $\vec{v}$ on the other, and add an edge between vote ρ_i and utility function v_j when $v_j \rhd \rho_i$. We report the percentage of matched votes, or the cardinality of the maximum matching divided by 170 (the number of vertices on each side).

The results are provided in Figure 3. k-approval is comfortably the best in terms of both internal and external con-

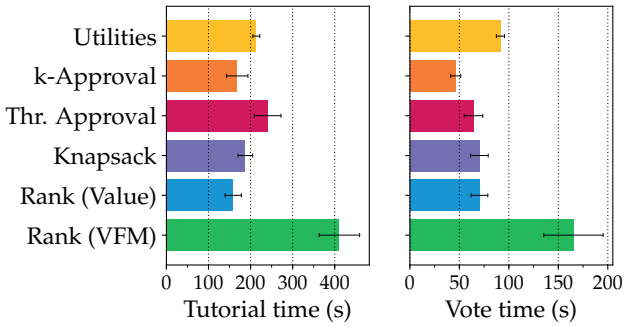

Figure 4: Average time taken (in seconds) to complete the pre-task tutorial and to cast a vote in each input format. Lower is better.

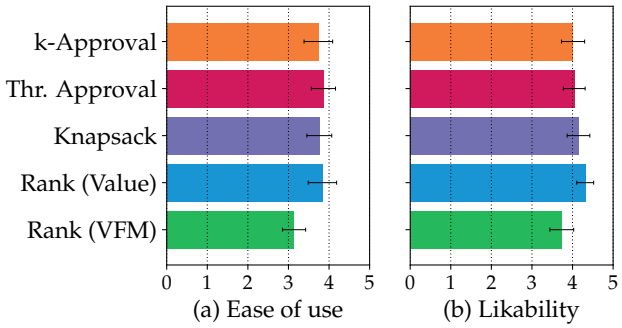

Figure 5: How easy to use each input format is, and how liked its user interface is, based on the subjective reports of the voters on a scale of 0 to 5, 5 being the best.

sistency (both above 50%). We find the internal consistency of knapsack to be surprisingly high: more than a third of the voters can report exact solutions to their personal knapsack problem, the computational hardness of the knapsack problem and the sheer number of budget-feasible subsets of alternatives notwithstanding. Ranking by value for money and ranking by value perform poorly in both forms of consistency. It is tempting to claim that this is due to the space of possible rankings being exponentially large, but as noted above, knapsack votes perform well despite this obstacle.

Finally, we remark that the high degree of similarity between internal and external consistency for each input format is yet another strong indication that the utility profile of one set of voters serves as a good substitute for the utility profile of another set of voters, which is a foundational assumption for the validity of our between-user study.

Response time. The response time to complete a task is recognized as a proxy for the objective difficulty (or cognitive load) associated with the task (Rauterberg 1992). For each input format, we report, in Figure 4, the average amount of time it took to learn how to vote in the format (complete the tutorial and pass the quiz) and to cast a vote in the format.

In terms of the difficulty of learning an input format, k-approval and ranking by value are the easiest (the difference between them is not statistically significant), followed by knapsack and threshold approval. Ranking by value for money is the most difficult by a wide margin.

In terms of the time taken to cast a vote, k-approval is also by far the fastest input format, at 45 seconds on average. Knapsack and ranking by value take about 70 seconds, while ranking by value for money is again the slowest by a wide margin, at almost 3 minutes.

For reference, we also report how long it takes for voters to submit their utility functions. At 96 seconds, this is slower than every input format except ranking by value for money. This largely supports the belief that it is taxing for voters to report their exact utility functions.

Summary. The objective indicators of usability overwhelmingly point to k-approval. It is distinctively the best at allowing voters to quickly learn the format and cast a vote, and results in votes that are by far the most consistent with the voters' utility functions. By contrast, ranking by value for money performs miserably. It takes voters more than three times longer to vote using this format than under k-approval, and the resulting votes have little in common with the voters' utility functions.

5.2 Subjective Indicators

In addition to computing objective indicators of usability, we asked 500 voters in study B to report their experiences with different input formats, and measured various subjective indicators of usability. When we say below that a result is statistically significant, we are referring to the Mann-Whitney and Wilcoxon signed-rank tests at the $p < 0.05$ level.

Ease of use. We asked voters to report how easy they found the voting task on a scale of 0 to 5 (5 being the easiest). The perceived (subjective) difficulty is reported in Figure 5(a). Ranking by value for money is significantly worse than every other input format, while the differences between the other input formats are not statistically significant.

User interface. We also asked voters to report how much they liked the user interface on a scale of 0 to 5 (5 being the most liked). As seen in Figure 5(b), ranking by value and knapsack are the most liked interfaces, followed by k-approval and threshold approval (with no significant difference between each pair). Ranking by value for money was again the least liked.

We believe that this is somewhat correlated with the inherent difficulty of an input format because our choice of user interface was standard in most cases. However, the results are subject to change with design of better user interfaces.

Perceived Expressiveness. We asked voters to report how well their assigned input format captured their preferences on a scale of 0 to 5. As seen in Figure 6, ranking by value is reported to be much more expressive than any other input format (by a statistically significant margin), while k-approval and threshold approval votes are the least expressive. Although voters dislike using ranking by value for money, they still feel that it captures their preferences well.

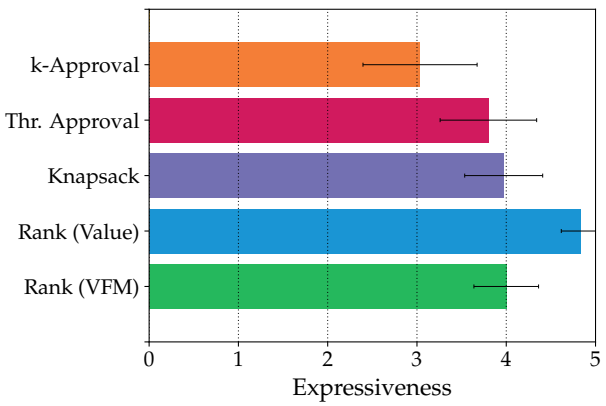

Figure 6: Voters' perceived expressiveness of different input formats. Higher is better.

Summary. In terms of subjective indicators, ranking by value seems the most preferred input format: voters feel that it best captures their preferences, and no other input format is more easy to use or liked (in a statistically significant manner). Ranking by value for money is again the most difficult to use and least liked, although voters feel it captures their preferences fairly well.

6 Discussion

Our results shed light on the efficiency and usability of five input formats used in participatory budgeting. Somewhat surprisingly, the most popular — and arguably the simplest — input format, k-approval, outperforms every other input format in terms of efficiency (welfare loss) and objective indicators of usability (consistency of votes and response time). In terms of the subjective indicators, no input format is statistically easier to use than k-approval, while the user interfaces of ranking by value and knapsack are only somewhat more liked than that of k-approval.

The results for the third subjective indicator, namely expressiveness, are the only ones that prevent k-approval from being dominant across the board. Indeed, voters feel that k-approval is the worst in capturing their preferences, while ranking by value is the best. Our efficiency experiments reveal that, in fact, the exact opposite is true: k-approval contains information that leads to the most efficient outcomes, while ranking by value leads to the least efficient ones. This highlights the distinction between what voters feel is important when casting a vote, and what is needed to enable efficient aggregation.

Ranking by value performs well in terms of subjective indicators of usability, and somewhat worse in terms of the objective indicators. However, it is especially concerning that it leads to outcomes that have relatively low social welfare.

Knapsack performs reasonably on all indicators, including surprisingly good response times, which corroborate the results of Goel et al. (2016). Based on our discussions with practitioners in Europe, it seems that the fact that this input format encourages voters to directly reason about the government's budget constraints is also seen as an advantage, which could potentially outweigh some of the disadvantages shown by our results.

The subjective and objective indicators agree that voters find ranking by value for money to be difficult to use and that these votes rarely reflect voters' true utility functions (but mysteriously lead to efficient outcomes). This cautions strongly against the use of ranking by value for money, although it is less clear what the implications are for pairwise value for money comparisons as advocated by Goel et al. (2016).

Finally, we discuss several limitations of our study, and point to directions for future work. First, our efficiency results use the distortion-minimizing aggregation method for each input format. While this provides a consistent choice across input formats, it would be interesting to use more realistic (e.g., greedy) aggregation methods to better understand the efficiency loss in practice. Second, our results are closely tied to our choice of user interfaces for eliciting voter preferences. Arguably, a better user interface can lead to increased measures of usability, including votes that are more consistent with voters' utility functions, which in turn can lead to greater efficiency. Hence, the design of improved, more intuitive user interfaces is an important direction for future research.

Next, in all of our experiments, except in the measurement of internal consistency, we only used data generated by asking voters to vote in a single input format. This choice was based on the assumption that asking voters to vote in multiple formats would not only be tiring, but can also affect the votes themselves. This was partially confirmed by our measurements of internal consistency. We observed that if we ask voters to report their utility functions and cast their votes using an input format, voters are generally far less consistent when utility functions are reported first. However, there is a need for more thorough experiments to identify and understand the effects of asking voters to report their preferences in multiple forms.

We note that our desert island setting uses 10 items (alternatives), while real participatory budgeting elections may require voters to compare more items. We limited the number of items to allow voters to accurately report their utility functions, which was necessary to measure consistency and efficiency loss. An important direction for future work is to study voter behavior when evaluating more than 10 items, which may require indirectly measuring consistency and efficiency loss without access to the utility functions.

More broadly, while our desert island setting provides a good abstraction of participatory budgeting and reduces the effect of voters' contextual background, it makes the voters a bit too homogeneous. In our setting, it is likely that *all* voters have similar preferences. By contrast, in participatory budgeting, it is likely that voter preferences are clustered based on factors such as personal interests and geographical location. Studying the structure of voter preferences and its effect on the choice of efficient outcomes in real participatory budgeting elections is perhaps the most compelling direction for future research.

References

Anshelevich, E., and Postl, J. 2016. Randomized social choice functions under metric preferences. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 46–52.

Anshelevich, E., and Sekar, S. 2016. Blind, greedy, and random: Algorithms for matching and clustering using only ordinal information. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, 390–396.

Anshelevich, E.; Bhardwaj, O.; and Postl, J. 2015. Approximating optimal social choice under metric preferences. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, 777–783.

Benadè, J. G.; Nath, S.; Procaccia, A. D.; and Shah, N. 2017. Preference elicitation for participatory budgeting. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, 376–382.

Boutilier, C.; Caragiannis, I.; Haber, S.; Lu, T.; Procaccia, A. D.; and Sheffet, O. 2015. Optimal social choice functions: A utilitarian view. *Artificial Intelligence* 227:190–213.

Brandt, F.; Conitzer, V.; Endress, U.; Lang, J.; and Procaccia, A. D., eds. 2016. *Handbook of Computational Social Choice*. Cambridge University Press.

Cabannes, Y. 2004. Participatory budgeting: A significant contribution to participatory democracy. *Environment and Urbanization* 16(1):27–46.

Camerer, C. 2011. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Caragiannis, I.; Nath, S.; Procaccia, A. D.; and Shah, N. 2017. Subset selection via implicit utilitarian voting. *Journal of Artificial Intelligence Research* 58:123–152.

Goel, A.; Krishnaswamy, A. K.; Sakshuwong, S.; and Aitamurto, T. 2016. Knapsack voting for participatory budgeting. Manuscript.

Hall, J., and Watson, W. H. 1970. The effects of a normative intervention on group decision-making performance. *Human Relations* 23(4):299–317.

Procaccia, A. D., and Rosenschein, J. S. 2006. The distortion of cardinal preferences in voting. In *Proceedings of the 10th International Workshop on Cooperative Information Agents (CIA)*, 317–331.

Rauterberg, M. 1992. A method of a quantitative measurement of cognitive complexity. In van der Veer, G.; Tauber, M.; Bagnara, S.; and Antalovits, M., eds., *Human-Computer Interaction: Tasks and Organisation*. CUD. 295–307.

Röcke, A. 2014. *Framing Citizen Participation: Participatory Budgeting in France, Germany and the United Kingdom*. Palgrave Macmillan.