Enhancing Collective Estimates by Aggregating Cardinal and Ordinal Inputs

Ryan Kemmer, Yeawon Yoo, Adolfo R. Escobedo*, Ross Maciejewski
School of Computing, Informatics, and Decision Systems Engineering, Arizona State University
{rwkemmer, yyoo12, adres, rmacieje}@asu.edu

Abstract

There are many factors that affect the quality of data received from crowdsourcing, including cognitive biases, varying levels of expertise, and varying subjective scales. This work investigates how the elicitation and integration of multiple modalities of input can enhance the quality of collective estimations. We create a crowdsourced experiment where participants are asked to estimate the number of dots within images in two ways: ordinal (ranking) and cardinal (numerical) estimates. We run our study with 300 participants and test how the efficiency of crowdsourced computation is affected when asking participants to provide ordinal and/or cardinal inputs and how the accuracy of the aggregated outcome is affected when using a variety of aggregation methods. First, we find that more accurate ordinal and cardinal estimations can be achieved by prompting participants to provide both cardinal and ordinal information. Second, we present how accurate collective numerical estimates can be achieved with significantly fewer people when aggregating individual preferences using optimization-based consensus aggregation models. Interestingly, we also find that aggregating cardinal information may yield more accurate ordinal estimates.

Introduction

Many crowdsourced activities utilize multiple people to collectively classify information, predict events, and make decisions (Galton 1907; Chittilappilly, Chen, and Amer-Yahia 2016). Because of varying subjective and numerical scales among humans, individual responses can be conflicting and it is usually unreasonable to trust a single person to provide a definitive result (Mao, Procaccia, and Chen 2013). However, results can be promising when individual tasks are completed independently by many people and properly aggregated to produce a collective decision/estimate. This is a principle commonly referred to as the "wisdom of crowds", which theorizes that the aggregated judgments of multiple people will be relatively accurate, even with error-prone individuals (Simmons et al. 2011; Surowiecki 2005). This concept has been recently been applied to crowdsourced estimation tasks, such as estimating the number of dots in images (Horton 2010), as well as many real-world domain problems, including determining collective human ethics to drive the decision making of self-driving cars (Noothigattu et al. 2017) and improving public health by gathering participant-reported information (Brabham et al. 2014).

Various ways of prompting questions to crowdsourced workers have been shown to affect the quality of data provided by individuals, and the efficiency of data collection (Chung et al. 2019). Furthermore, the modality of data that is collected can play a role in determining individual and group opinions. Studies in psychology have shown that eliciting rankings to compare objects can yield different results than using rating scales, the latter of which tend to exhibit higher variability (Rankin and Grube 1980; Ovadia 2004). Moreover, the size of the task given to each individual can also have an impact on the overall accuracy and efficiency of crowdsourced computation. In crowdsourced comparison tasks, it has been found that there are big trade-offs between problem size, worker effort, and the quality of data collected. (Wilber, Kwak, and Belongie 2014).

A number of works have demonstrated that the quality of a collective decision/estimate is highly dependent on the aggregation method employed (Mao, Procaccia, and Chen 2013). Such concerns fall under the purview of computational social choice, a field dedicated in part to the rigorous design of preference data aggregation mechanisms (Brandt et al. 2016). An ongoing debate in this field, and in group decision-making in general, centers on the selection of *ordinal data* (i.e., rankings) or *cardinal data* (i.e., ratings or scores) to elicit and aggregate preferences. Each of the two modalities is said to capture key distinct characteristics (e.g., ability to express indifference, intensity of preference, resp.), but each also possesses well-known theoretical shortcomings (e.g., Arrow's Impossibility Theorem (Arrow 1951), subjectivity of scales, resp.). The multitude of aggregation methods that exist can be divided roughly into standard statistical methods (e.g., average), efficient "voting rules" (e.g., Borda rule (Brandt et al. 2016)), and optimization-based consensus aggregation models (e.g., ranking aggregation (Cook 2006)). While optimization-based aggregation models can be more computationally demanding, they also tend to be more resistant to individ-

*Corresponding author

ual voter error, bias, and manipulation (Brandt et al. 2016; Dwork et al. 2001).

This work investigates the effects of asking participants questions whose answers require different input modalities, specifically ordinal and cardinal estimates, and tests how well different aggregation techniques perform in approximating an underlying ground truth. We aim to answer four main questions:

- How is the quality of crowdsourced computation affected when data is collected in ordinal form (ranking) and/or cardinal form (rating/numerical estimation)?
- Does aggregating ordinal and cardinal data together using multimodal aggregation models produce more accurate collective estimates?
- How does the size of the problem distributed to participants influence different aggregation methods?
- What is the size of the crowd required to achieve good results, and how does this differ depending on the aggregation method used?

To address these questions, we design a crowdsourced experiment that elicits ordinal and cardinal inputs to perform two related but distinct estimation tasks: ordering a set of images based on the number of dots they contain and estimating the number of dots contained in each of the individual images from the image set. Our results indicate that integrating ordinal and cardinal estimates can improve accuracy and efficiency of crowdsourced computation. Henceforth, the terms ranking (resp., rating) and ordinal (resp., cardinal) information are used interchangeably.

Related Works

Crowdsourcing has many unique challenges related to the way in which data is collected and aggregated. Cognitive biases such as anchoring effect, bandwagon effect, and decoy effect have been found to be present while collecting data in crowdsourced tasks, and negatively affect results (Eickhoff 2018). Furthermore, studies have shown that in subjective labeling tasks, systematic biases stemming from worker opinions can produce overall biased results (Hube, Fetahu, and Gadiraju 2019).

Researchers have developed a variety of methods to cope with human unreliability. Different methods of quality control such as honeypot traps and expectation-maximization algorithms have been used to mitigate worker error (Quoc Viet Hung et al. 2013). In the context of image annotation, researchers have developed multiple quality control systems, including ways to have other participants perform quality checks (Su, Deng, and Fei-Fei 2012). Machine learning techniques have also been employed to aid in agent selection and data post-processing, and systems have been developed to extract high-quality collective intelligence at low cost (Davis-Stober et al. 2015; Goldstein, McAfee, and Suri 2014). Along with quality control systems, stopping rules have also been proposed in a number of human computation tasks to control the number of worker responses needed for individual questions and to achieve maximum efficiency with respect to accuracy (Abraham et al. 2014).

While there are many methodologies to cope with noisy and unreliable human data in crowdsourcing, there has not been much research specifically on broadening and experimenting with multiple types of *input elicitation*, that is, the type of information asked from crowd members regarding a particular question. This direction appears to be promising. One study investigating crowdsourcing to annotate data into clusters found that simple changes to the worker interface could have a significant impact on the quality of data collected and on the cost spent by crowdsourcers (Wilber, Kwak, and Belongie 2014). Furthermore, a study in the field of computational social choice investigated input elicitation in the context of a crowdsourced participatory budgeting problem, and found that presenting participants with different subjective scales and interfaces produced varying results (Benade 2018). Another study about collecting top group preferences found that using rankings to help refine the scores of ratings can better infer top-k results (Li, Zhang, and Li 2018).

The fields of sociology and economics have investigated other phenomena that may play an influence on how to better collect data from humans. There have been a variety of interesting predictive abilities of crowds discovered, including accurate predictions by mere recognition (Herzog and Hertwig 2011). A study related to crowd wisdom found that individuals perform better at predictive tasks with a process called "dialectical bootstrapping", which involves individual participants answering the same question multiple times (Herzog and Hertwig 2009). Another study found that when predicting the outcomes of NFL games, crowds performed better when participants were prompted to estimate the final score instead of which team would cover the point spread (Simmons et al. 2011). This suggests the way that questions are presented to individual participants could influence the quality of the resulting collective estimates.

Researchers have recently investigated how aggregation methods from computational social choice can enhance the accuracy of collective estimates. For example, in (Mao, Procaccia, and Chen 2013) popular voting rules such as plurality, Borda rule (Brandt et al. 2016), and Thurstone's model (Thurstone 1927) and the Kemeny consensus aggregation model (Kemeny and Snell 1962) were applied to solve ranking problems including ordering images based on the number of dots they contain. This and other studies (e.g., (Werbin-Ofir, Dery, and Shmueli 2019)) have concluded that the quality of the collective ranking depends on the method used and the context of the problem. Other works have devised specialized consensus aggregation methodologies to enable the extraction of accurate collective decisions in highly distributed decision-making contexts (Dwork et al. 2001; Yoo, Escobedo, and Skolfield 2020). Of particular interest to the present study are (Moreno-Centeno and Escobedo 2016) and (Fishbain and Moreno-Centeno 2016), which devised extensions of the Kemeny-Snell ranking distance and the Cook and Kress rating distance, respectively. These two generalized measures were derived from axiomatic foundations for guaranteeing that individual rankings and ratings, respectively, receive equal voting power in the aggregation, independent of how many objects they eval-

uate. The implications of these axiomatic foundations could be beneficial in crowdsourcing, where participants often do not evaluate exactly the same set of objects due to time limitations and where a high number of objects must be evaluated altogether.

Overall, research has indicated many potential ways to improve the process of collecting and aggregating crowdsourced data. While there are many strategies being currently employed, a review of the literature suggests that this challenge can be addressed by collecting richer data from individuals. Additionally, previous work suggests that specialized aggregation methods for integrating this data should be considered for making good use of this information.

Experimental Design

In order to evaluate the research questions proposed, we developed a human computation experiment that involved collecting cardinal and ordinal inputs. Estimating the number of dots in an image is considered a benchmark task in human computation, as it produces a high amount of variable results among participants, but, nevertheless, tends to return highly accurate results when a large enough number of participants evaluates the same images (Horton 2010; Mao, Procaccia, and Chen 2013). We therefore chose a dot estimation activity to measure the impact of different input elicitation modalities and aggregation methods.

When prompting a group of people to collectively answer a question, the way in which the question is framed can influence the collective decision. To test this theory, we prompted participants questions in both cardinal and ordinal formats to see how the way questions are framed influences collective results. There are multiple logically equivalent ways to collect ordinal information (Chen et al. 2013). One efficient way to do this is to collect a ranking of multiple alternatives, as this method obtains multiple pairwise comparisons at a time. Hence, ordinal inputs were elicited by asking participants to rank a subset of images by the number of dots they contain. Cardinal inputs were elicited by asking participants to provide a numerical estimate of the number of dots contained in each of the same images. Participants' inputs contributed to four larger problems of ranking 30 images from least to greatest number of dots contained and of estimating the number of dots in each of the 30 images.

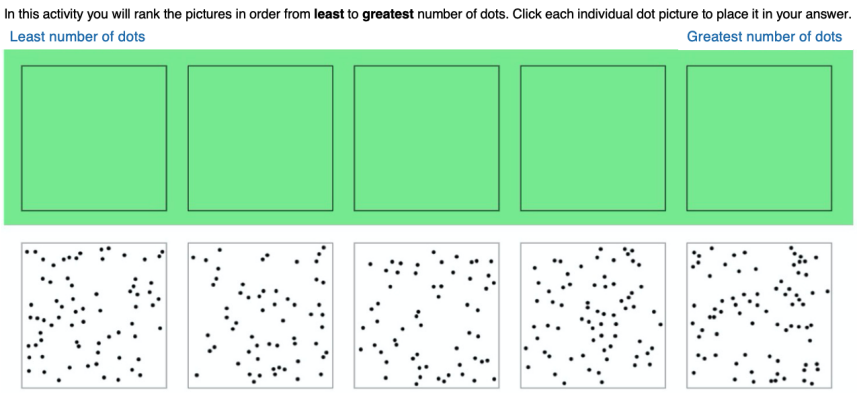
The experiments were designed to observe how collective estimates change with problem size. The study asked participants four problems of varying sizes, specifically 2-image, 3-image, 5-image, and 6-image problems, each with its own data set of 30 images. For a specific problem size, all 30 images are seen the same number of times; however, this frequency varies across the problem sizes. When the data obtained from all 300 participants are considered, all images are seen 20 times in the 2-image problem, calculated as $x =$ 300 participants/ (30 images/ 2 images seen per participant). Similarly, all images are seen 30, 50, and 60 times in the 3-image, 5-image, and 6-image problems, respectively.

In each problem data set, each image has a unique number of dots, ranging from 50 to 79 dots. This range was chosen primarily to make the ordinal estimation task roughly as difficult as the cardinal estimation task. This can be explained as follows. Participants see only a small subset of the 30 images in each problem data set—specifically 2, 3, 5, or 6 images. Because these subsets are assigned randomly, it is possible for some participants to be assigned ordinal estimation tasks requiring comparatively less effort. For example, in the 2-image problem, one participant could be assigned the 50-dot and 79-dot images, which would represent a relatively simpler task than that of another participant who is assigned the 50-dot and 55-dot images, whose separation is narrower. Along this line of reasoning, a broader dot range, say of 50 to 200, may engender ordinal estimation tasks that are relatively trivial to complete. In summary, the range of dots contained in each of the four problem data sets was 30, which tended to provide nontrivial ordinal estimation tasks, as can be seen for example in Figure 1a. It is also worthwhile to remark that participants were not explicitly incentivized to manually count the number of dots in each image, i.e., to do well in the cardinal estimation task. That is, workers did not receive extra compensation for providing accurate results.

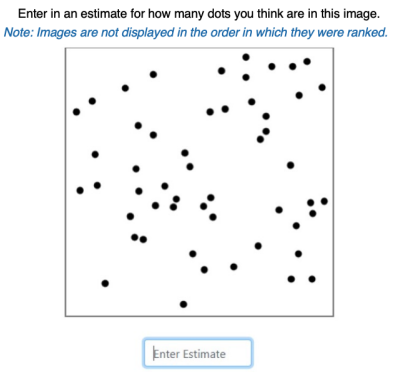
Web application

A web application and user interface were developed to collect ordinal and numerical estimates from participants in the experiment. The application was then published as an activity on Amazon MTurk, a popular crowdsourcing platform (Buhrmester, Kwang, and Gosling 2011). Each person who accepted our activity was shown an experiment briefing page and then was prompted to type their MTurk ID to enter the activity. Once a new participant entered the activity, the participant was asked to complete all four varying-size problems; the order the problems appear was randomized. The images shown to participants under a specific problem were drawn from the next available segment of a previously generated random permutation of the 30 images. For example, if the previously generated random permutation of the 30 images was (79,78,...,50) in the 2-image problem and a previous participant received images 79 and 78, then the next participant would be assigned images 77 and 76.

Each problem asked the participant to first rank the images by the number of dots they contain and then to provide numerical estimates of how many dots are in each individual image. Numerical estimates were elicited by showing the same images from the ordinal estimation tasks, but shown one image at a time and in a randomized order—this was done to ensure that cardinal estimates are not necessarily anchored to the order indicated by the participant answers to the ordinal estimation task. The user interface for the ordinal estimation of each question is shown in Figure 1a, and the numerical estimation of each question is shown in Figure 1b. If a participant did not complete the task correctly, he or she was prompted to try the same question again. For the cardinal and ordinal estimation questions, participants had a chance to edit or change their answers before submitting them in case they made a mistake. Throughout the duration of the activity, the time it took each participant to complete each individual task was recorded and stored for analysis. At the end of the study, participants were asked to fill out a brief demographic questionnaire. Finally, they received a code to claim their financial compensation (approximately $1.00 for



(a) Interface for ordinal estimation



(b) Interface for cardinal estimation

Figure 1: User interface for estimation tasks

5 minutes of work).

Participants

A total of 300 participants completed the study. Participants were recruited from Amazon MTurk. Data from participants that started the study, but did not finish all of the questions, were removed. The demographic survey was completed by 288 of the 300 participants. The average age reported was 36.8, with a median age of 34. A total of 170 of them reported their gender as male, 117 reported female, and 2 reported their gender as other. Education level varied significantly among participants: 1 participant had education equivalent to less than a high school degree, 43 had completed high school/GED, 51 had completed some college, 29 had a 2-year degree, 127 had a 4-year degree, 32 had a master's degree, 1 had a doctoral degree, and 5 had a professional degree. A total of 236 participants reported they were employed, and 52 reported they were unemployed. A total of 286 participants were native English speakers and 2 were non-native speakers. Finally, a total of 83 participants reported they had completed similar estimation tasks and 205 reported they had not.

Aggregation methods

This section introduces the aggregation methods used in the experiments. Beforehand, some notation conventions used throughout the paper are described. Let $\boldsymbol{a}^\ell$ and $\boldsymbol{b}^\ell$ denote the ordinal and cardinal estimate vectors, respectively, gathered from participant ℓ. Also, a_i^ℓ and b_i^ℓ represents the rank position and cardinal estimation value of alternative i from participant ℓ, respectively. The subset of alternatives (i.e., images) evaluated in $\boldsymbol{a}$ (resp., $\boldsymbol{b}$) is denoted as $V_{\boldsymbol{a}}$ (resp., $V_{\boldsymbol{b}}$) and the set of participants is denoted as L. Ordinal estimates are assumed to be strict (i.e., ties are not allowed).

Optimization-based models

Ordinal Aggregation (Ranking Aggregation) The Ordinal Aggregation (OA) model is a ranking-based aggregation model, which minimizes the Normalized Projected Kemeny-Snell distance for incomplete rankings, written here succinctly as d_{NPKS} (Moreno-Centeno and Escobedo 2016) and defined as follows:

$$d_{NPKS}(\boldsymbol{a}^1,\boldsymbol{a}^2)=\begin{cases}\frac{d_{KS}(\boldsymbol{a}^1|_{(V_{\boldsymbol{a}^1}\cap V_{\boldsymbol{a}^2})},\boldsymbol{a}^2|_{(V_{\boldsymbol{a}^1}\cap V_{\boldsymbol{a}^2})})}{\bar{n}(\bar{n}-1)/2} & \text{if } \bar{n}\geq 2,\\ 0 & \text{else,}\end{cases}$$

where $\bar{n} := |V_{\boldsymbol{a}^1} \bigcap V_{\boldsymbol{a}^2}|$ and $\boldsymbol{a}^1|_{(V_{\boldsymbol{a}^1}\cap V_{\boldsymbol{a}^2})}$ and $\boldsymbol{a}^2|_{(V_{\boldsymbol{a}^1}\cap V_{\boldsymbol{a}^2})}$ denote the projections of each ranking onto the subset of alternatives evaluated in both rankings $\boldsymbol{a}^1$ and $\boldsymbol{a}^2$. The original Kemeny-Snell distance (Kemeny and Snell 1962), defined as $d_{KS}(\boldsymbol{a}^1,\boldsymbol{a}^2) = \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \left|\text{sign}(a_i^1-a_j^1)-\text{sign}(a_i^2-a_j^2)\right|$, counts the number of pairwise inversions between two rankings. The OA consensus aggregation (i.e., Kemeny) model is defined as follows:

$$\min_{\boldsymbol{u}} \sum_{\ell=1}^{|L|} d_{NPKS}(\boldsymbol{a}^\ell,\boldsymbol{u})$$

where $\boldsymbol{u}$ is a candidate aggregate ranking.

Cardinal Aggregation (Rating Aggregation) The Cardinal Aggregation (CA) model is a rating-based aggregation model, which minimizes the Normalized Projected Cook-Kress distance for incomplete ratings, written here succinctly as d_{NPCK} (Fishbain and Moreno-Centeno 2016) and defined as follows:

$$d_{NPCK}(\boldsymbol{b}^1,\boldsymbol{b}^2)=\begin{cases}\frac{d_{CK}(\boldsymbol{b}^1|_{(V_{\boldsymbol{b}^1}\cap V_{\boldsymbol{b}^2})},\boldsymbol{b}^2|_{(V_{\boldsymbol{b}^1}\cap V_{\boldsymbol{b}^2})})}{4R\cdot\left\lceil\frac{|V_{\boldsymbol{b}^1}\cap V_{\boldsymbol{b}^2}|}{2}\right\rceil\cdot\left\lfloor\frac{|V_{\boldsymbol{b}^1}\cap V_{\boldsymbol{b}^2}|}{2}\right\rfloor} & \text{if } \bar{n}\geq 2,\\ 0 & \text{else,}\end{cases}$$

where R is the range of the ratings. The original Cook-Kress distance (Cook and Kress 1985), defined as $d_{CK}(\boldsymbol{b}^1,\boldsymbol{b}^2) = \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \left|(b_i^1-b_j^1)-(b_i^2-b_j^2)\right|$, calculates the pairwise differences of intensity between ratings. The CA consensus aggregation model is defined as follows:

$$\min_{\boldsymbol{r}} \sum_{\ell=1}^{|L|} d_{NPCK}(\boldsymbol{b}^\ell,\boldsymbol{r}).$$

where $\boldsymbol{r}$ is a candidate aggregate rating.

Cardinal and Ordinal Aggregation The Cardinal and Ordinal Aggregation (COA) model jointly aggregates a set of ratings and a set of rankings by utilizing both d_{NPKS} and d_{NPCK} and giving equal weights to the two modalities of information. The aggregate rating-ranking solution obtained is logically coupled. Specifically, this consensus aggregation model finds the rating $\boldsymbol{r}$ and the ranking induced by ordering the values of $\boldsymbol{r}$ in non-increasing order, written as rank($\boldsymbol{r}$), which together yield the minimum cumulative ranking-rating distances to the multimodal inputs. The COA consensus aggregation model is defined as follows:

$$\min_{\boldsymbol{r}} \sum_{\ell=1}^{|L|} d_{NPCK}(\boldsymbol{b}^{\ell}, \boldsymbol{r}) + \sum_{\ell=1}^{|L|} d_{NPKS}(\boldsymbol{a}^{\ell}, \text{rank}(\boldsymbol{r})).$$

Separation-Deviation Aggregation The Separation-Deviation (SD) model is another multimodal model that takes into account the difference between the pairwise comparison of two alternatives i and j in the aggregated outcome and each participant's evaluations (separation)—given by the difference of intensities in ratings, as in d_{NPCK}—and the difference between the value of alternative i in the aggregated outcome and in each participant's evaluation (deviation) (Hochbaum 2010). In the consensus aggregation model, the separation is penalized by a function s_{ij}^{ℓ} and the deviation is penalized by a function d_i^{ℓ}. The model can be mathematically formulated as follows:

$$\min_{\boldsymbol{r}} \sum_{\ell=1}^{|L|} \Big(\sum_{i,j=1}^{n} s_{ij}^{\ell}((r_i - r_j) - (b_i^{\ell} - b_j^{\ell})) + \sum_{i=1}^{n} d_i^{\ell}(r_i - b_i^{\ell}) \Big)$$

where r_i represents the rating value of alternative i in the candidate solution.

All consensus aggregation models were solved using specialized mixed-integer and integer linear programming models specified in (Hochbaum and Levin 2006a; Fishbain and Moreno-Centeno 2016; Escobedo, Hochbaum, and Moreno-Centeno 2020; Yoo and Escobedo 2020).

Traditional voting rule-based methods

Plurality Rule The plurality rule selects an alternative with the most amount of first place votes. To obtain a complete ranking with this rule, the candidates are ordered based on the number of first-place votes they receive.

Borda Rule The Borda rule assigns a score to each candidate in a ballot according to how many candidates it defeats and then determines a complete ranking from the ordered candidate scores summed over all ballots (Brandt et al. 2016).

Copeland Rule The Copeland rule chooses the alternative with the highest number of pairwise wins minus defeats, i.e., Copeland scores. To obtain a complete ranking, candidates are ordered by non-increasing Copeland scores.

Average

The aggregated rating $\boldsymbol{r}$ from the average method is:

$$\boldsymbol{r} = \left(\frac{\sum_{\ell=1}^{|L|} b_1^{\ell}}{|L(1)|}, \frac{\sum_{\ell=1}^{|L|} b_2^{\ell}}{|L(2)|}, ..., \frac{\sum_{\ell=1}^{|L|} b_n^{\ell}}{|L(n)|} \right).$$

where $L(i) \subset L$ denotes the subset of participants who evaluated candidate (image) i.

Median

The median method finds the halfway point of the numerical estimates after arranging the estimates in order from least to greatest. Specifically, assuming $|L(i)|$ is odd, the aggregated ranking $\boldsymbol{r}$ from the median method is:

$$\boldsymbol{r} = \left(\bar{b}_{1\,\frac{|L(1)|+1}{2}}, \bar{b}_{2\,\frac{|L(2)|+1}{2}}, ..., \bar{b}_{n\,\frac{|L(n)|+1}{2}} \right)$$

where $\bar{b}_{ij}$ is the jth value in the list of arranged estimates of alternative i, sorted from least to greatest.

Results

In this section, we present and analyze the results of the experiment by measuring how close the collective ordinal estimates and collective cardinal estimates obtained under each aggregation method are to the respective ground truths. All featured aggregation models are used for the ordinal estimation task; however, only the cardinal aggregation, cardinal and ordinal aggregation, separation-deviation, average, and median methods are tested for the cardinal estimation task since it was not immediately evident how to transform ordinal inputs for use by the OA model and by the traditional voting methods.

Before proceeding, we discuss a useful concept for gauging the goodness of the participant-to-image distributions in the experiments. First, let $G = (V, E)$ be the pairwise-comparison (undirected) graph associated with the set of evaluation inputs. The node set V is comprised of the objects (i.e., images) being evaluated; its edge set E is constructed by drawing an edge (i, j) when at least one of the participants directly evaluates both i and j. The *number of hops* between $i \in V$ and $j \in V$ in G is defined as the length of the shortest path between the two nodes (Hochbaum and Levin 2006b). A maximum number of two hops between all pairs of objects is recommended for obtaining good collective results (Hochbaum and Levin 2010). The robustness of the relative comparison between i and j is expected to deteriorate for higher numbers of hops. As an example application of this concept, when images i and j are evaluated by one person, j and k are evaluated by a different person, and no one evaluates i and k, there is one hop between i and j (and j and k) and two hops between i and k. Table 1 shows the number of hops in the experiment. It shows that displaying more images at a time allows more pairs of alternatives to be evaluated, which tends to increase the connectedness of the pairwise-comparison graphs. Note that the minimum number of hops is always 1.0, since there is at least one comparison of two images completed by each participant.

		Number of times each image was seen											
Problem size		5	10	15	20	25	30	35	40	45	50	55	60
2-image	avg	2.29	1.75	1.60	1.51								
	max	4.00	3.00	2.75	2.00								
3-image	avg	1.74	1.49	1.33	1.23	1.16	1.11						
	max	3.00	2.00	2.00	2.00	2.00	2.00						
5-image	avg	1.48	1.23	1.12	1.06	1.03	1.01	1.01	1.00	1.00	1.00		
	max	2.00	2.00	2.00	2.00	2.00	2.00	1.93	1.71	1.40	1.00		
6-image	avg	1.39	1.15	1.06	1.02	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	max	2.00	2.00	2.00	2.00	1.97	1.71	1.34	1.14	1.05	1.02	1.00	1.00

Table 1: Average number of hops in pairwise comparison graphs calculated for different segments of the data

Performance of aggregation methods

The overall experimental results are shown in Figures 2 and 3. The performance of the aggregation methods on ordinal estimation is quantified via the normalized projected Kemeny-Snell distance between the ground truth and the collective ordinal estimates (aggregate ranking); the performance of the aggregation methods on cardinal estimation is quantified via the normalized Euclidean distance between the ground truth and the collective cardinal estimates (aggregate rating). Hence, the shorter the distance, the better an aggregation model performs. To calculate the number of pairwise reversals between each collective ordinal estimate and the ground truth ranking, the normalized Kemeny-Snell distance should be multiplied by $435 = n(n-1)/2$ (where $n = 30$, the number of images in each problem data set). The difference in the number of dots between each collective cardinal estimate and the ground truth numerical values is obtained by multiplying the normalized Euclidean distance value by the range of the ground truth values, in this case 29.

Figure 2 demonstrates that the performance of the average and median methods in the ordinal estimation tasks is consistently in the bottom-three positions. Overall, the traditional voting rules did not perform as well as consensus aggregation methods. The top-three performances in these tasks were achieved by the two multimodal models, SD and COA, and interestingly by the cardinal-input consensus aggregation model, CA. The latter results suggest that, when the intensities of preference are factored into the aggregation, cardinal inputs can be valuable at performing ordinal estimation tasks. Figure 3 demonstrates that in the cardinal estimation tasks, COA and the average method outperform other cardinal aggregation methods; specifically, COA performed better when each image was evaluated fewer than 30 times, and the average method did better when each image was evaluated by 30 or more times. Overall, COA outperforms other consensus aggregation models in both types of tasks. This suggests that integrating multimodal information (ordinal and cardinal inputs) can help attain more accurate collective estimates.

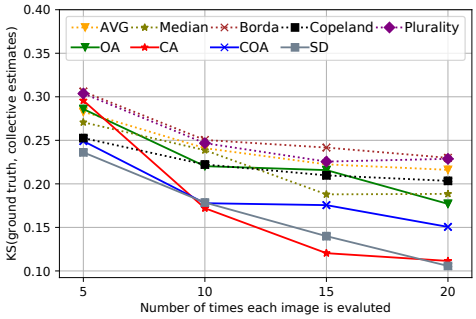
Effect of problem size

Recall that, for each exercise, a participant is looking at 2, 3, 5, or 6 images at a time. As illustrated in Figure 2, showing more images at a time (i.e., a larger problem size) helps attain a collective ordinal estimate that is closer to the ground truth. Specifically, Figure 2(a) shows the distance between the ground truth ranking and the aggregate ranking when two images are shown at a time and Figure 2(d) shows the distance when six images are shown at a time. A plausible reason for this is that, when two images are seen, there are only two possibilities of ordering them: the correctly ordered pair or the inverse ordering of the pair. In contrast, when six images are seen, there are 720 possibilities of ordering them. Although it is very difficult to order the six images perfectly, participants may order a large number of pairs correctly due to the implicit pairwise comparisons in a ranking. Furthermore, a more highly connected graph of comparisons is obtained by exercises where more images are evaluated. As shown in Table 1, with all 300 users, the average number of hops is 1.0 in the 5-image and 6-image estimation tasks, which means every pair of alternatives has been evaluated by at least one person (i.e., the pairwise comparison graph is fully connected). Hence, despite the higher cognitive load of ordering more images, the collective ranking obtained from these larger size problems yielded closer approximations to the ground truth. Problem size did not appear to affect the quality of cardinal estimation to the same degree, due to the fact that numerical estimation was done one image at a time.

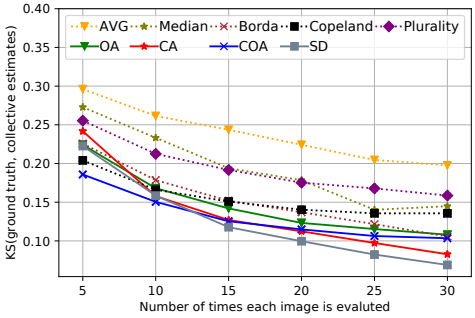
While showing more images at once improves ordinal estimation tasks in our experiment, there may be a maximum on how many objects can be presented at once before the quality of estimates deteriorates. Psychologists termed this phenomenon "choice overload", which suggests that presenting people too many options at a time can result in the depletion of cognitive resources (Reed et al. 2011). Presenting people 6 images to rank at a time provided the highest accuracy in this experiment, however, there is likely a limit not reached in this study on how many images can be efficiently ranked at once. The joint aggregation model was particularly effective at completing the ordinal estimation task as problem size increased.

Efficiency of crowdsourcing

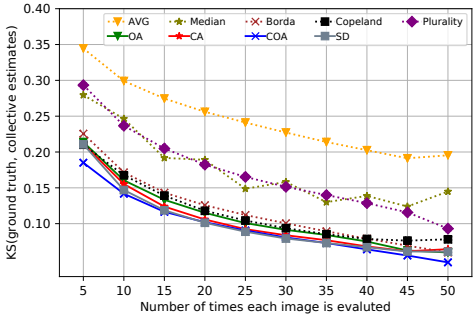
Based on the concept of the wisdom of crowds, the quality of the collective estimates obtained by each aggregation method should improve as the number of times each image is seen increases. To test this hypothesis, we segmented the data based on the number of times each image in the problem set is evaluated, in increments of 5, in order to show the effects of varying crowd size. For each segment, ${}_{2k}C_i$ collective estimates are calculated, where k is the problem size and $i = 1, 2, ..., k$ (i.e., we choose i subgroup of users from total of $2k$ groups and calculate collective estimates for



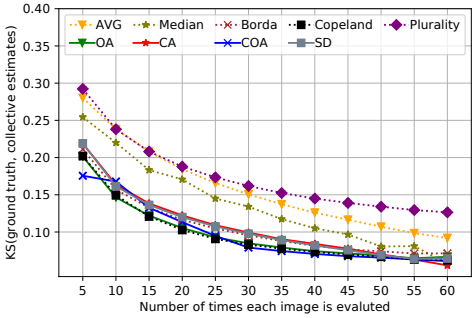
(a) 2-image Ordinal Estimation Task



(b) 3-image Ordinal Estimation Task

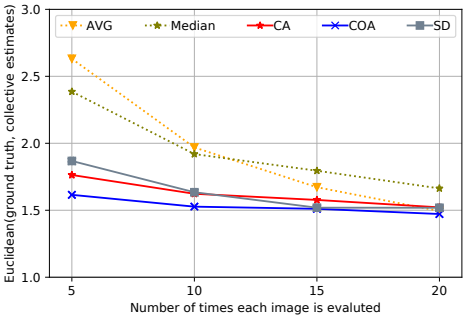


(c) 5-image Ordinal Estimation Task

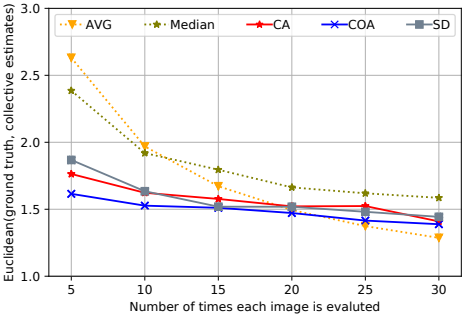


(d) 6-image Ordinal Estimation Task

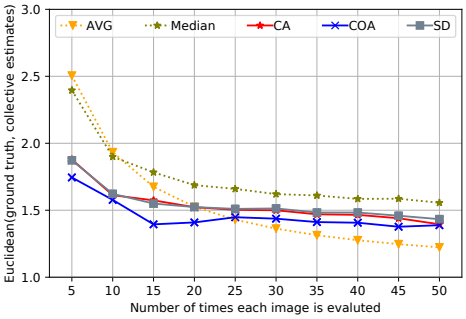
Figure 2: Accuracy of collective ordinal estimation



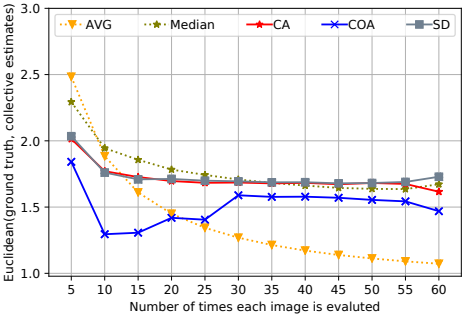
(a) 2-image Cardinal Estimation Task



(b) 3-image Cardinal Estimation Task



(c) 5-image Cardinal Estimation Task



(d) 6-image Cardinal Estimation Task

Figure 3: Accuracy of collective cardinal estimation

each subgroup), their distances to the ground truth are calculated, and the corresponding average is reported. Note that the possible range of the number of times each individual image was evaluated increases with problem size (since increasing the number of images seen by each participant also increases the times each image from the problem set is seen overall). As shown in Figure 2 and 3, the more times each image was evaluated, the more accurate the aggregate estimate generally became for both the ordinal and cardinal estimation tasks. As can be seen in the figures, the improvement of the collective ordinal estimates followed a more steady and consistent trend than the collective cardinal estimates.

In the cardinal estimation task, the average method significantly outperforms other methods when images are evaluated roughly over 30 times. This suggests that in this activity, averaging may be the best way to determine the cardinal estimates when there are sufficient resources available to recruit a large number of participants. Conversely, consensus aggregation models performed significantly better than the average and median methods when each image was evaluated by very few participants. However, their performance did not increase as sharply as the average method for higher numbers of participants. Indeed, in the 5-image cardinal estimation task, there is only a slight increase in performance by the CA, SD, and COA models between the collective estimates obtained when images are viewed 50 times instead of 15 times. This suggests that for these three consensus aggregation models, it may not be worthwhile to add participants beyond a certain point. For example, in the 6-image cardinal estimation task, the COA collective estimate obtained when images were viewed only 10 times was nearly identical to the performance achieved by the average method with three times the number of individual image views; however, in this case the COA performance worsened when more participants viewed the images. While such observations support the increased efficiency of consensus aggregation methods, they also raise a question of how the appropriate crowd size should be determined when eliciting multimodal inputs. In particular, crowdsourcers may consider applying multimodal methods when fewer resources and/or eligible participants are available for performing tasks. An additional future question is to consider integrating the average method with a consensus aggregation model and testing how the resulting multimodal model performs at determining collective estimates in other similar tasks.

Discussion

This section reviews important observations related to the research questions, and limitations of the research.

Benefits of utilizing multimodal information

Our experiment results demonstrated that having multimodal information helps better approximate the ordinal and cardinal ground truths. Because certain modalities may be more useful to different people (e.g., people may provide more correct ordinal estimates but less accurate numerical estimates, or vice versa), having both cardinal and ordinal estimation complements differing strengths. Considering all tested estimations tasks, the cardinal and ordinal aggregation model exhibited the most consistent performance among the tested methods. This supports the notation that eliciting and aggregating multimodal information can help complete the separate types of estimation tasks. Furthermore, we found that aggregating cardinal information may yield more accurate ordinal estimates (using CA, a consensus aggregation model that focuses on the differences between pairwise estimates rather than the actual numerical values).

Displaying more images at a time

Showing more images at a time leads to more accurate collective outcomes. It is reasonable to hypothesize that displaying more images could burden cognitive ability so that the estimating accuracy may deteriorate. However, our experiments did not reach the point at which the collective ordinal estimates markedly deteriorate. We suspect that may occur when 7-9 images (or more) are displayed, based on Millers law (Miller 1956).

Supporting the idea of the wisdom of crowds

Our research empirically supports the concept of the wisdom of crowds. Specifically, the results demonstrated that gathering information from more participants led to more accurate collective estimates under a variety of aggregation methods. That said, optimization-based aggregation methods were able to attain better estimates most efficiently.

Limitations

Our findings are admittedly limited in terms of scope. The main takeaways apply so far only to the featured experiment on dot estimation, and in the future these methods will need to be extended to other crowdsourced activities. To extend these results, it will be useful to evaluate how input elicitation and aggregation methodologies affect prediction activities, grading tasks, and other classification problems. Moreover, no time limit was given in the experiment, which may have allowed participants to manually count the number of dots instead of estimating them. According to (Maddalena et al. 2016), the quality of responses could actually increase when participants are asked to complete a task in a predefined amount of time. Future work will look to limiting completion times in order to test this hypothesis.

Conclusions

This paper investigates how the quality and efficiency of crowdsourced collective estimates can be improved by integrating multiple modalities of input. To the best of our knowledge, this is the first experiment which applied both optimization-based consensus aggregation models and traditional voting-based methods to aggregate estimations from hundreds of people. The main contributions of the paper are showing empirically that (i) given a generalizable benchmark human computation task, collecting and aggregating both ordinal and cardinal information has the potential to improve crowdsourcing results and (ii) all aggregation models tend to perform better as crowd size increases, which aligns with the idea of the wisdom of crowds, but optimization-based models can achieve this most efficiently.

Acknowledgments. The authors thank all participants in this study, which received institutional IRB approval prior to deployment. The PI (the third author) and two students (the first and second authors) gratefully acknowledge funding support from the National Science Foundation (Award 1850355) and the Army Research Office (Award 74113NSII). The first and fourth author also gratefully acknowledge support by the National Science Foundation (Award 1350573 and 1939725).

Data Availability

Experimental data, including the web application, images, and recorded estimates, are available at github.com/ryankemmer/dotsactivity-data or at github.com/adolfoescobedo.

References

Abraham, I.; Alonso, O.; Kandylas, V.; Patel, R.; Shelford, S.; and Slivkins, A. 2014. How many workers to ask? adaptive exploration for collecting high quality labels.

Arrow, K. J. 1951. Social choice and individual values.

Benade, G. 2018. Efficiency and usability of participatory budgeting methods.

Brabham, D. C.; Ribisl, K. M.; Kirchner, T. R.; and Bernhardt, J. M. 2014. Crowdsourcing applications for public health. *Am. J. Prev. Med.* 46(2):179–187.

Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D. 2016. *Handbook of computational social choice*. Cambridge University Press.

Buhrmester, M. D.; Kwang, T. N.; and Gosling, S. D. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6 1:3–5.

Chen, X.; Bennett, P. N.; Collins-Thompson, K.; and Horvitz, E. 2013. Pairwise ranking aggregation in a crowd-sourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, 193–202. New York, NY, USA: Association for Computing Machinery.

Chittilappilly, A.; Chen, L.; and Amer-Yahia, S. 2016. A survey of general-purpose crowdsourcing techniques. *IEEE Trans Knowl Data Eng* 28:1–1.

Chung, J. J. Y.; Song, J. Y.; Kutty, S.; Hong, S. R.; Kim, J.; and Lasecki, W. S. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proc. ACM Hum.-Comput. Interact.* 3(CSCW).

Cook, W. D., and Kress, M. 1985. Ordinal ranking with intensity of preference. *Management Science* 31(1):26–32.

Cook, W. D. 2006. Distance-based and ad hoc consensus models in ordinal preference ranking. *Eur. J. Oper. Res.* 172(2):369–385.

Davis-Stober, C.; Budescu, D.; Broomell, S.; and Dana, J. 2015. The composition of optimally wise crowds. *Decision Analysis* 12.

Dwork, C.; Kumar, R.; Naor, M.; and Sivakumar, D. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on the World Wide Web*, 613–622. New York, NY, USA: ACM.

Eickhoff, C. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, 162–170. New York, NY, USA: Association for Computing Machinery.

Escobedo, A. R.; Hochbaum, D. S.; and Moreno-Centeno, E. 2020. An axiomatic distance methodology for aggregating multimodal evaluations.

Fishbain, B., and Moreno-Centeno, E. 2016. Self calibrated wireless distributed environmental sensory networks. *Scientific Reports* 6:24382.

Galton, F. 1907. Vox populi.

Goldstein, D. G.; McAfee, R. P.; and Suri, S. 2014. The wisdom of smaller, smarter crowds. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, 471–488. New York, NY, USA: Association for Computing Machinery.

Herzog, S. M., and Hertwig, R. 2009. The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science* 20(2):231–237. PMID: 19170937.

Herzog, S., and Hertwig, R. 2011. The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition. *Judgment and Decision Making* 6.

Hochbaum, D., and Levin, A. 2006a. Methodologies and algorithms for group-rankings decision. *Management Science* 52:1394–1408.

Hochbaum, D. S., and Levin, A. 2006b. The k-allocation problem and its variants. In *International Workshop on Approximation and Online Algorithms*, 253–264. Springer.

Hochbaum, D. S., and Levin, A. 2010. How to allocate review tasks for robust ranking. *Acta informatica* 47(5-6):325–345.

Hochbaum, D. S. 2010. The separation, and separation-deviation methodology for group decision making and aggregate ranking. In *Risk and Optimization in an Uncertain World*. INFORMS. 116–141.

Horton, J. J. 2010. The dot-guessing game: A 'fruit fly' for human computation research.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19. New York, NY, USA: Association for Computing Machinery.

Kemeny, J. G., and Snell, L. J. 1962. Preference ranking: An axiomatic approach. In *Mathematical Models in Social Science*. Boston: Ginn. 9–23.

Li, K.; Zhang, X.; and Li, G. 2018. A rating-ranking method for crowdsourced top-k computation. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, 975–990. New York, NY, USA: ACM.

Maddalena, E.; Basaldella, M.; De Nart, D.; Degl'Innocenti, D.; Mizzaro, S.; and Demartini, G. 2016. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Fourth AAAI conference on human computation and crowdsourcing*.

Mao, A.; Procaccia, A. D.; and Chen, Y. 2013. Better human computation through principled voting. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, 1142–1148. AAAI Press.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* 63(2):81–97.

Moreno-Centeno, E., and Escobedo, A. R. 2016. Axiomatic aggregation of incomplete rankings. *IIE Transactions* 48(6):475–488.

Noothigattu, R.; Gaikwad, S. N. S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. D. 2017. A voting-based system for ethical decision making.

Ovadia, S. 2004. Ratings and rankings: Reconsidering the structure of values and their measurement. *Int. J. Soc* 7:403–414.

Quoc Viet Hung, N.; Tam, N. T.; Tran, L. N.; and Aberer, K. 2013. An evaluation of aggregation techniques in crowdsourcing. In Lin, X.; Manolopoulos, Y.; Srivastava, D.; and Huang, G., eds., *Web Information Systems Engineering – WISE 2013*, 1–15. Berlin, Heidelberg: Springer Berlin Heidelberg.

Rankin, W. L., and Grube, J. W. 1980. A comparison of ranking and rating procedures for value system measurement. 233–246. European Journal of Social Psychology.

Reed, D.; DiGennaro Reed, F.; Chok, J.; and Brozyna, G. 2011. The "tyranny of choice": Choice overload as a possible instance of effort discounting. *The Psychological record* 61:547–560.

Simmons, J. P.; Nelson, L. D.; Galak, J.; and Frederick, S. 2011. Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *J. Consum. Res* 38(1):1–15.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing annotations for visual object detection. 40–46.

Surowiecki, J. 2005. *The Wisdom of Crowds*. Anchor.

Thurstone, L. L. 1927. The method of paired comparisons for social values. *J. Abnorm. Psychol* 21(4):384.

Werbin-Ofir, H.; Dery, L. N.; and Shmueli, E. 2019. Beyond majority: Label ranking ensembles based on voting rules. *Expert Syst. Appl.* 136:50–61.

Wilber, M. J.; Kwak, I. S.; and Belongie, S. J. 2014. Cost-effective hits for relative similarity comparisons.

Yoo, Y., and Escobedo, A. R. 2020. A new binary programming formulation and social choice propertyfor expediting the solution to Kemeny rank aggregation. *Under review*. [Available at `optimization-online.org/DB_HTML/2020/08/7958.html`].

Yoo, Y.; Escobedo, A. R.; and Skolfield, J. K. 2020. A new correlation coefficient for comparing and aggregating non-strict and incomplete rankings. *European Journal of Operational Research* 285(3):1025–1041.