The Wisdom of Crowds in Rank Ordering Problems

Brent Miller (brentm@uci.edu)
Pernille Hemmer (phemmer@uci.edu)
Mark Steyvers (mark.steyvers@uci.edu)
Michael D. Lee (mdlee@uci.edu)

Department of Cognitive Sciences
3151 Social Science Plaza
Irvine, CA 92697-5100

Abstract

When averaging the estimates of individuals, the aggregate can often come surprisingly close to the true answer. We are interested in extending this "wisdom of crowds" phenomenon to more complex situations where a simple strategy like taking the mode or mean of responses is inappropriate, or might lead to bad predictions. We report the performance of individuals in a series of ordering tasks, where the goal is to reconstruct from memory the order of time-based events, or the magnitude of physical properties. We introduce a Bayesian version of a Thurstonian model that aggregates orderings across individuals, and compare it to heuristic aggregation techniques inspired by existing models of social choice and voting theory. The Bayesian model performs as well as the heuristics in reconstructing the true ordering, and has the advantage of being well calibrated, in the sense that it gives more confident responses the closer it is to the truth.

Keywords: Bayesian Modeling; Rank Ordering; Consensus; Wisdom of Crowds; Rank aggregation.

Introduction

When Galton first surveyed English fair-goers in 1906, it was a novel curiosity that their estimates of the dressed weight of an ox, when averaged, closely approximated the true weight (Galton, 1907). Subsequently, many demonstrations have shown that aggregating the judgments of a number of individuals often results in an estimate that is close to the true answer. This phenomenon has come to be known as the "wisdom of crowds" (Surowiecki, 2004). The wisdom of crowds idea is currently used in several real-world applications, such as prediction markets (Dani et al., 2006), spam filtering, and the prediction of consumer preferences through collaborative filtering.

Many wisdom of crowds demonstrations have involved situations where a single numerical quantity needs to be estimated. In these cases, a robust estimate of the central tendency of individual estimates can be an effective aggregation method (Yaniv, 1997). Other situations have involved recovering the answers to multiple choice questions. For example, on the game show "Who Wants to be A Millionaire", contestants are given the opportunity to ask all members of the audience to answer a multiple choice question. In this case, an aggregation method based on the modal response can be quite effective. Over several seasons of the show, the modal response of the audience corresponded to the correct answer 91% of the time. More sophisticated approaches have been developed, such as Cultural Consensus Theory (e.g., Romney, Batchelder, Weller, 1987), that additionally take differences across individuals and items into account when aggregating multiple choice answers.

In this paper, we extend the wisdom of crowds idea to the more complex problem of rank ordering. Is it possible to recover the correct order of events or physical properties from a large number of independent individual responses? How confident can we be that these aggregations represent the ground truth?

Aggregating rank order data is not a new problem. In social choice theory, a number of systems have been developed for aggregating rank order preferences for groups (Marden, 1995). Preferential voting systems, where voters explicitly rank order their candidate preferences, are designed to pick one or several candidates out of a field of many. These systems, such as the Borda count, perform well in aggregating the individuals' rank order data, but with an inherent bias towards determining the top members of the list.[1] However, as voting is a means for expressing individual preferences, there is no ground truth. The goal for these systems is to determine an aggregate of preferences that is in some sense "fair" to all members of the group.

Relatively little research has been done on the rank order aggregation problem with the goal of approximating a known ground truth. In follow-ups to Galton's work, Gordon (1924) and Bruce (1935) tested a large number of individuals in psychophysical ordering tasks. They found that the group estimate approximates the ground truth better as the size of the group increases. Interestingly, these authors used the Borda count voting method (without making this connection to voting theory explicit in their work) to aggregate the rank orderings of individuals. Romney et al. (1987) also developed an informal aggregation model for rank order data based on Cultural Consensus Theory, using factor analysis of the covariance structure of rank order judgments. With this, they were able to partially recover the correct order of 34 causes of death in

[1] This is necessary to satisfy the *Condorcet Criterion*, which requires that a top ranked candidate selected by a voting system should be a candidate who has more votes when compared to every other voter on the ballot (Shepsle & Bonchek, 1997)

the US on the basis of the individual orderings of 36 subjects.

We present empirical and theoretical research on the wisdom of crowds phenomenon for rank order aggregation. We conduct an empirical study where people are asked to rank order the occurrence of events (e.g., US presidents by term of office[2]) or the magnitude of some physical property (e.g., rivers by length). Most importantly, no communication between people is allowed for these tasks, and therefore the aggregation method operates on the data produced by independent decision-makers.

Importantly, for all of the problems there is a known ground truth. The ground truth might only be partially known to the tested individuals. If different individuals have knowledge of different parts of the ordering problems, aggregation across individuals can yield a group answer that comes closer to the ground truth than any of the individuals in the group. For example, if some individuals know that the Congo is longer than the Parana River, and other individuals know that the Parana River is longer than the Mekong River, aggregation might lead to the correct overall ordering (i.e., Congo > Parana > Mekong). Therefore, for the wisdom of crowd phenomenon to work, the errors in semantic memory need to have some degree of independence. If all individuals have access to the same knowledge, there will be no advantage to aggregating their answers.

We compare several heuristic computational approaches—based on voting theory and existing models of social choice—that analyze the individual judgments and provide a single answer as output, which can be compared to the ground truth. We refer to these synthesized answers as the "group" answers because they capture the collective wisdom of the group, even though no communication between group members occurred.

We also develop a probabilistic model based on a Thurstonian approach that represents items as distributions on an interval dimension. We make inferences about the parameters of the model using Markov chain Monte Carlo (MCMC). The advantage of MCMC estimation procedure is that it gives a probability distribution over group orderings, and we can therefore assess the likelihood of any particular group ordering. We use this likelihood as a confidence measure to test whether the model is calibrated, in the sense that the group answers with high confidence are close to the ground truth.

Experiment

Method

Participants were 78 undergraduate students at the University of California, Irvine. The experiment was composed of 20 questions (3 were excluded from analysis; one because participants misunderstood the question, one because of the lack of a proper ground truth, and the last for consistency as it only included 5 elements for ordering, whereas all the others included 10). The remaining questions involved general knowledge regarding: population statistics (4 questions), geography (3 questions), dates, such as release dates for movies and books (7 questions), U.S. Presidents, material hardness, the 10 Commandments, and the first 10 Amendments of the U.S. Constitution

All questions had a ground truth obtained from Pocket world in figures and various online sources. An interactive interface was presented on a computer screen. Participants were instructed to order the presented items (e.g., "Order these books by their first release date, earliest to most recent"), and responded by dragging the individual items on the screen using the computer mouse, and "snapping" the item into the desired location in the ordering. Once participants were satisfied with their response they clicked on the submit button. They were prompted to confirm that they wished to proceed before being presented with the next question. Once their response was submitted it was not possible to return to that question. The questions were presented in a fixed order. Half the participants received the forward ordering of questions, the other half received the backwards ordering of questions. The initial ordering of the 10 items within a question was randomized across all questions and all participants.

Results

We first evaluated participants' responses based on whether or not they reconstructed the correct ordering. Table 1 shows the proportion of individuals who got the ordering exactly right (PC) for each of the ordering task questions. On average, about one percent of participants recreated the correct rank ordering perfectly. We also analyzed the performance of participants with a more fine-grained measure, using Kendall's τ distance. This distance metric is used to count the number of pair-wise disagreements between the reconstructed and correct ordering. The larger the distance, the more dissimilar the two orderings are.

Table 1: *Participant performance statistics*.

		Percentiles of τ				
Problem	PC	25	50	75	90	100
books	0.000	15	10	8	5	3
city population europe	0.000	19	15	12	10	7
city population us	0.000	20	14	11	8	6
city population world	0.000	23	18	15	12	5
country landmass	0.000	12	9	7	5	2
country population	0.000	17	15	11	9	4
hardness	0.000	18	15	12	11	7
holidays	0.051	12	8	5	3	0
movies releasedate	0.013	9	6	4	2	0
oscar bestmovies	0.013	14	10	6	4	0
oscar movies	0.000	16	10	5	2	1
presidents	0.064	10	7	3	1	0
rivers	0.000	19	15	13	11	3
states westeast	0.026	10	6	3	1	0
superbowl	0.000	24	17	14	11	6
ten ammendments	0.013	19	13	10	4	0
ten commandments	0.000	23	17	11	7	1
AVERAGE	0.011	16.5	12.1	8.8	6.2	2.6

[2] The ordering of US Presidents has been studied before in the context of memory research by Healy, Havas, and Parker (2000).

Values of τ range from: $0 \leq \tau \leq N(N-1)/2$, where N is the number of items in the order (10 for all of our questions). A value of zero means the ordering is exactly right, and a value of one means that the ordering is correct except for two neighboring items being transposed, and so on up to the maximum possible value of 45.

Table 1 shows the distribution of τ values over the ranked population of participants for each of the 17 sorting task questions, in terms of values at the 25th, 50th, 75th, 90th and 100th percentiles. For six of the questions, one or more participants get the ordering exactly right, as indicated by a τ of 0 for the 100th percentile. The best individuals on each question achieve good performance, and solve the problem exactly, or are within a few pair transposes, for most questions. As this is a prior knowledge task, it is interesting to note the best performance overall was achieved on the *Presidents*, *States from west to east*, *Oscar movies*, and *Movie release dates* tasks. These four questions relate to educational and cultural knowledge that seems most likely to be shared by our undergraduate subjects.

Modeling

We evaluated a number of heuristic aggregation models and compared the performance of these methods against a probabilistic model based on a Thurstonian approach. For each model, the set of orderings from individuals is analyzed in order to create a single group ordering, which is then compared to the ground truth.

Heuristic Models

We tested four heuristic aggregation models. The simplest heuristic, based on the mode, has been used since the earliest rank order experiments (Lorge et al. 1957). For this heuristic, the group answer is based on the most frequently occurring sequence of all observed sequences. In cases where several different sequences correspond to the mode, a randomly chosen modal sequence was picked.

The second method, which we refer to as the "greedy count", counts the number of participants responses for each item in each position. The item and the position with the largest agreement among participant is selected first. The selection of items then proceeds in a greedy algorithm fashion, making sure that each item and position is not already filled.

The third method takes the group answer as the participant ranking that is "closest", as determined by a distance measurement metric, to the rankings of all participants. This is known as the Kemeny-Young method (e.g., Marden, 1997). It is implemented here by finding the participant ordering that has the smallest distance, measured by the sum of Kendall's τ's between strings, to the orderings of all other participants. Note that we restrict ourselves to finding a ranking from the *existing* set of participants' responses. This method can be extended to find *any* arbitrary rank order that is closest to the "middle" of observed rankings, but that approach suffers from well-known computational complexity problems.

The fourth method uses the Borda count method, a widely used technique from voting theory. In preferential voting systems, voters express their candidate choices in terms of an ordering of all ballot candidates. In the Borda count method, weighted counts are assigned such that the first choice "candidate" receives a count of N (where N is the number of candidates), the second choice candidate receives a count of N-1, and so on. These counts are summed across candidates and the candidate with the highest count is considered the "most preferred". Here, we use the Borda count to create an ordering over all items by ordering the Borda counts.

Table 2 reports the performance of all of the aggregation models. For each, we checked whether the inferred group order is correct (C) and measured Kendall's τ. We also report in the Rank column the percentage of participants who perform worse or the same as the group answer, as

Table 2: *Performance of the four heuristic models and the Thurstonian model*

	Kemeny-Young			Thurstonian Model			Borda Counts			Greedy Count			Mode		
Problem	C	τ	Rank	C	τ	Rank	C	τ	Rank	C	τ	Rank	C	τ	Rank
books	0	4	96	0	6	88	0	7	82	0	7	82	0	12	40
city population europe	0	11	81	0	11	81	0	11	81	0	13	69	0	17	42
city population us	0	10	87	0	11	79	0	12	67	0	9	90	0	16	45
city population world	0	18	59	0	16	73	0	15	77	0	16	73	0	19	44
country landmass	0	7	76	0	5	95	0	5	95	0	5	95	0	7	76
country population	0	11	82	0	11	82	0	11	82	0	13	67	0	15	53
hardness	0	11	91	0	11	91	0	11	91	0	18	31	0	15	46
holidays	0	5	77	0	4	78	0	4	78	0	4	78	1	0	100
movies releasedate	0	2	95	0	2	95	0	2	95	0	2	95	0	2	95
oscar bestmovies	0	3	97	0	4	90	0	3	97	0	5	90	0	3	97
oscar movies	0	2	96	0	1	100	0	2	96	0	3	88	0	2	96
presidents	0	1	94	0	2	87	0	3	79	0	1	94	1	0	100
rivers	0	11	91	0	12	86	0	11	91	0	13	77	0	16	42
states westeast	0	1	97	0	2	88	0	3	78	0	1	97	0	1	97
superbowl	0	10	96	0	12	88	0	10	96	0	15	71	0	19	40
ten ammendments	0	2	97	0	4	95	0	5	90	0	4	95	0	4	95
ten commandments	0	11	82	0	11	82	0	12	74	0	12	74	0	17	51
AVERAGE	0.00	7.03	87.9	0.00	7.35	87.0	0.00	7.47	85.3	0.00	8.29	80.3	0.12	9.67	68.2

measured by τ. With the Rank statistic, we can verify the wisdom of crowds effect. In an ideal model, the group heuristic should perform as well as or better than all of the individuals in the group. Table 2 shows the results separately for each problem, and averaged across all the problems.

These results show that the mode heuristic leads to the worst performance overall in rank. On average, the mode is as good or better of an estimate than 68% of participants. This means that 32% of participants came up with better solutions individually. This is not surprising, since, with an ordering of 10 items, it is easily possible that only a few participants will agree on the ordering of items. The difficulty in inferring the mode makes it an unreliable method for constructing a group answer. This problem will be exacerbated for orderings involving more than 10 items, as the number of possible orderings grows combinatorially. The greedy count heuristic performs better than the mode overall, but it does not lead to the correct answer for any individual problem.

The Borda count and Kemeny-Young methods perform relatively well in terms of Kendall's τ and overall rank performance. On average, these methods perform with ranks of 85% and 88% respectively, indicating that the group answers from these methods score amongst the best individuals, although 10% of individuals still perform better.

A Thurstonian Model

Despite comparable statistical performances, the heuristic aggregation models create no explicit representation of each individual's working knowledge. Therefore, even though the methods can aggregate the individual pieces of knowledge across individuals, they cannot explain *why* individuals rank the items in a particular way, or how much confidence should be placed in the overall group ranking. To address this potential weakness, we develop a simple probabilistic model based on the seminal Thurstonian approach. Although the Thurstonian approach has often been used to analyze preference rankings (see Marden, 1997 for an overview), it has not been applied, as far as we are aware, to ordering problems where there is a ground truth.

In the Thurstonian approach, the overall item knowledge for the group is represented explicitly as a set of coordinates on an interval dimension. The interval representation is justifiable given that all the problems in our study involve one-dimensional concepts (e.g., the relative timing of events, or the lengths of items). Specifically, each item is represented as a value μ_i along this dimension, where $i \in \{1, \ldots, N\}$. Each individual is assumed to have access to the group-level information. We assume, however, that individuals do not have precise knowledge about the exact location of each item. We model each individual's location of the item by a single sample from a distribution, centered on the item's group location. We represent the uncertainty associated with this value, μ_i, with a Normal distribution, $N(\mu_i, \sigma_i)$. In a fully specified Thurstonian model, once an individual draws samples for each item, the ordering for that individual is based on the ordering of the samples. Figure 1 shows an example of the group-level information for six items, A to G. A particular individual might sample values from these distributions such that some items are ranked correctly, but other items are transposed. In Figure 1, there is a larger degree of uncertainty for item C, making it likely that item C is placed incorrectly in the ordering.

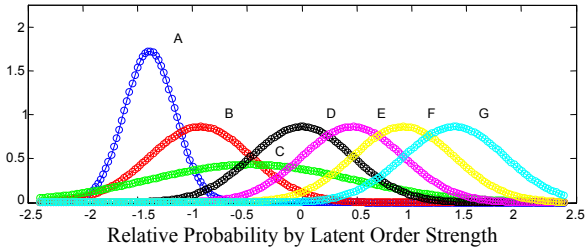

Figure 1. Example of group-level information for six items.

We apply Bayesian estimation techniques to infer the group representation from the individual orderings. Bayesian methods have been applied to Thurstonian models before (Yao, & Böckenholt, 1999), but here we present a simplified version of the Thurstonian model that facilitates more efficient Bayesian inference.

In the simplified model, we do not attempt to explain the particular orderings for each individual, but rather the pairwise orderings across all individuals. The data for this model consist of a N x N count matrix R, where $R(i, j)$ contains the number of participants who ordered item i later than item j. For example, Figure 2 shows the matrix for the *Presidents* question with the Presidents in the correct order. Note that nearly all of the 78 participants correctly place George Washington earlier than any of the other Presidents, but that Dwight D. Eisenhower, who should be ranked last, is often placed earlier than other Presidents. The pairwise data therefore indicate some uncertainty about the ranking of Eisenhower relative to other Presidents.

In our model, when determining the relative order of two items i and j, a person samples a value from item i, $x_i \sim N(\mu_i, \sigma_i)$, and also a value from item j, $x_j \sim N(\mu_j, \sigma_j)$. These values are then compared to each other and item i is ranked above j whenever $x_i > x_j$. Let θ_{ij} represent the probability of the outcome $x_i > x_j$. This probability can be

		A	B	C	D	E	F	G	H	I	J
George Washington	A	0	0	2	1	1	1	2	1	1	2
John Adams	B	78	0	29	10	14	7	6	5	4	5
Thomas Jefferson	C	76	49	0	10	10	1	6	2	3	2
James Monroe	D	77	68	68	0	45	15	18	14	13	15
Andrew Jackson	E	77	64	68	33	0	11	9	10	9	11
Theodore Roosevelt	F	77	71	77	63	67	0	37	18	24	23
Woodrow Wilson	G	76	72	72	60	69	41	0	22	29	27
Franklin D. Roosevelt	H	77	73	76	64	68	60	56	0	40	34
Harry S. Truman	I	77	74	75	65	69	54	49	38	0	38
Dwight D. Eisenhower	J	76	73	76	63	67	55	51	44	40	0

Figure 2. Count matrix R for the 'Presidents' question.

determined exactly:

$$\theta_{ij} = p(x_i > x_j) = \Phi\left((\mu_i - \mu_j)/\sqrt{\sigma_i^2 + \sigma_j^2}\right), \quad (1)$$

where Φ is the cumulative normal distribution. This sampling process is repeated for each individual and all item pairs. Therefore, the number of times that item i is ranked before item j, across all individuals, is based on the binomial distribution:

$$R_{ij} \sim B(\theta_{ij}, K), \quad (2)$$

where K is the number of individuals.

In this probabilistic model μ_i and σ_i are the latent variables that can be estimated on the basis of the observed data R.[3] We applied MCMC techniques to estimate the latent parameters using a sequence of Metropolis Hasting steps. In order to prevent a drift in the items during estimation (as there is no natural zero point), we fixed the minimum of μ_i to 0 and the maximum of μ_i to 1. We ran 20 chains with a burn-in of 200 iterations. From each chain, we drew 20 samples with an interval of 10 iterations. In total, we collected 400 samples. To construct a single group answer, we analyzed the ordering of the items according to μ_i, separately for each sample, and then picked the mode of this distribution. This corresponds to the most likely order in the distribution over orders inferred by the model.

The result of this Thurstonian model is shown in Table 2. The model performs approximately as well as the Borda count method, but not quite as well as the Kemeny-Young method. The model does not recover the exact answer for any of the 17 problems, based on the knowledge provided by the current 78 participants. It is possible that a larger sample size is needed in order to achieve perfect reconstructions of the ground truth.

Visualization of Group Knowledge One advantage of the Thurstonian approach is that it allows a visualization of group knowledge not only in terms of the order of items, but also in terms of the uncertainty associated with each item on the interval scale. Figure 3 shows the inferred distributions for four problems where the model performed relatively well. The crosses correspond to the mean of μ_i across all samples, and the error bars represent the standard deviations σ_i based on a geometric average across all samples.

These visualizations are intuitive, and show how some items are confused with others in the group population. For instance, nearly all participants were able to identify George Washington as the first President of the U.S., but many confused later Presidents whose terms occurred close to each other. Likewise, there was a large agreement on the proper placement of the right to bear arms in the amendments question — this amendment is often popularly referred to as "the second amendment".

[3] Because of the simplified nature of the model, there is no need to explicitly estimate the particular draws x. These have been integrated out of the model by virtue of Equation (1)

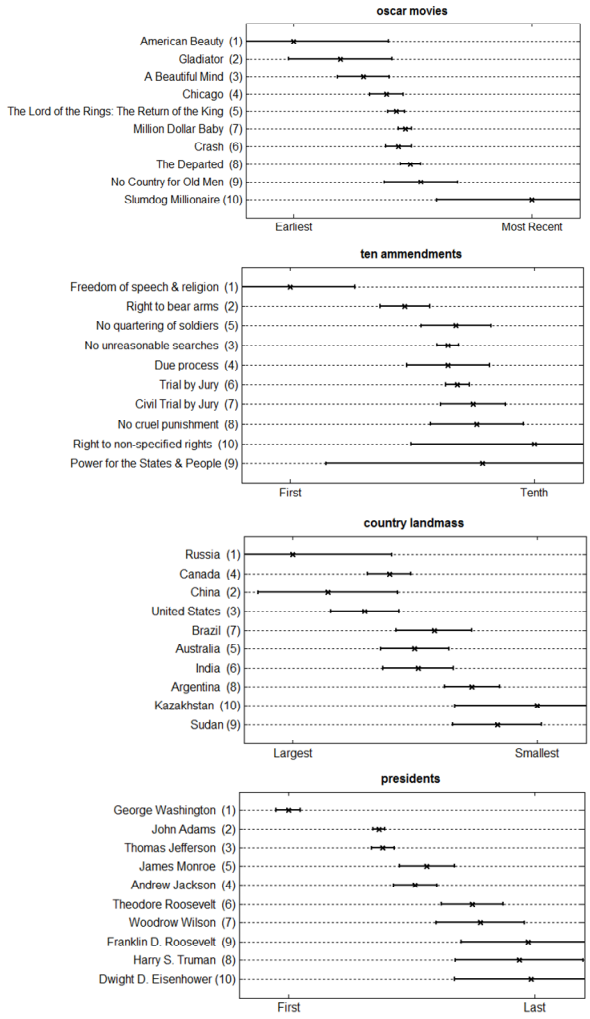

Figure 3. Sample Thurstonian inferred distributions. The actual order is the ground truth ordering, while the numbers in parentheses show the group answer.

Model Calibration Since the probabilistic model is estimated with MCMC techniques, we derive a posterior distribution over all group orderings, from which we select the mode as the best group answer. Because of this, we can also assess the posterior probability of this group answer. This probability has a natural interpretation as the model's measure of confidence. If the distribution over orders is very peaked, most posterior probability is concentrated on the modal answer, indicating a high confidence. If, on the other hand, the model is uncertain about any of the orderings, a low posterior probability, and therefore a low confidence, is given to the modal answer. We can then use this confidence measure to assess to what extent the model is calibrated. That is, we can ask: do confident answers come close to the ground truth?

Figure 4 shows an ordering of the problems according to their confidence values (i.e., the posterior probability of the modal answer). The right panel shows the Kendall τ distance between the group answer and the true answer. The correlation between confidence and Kendall τ is -.63,

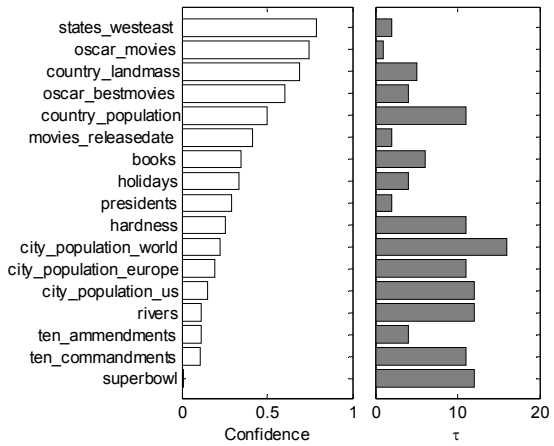

Figure 4. The relation between the confidence in the group answer and the Kendall τ distance of the group answer to the true answer.

showing the expected relationship: high confidence responses are associated with orderings that are closest to the correct ordering. Calibration is important because, in practical situations, the ground truth is not available and a decision maker need to know how confident to be in the aggregated group answer.

Conclusion

We have presented four heuristic aggregation approaches, as well as a Thurstonian approach, for the problem of aggregating rank orders to uncover a ground truth. The model comparison showed that the mode is not a reliable approach for extracting the ground truth, because few individuals agree on the same ordering. We expect that in larger ordering tasks, involving more than 10 items, there might be no individuals that agree with any other on the item ordering. The other heuristic methods, such as the greedy count and the Borda count, analyze the orderings locally by counting the number of times items each occur at each position. This strategy seems to overcome some of the problems with using the mode. The Kemeny-Young method extracted a group answer by finding an existing answer in the data that had the smallest combined distance to all other answers, as measured by Kendall's τ. This result suggests that the idea of finding "prototypical" orderings can lead to effective group answers.

We also presented a Bayesian model based on the classic Thurstonian approach. While this model did not outperform the heuristic models, it did perform well, and has some advantages over the heuristic models. The Bayesian model not only extracts a group ordering, but also a representation of the uncertainty associated with the ordering. This can be visualized to gain insight into mental representations and processes. The MCMC estimation procedure used for the Bayesian model leads naturally to a distribution over orderings. This distribution can be used to measure the confidence in any particular group answer. We found that this confidence relates to how close the group answer is to the true answer. Additionally, although not explored here, the Bayesian approach potentially offers advantages over heuristic approaches because the probabilistic model can be easily expanded with additional sources of knowledge, such as confidence judgments from participants and background knowledge about the items.

References

Bruce, R. (1935). Group Judgments in the Fields of Lifted Weights and Visual Discrimination. *The Journal of Psychology*, 1, 117-121.

Dani, V., Madani, O., Pennock, D.M., Sanghai, S.K., & Galebach, B. (2006). An Empirical Comparison of Algorithms for Aggregating Expert Predictions. In Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI).

Galton, F. (1907). Vox Populi. *Nature*, 75, 450-451.

Gordon, K. (1924). Group Judgments in the Field of Lifted Weights. *Journal of Experimental Psychology*, 7, 398-400.

Healy, A. F., Havas, D. A., Parker, J. T. (2000). Comparing Serial Position Effects in Semantic and Episodic Memory Using Reconstruction of Order Tasks. *Journal of memory and language*, 42, 147-167.

Lorge, I, Fox, D., Davitz, J., Brenner, M., (1957). A Survey of Studies Contrasting the Quality of Group Performance and Individual Performance. *Psychological Bulletin*, 55, 337-372.

Romney, K. A., Batchelder, W. H., Weller, S. C. (1987). Recent Applications of Cultural Consensus Theory. *American Behavioral Scientist*, 31, 163-177.

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. New York, NY: Chapman & Hall USA.

Shepsle, K. A., Bonchek, M. S. (1997*). Analyzing Politics*. New York, NY: Doubleday

Surowiecki, J. (2004*). The Wisdom of Crowds*. New York, NY: W. W. Norton & Company, Inc.

Yaniv, I. (1997). Weighting and trimming: Heuristics for Aggregating Judgments under Uncertainty. *Organizational behavior and human decision processes, 69*(3), 237-249.

Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology, 52*, 79–92.