A hierarchical Bayesian model for improving wisdom of the crowd aggregation of quantities with large between-informant variability

Saiwing Yeung (saiwing.yeung@gmail.com)
Institute of Education, Beijing Institute of Technology, China

Abstract

The wisdom of the crowd technique has been shown to be very effective in producing judgments more accurate than those of individuals. However, its performance in situations in which the intended estimates would involve responses of greatly differing magnitudes is less well understood. We first carried out an experiment to elicit people's estimates in one such domain, populations of U.S. metropolitan areas. Results indicated that there were indeed vast between-subjects differences in magnitudes of responses. We then proposed a hierarchical Bayesian model that incorporates different respondents' biases in terms of the overall magnitudes of their answers and the amount of individual uncertainties. We implemented three variations of this model with different ways of instantiating the individual differences in overall magnitude. Estimates produced by the variation that accounts for the stochasticities in response magnitude outperformed those based on standard wisdom of the crowd aggregation methods and other variations.

Keywords: wisdom of the crowd; graphical model; hierarchical Bayesian model; human judgments; individual differences.

Introduction

The wisdom of the crowd (WoC) technique involves aggregating decisions or estimates made by a group of people. Much research has found that the crowd as a whole can produce estimates that are much more accurate than those by a random informant (Surowiecki, 2005). However, most research focused on types of quantities that are naturally bounded. For example, if the targets of estimation were probabilities of events, all responses would need to be between 0 and 1. This restriction constrains the plausible range of responses and could, as a result, potentially help produce more accurate estimates. Other similarly naturally bounded quantities, some to a lesser degree, include year of events, temperature of cities, etc. In contrast, many real life estimation problems involve values that are not naturally bounded, such that estimates given by different informants could vary by multiple orders of magnitude.

Previous studies have found that when applied to quantities that are not naturally bounded, traditional WoC aggregation methods, such as the mean or median of a crowd's estimates, yield relatively smaller improvement, compared to those that are naturally bounded. For example, Yeung (2013) reported that neither mean nor median improved confidence interval estimates for questions without natural bounds, while they did improve those with natural bounds. Rauhut and Lorenz (2011) also reported that averaging an individual's multiple responses to the same questions did not improve estimates for general numerical questions, in contrast to similar previous research using questions about percentage values (Vul & Pashler, 2008).

Estimates about quantities without natural bounds are commonly encountered because many naturally occurring quantities can be described by distributions without a natural maximum and are severely right skewed. For example, Gibrat's law suggested that the distribution of populations of cities follows a log-normal distribution (Eeckhout, 2004). Other distributions with similar characteristics include power-law, Pareto, and exponential distributions. They naturally occur in many different contexts, including income and wealth, number of friends, waiting time, time till failure, etc. (Barabási, 2005). How to best aggregate these quantities in a WoC context is not very well understood. In the present research we demonstrate a hierarchical Bayesian approach to the problem.

Hierarchical Bayesian models formally express the relationships between psychological constructs, stimuli, and observations. They produce quantitative predictions that can be compared with empirical data, providing a way to test psychological theories encapsulated in the models (Lee, 2011).

In this paper our main objectives are to improve the WoC estimates for quantities without natural bounds using such models, and to investigate the psychological assumptions on which these models rely. We first carried out an empirical experiment to obtain the data on which we will base our analyses. Standard WoC procedures will be applied and their performances will be evaluated. We will then propose and implement a family of computational models that is based on assumptions made about the structure of people's responses. We will then compare the performance of these models against those of the standard WoC methods. Finally we will discuss the implication of our findings.

The experiment

We recruited 101 participants from Amazon Mechanical Turk. Workers were required to be 18 years or older, be residing in the U.S., and have a lifetime acceptance rate on MTurk of over 95%. Each participant was paid US$0.40.

We first reminded participants to not use any external resources during the experiment. We then asked participants to rate the level of their knowledge about geography and population on a 7-point scale (from "Very Good" to "Very Poor").

The participants then completed a set of trivia questions on U.S. geography taken from the experiment in Moore and Healy (2008). We used all nine geography questions there that were about the U.S. Out of those, three were classified by Moore and Healy as easy, four medium, and one hard. Two of the questions were changed slightly to make them more difficult in order to increase the discriminatory power about the participants' knowledge.

The participants would then proceed to the main part of the experiment. Here they were asked to make estimates about the population of 20 U.S. metropolitan areas (the full list can be found in Figure 2). The definition of metropolitan areas were defined for them in the instruction: “a metropolitan area refers to a densely populated urban core and its less-populated surrounding territories”. We selected every three metro areas from the list of U.S. metropolitan areas in Wikipedia[1]. That is, we use the top-ranked (one with the highest population) metro area, the 4th, the 7th, and so on, with the last one being the 58th ranked one. Each metro area was specified using their Metropolitan Statistical Areas name (e.g. “New York-Newark-Jersey City metropolitan area”) and was elicited using the prompt “I think it’s equally likely that the population of the metropolitan area is above or below”.

We allowed participants to respond in units of thousands or millions — responses could be entered with suffixes of “k” or “K”, and “m” or “M”, indicating thousands and millions, respectively. As the participants entered their responses, the experiment automatically convert the responses into thousands (if greater than 1,000) and into millions (if greater than 1,000,000) and display them so that the participants could visually inspect their answers after conversion, and confirm that the responses were entered as intended. For example, if the participant had entered “1.23m” in the input text field, the experiment would display “1,230,000”, “1,234 thousand(s)”, and “1.23 million(s)” immediately above. This should particularly be useful for minimizing input errors. The orders of the questions were randomized between participants. After the estimation task the participants were asked to self-rate their level of knowledge about geography and population again. Finally they completed a short demographic survey.

Basic results

Of the 101 participants, 37 (36.6%) were female. The average age was 31.8, with $s.d.$ of 10.1. For the sake of brevity, all population figures in this paper represent units of one thousand. That is, a value shown as 1,230 corresponds to an estimate or model prediction of 1,230 thousand, or 1.23 million.

We first inspected the responses visually. Figure 1 displays the responses of four participants of varying performances. The two graphs on top display participants who performed well in terms of getting the overall magnitude of the answers correct; the two at the bottom were two who performed poorly. The two graphs on the left display participants who performed well in terms of getting the relative size of the metros correctly; the two on the right were two who performed poorly. The wide disparity in performance along these two dimensions can be easily seen.

We then checked whether the magnitudes of the estimates did vary widely, both between participants and between items. The median estimates of individuals varied from 25 to 37,500 (25/75 percentiles at 375 and 1600, respectively), suggesting that the between-subjects estimates did vary widely. Similarly, the median ratio of an individual’s highest and lowest estimates was 36.67 (25/75 percentiles at 11 and 80), showing that even within the same individual, the differences between items were very large. Both the range of mean estimates and the range of within-subjects ratios between estimates were much bigger than those possible for estimates about other types of quantities such as percentage or year of events. More importantly, the large differences in magnitude suggest that using the arithmetic mean to aggregate people’s estimates might produce crowd estimates that are unstable from one sample to another. In particular, the variability in the numbers of informants who make estimates on the high end could severely impact the mean. However, we might be able to produce more stable estimates if we can better account for the differences in the magnitudes of individuals’ estimates.

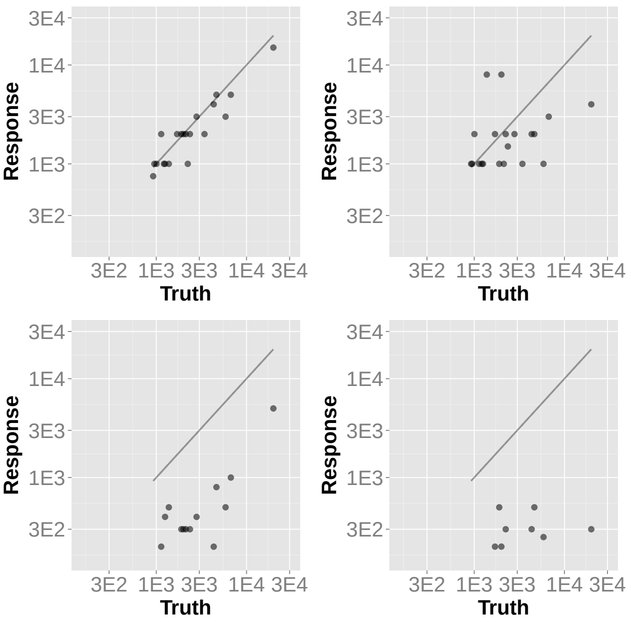

Figure 1: Varieties in individual performance. This figure contains the correlations between the estimates and the ground truth for four participants, highlighting the variety of performance on two dimensions: overall magnitude and accuracy in terms of relative size. The diagonal line spans the line of perfect correlation.

Overall the participants correctly assigned higher estimates to larger metros, and vice versa. The median Pearson’s correlation coefficient between participants’ estimates and the truth was 0.831, with the 25/75 percentiles at 0.658 and 0.931.

We first investigated the results using the standard WoC methods. Two of the most commonly used metrics of prediction performance are the mean absolute distance (MAD) and root mean square error (RMSE). The key difference between the two metrics is the use of the squared loss in RMSE. Squared loss is convex and it means that large prediction errors are overweighted over smaller ones.

One issue with these metrics is that the truths and, to a greater extent, the participants’ estimates have large variability. Thus, performance metrics that rely on the arithmetic difference between estimates and the truth would overweigh items with larger truth values. For example, let’s say participant A made an estimate that is 10% off the truth for the New York question (the largest metro), and 5% for the Las Vegas

[1] `http://en.wikipedia.org/wiki/List_of_Metropolitan_Statistical_Areas` (retrieved Dec 31, 2013)

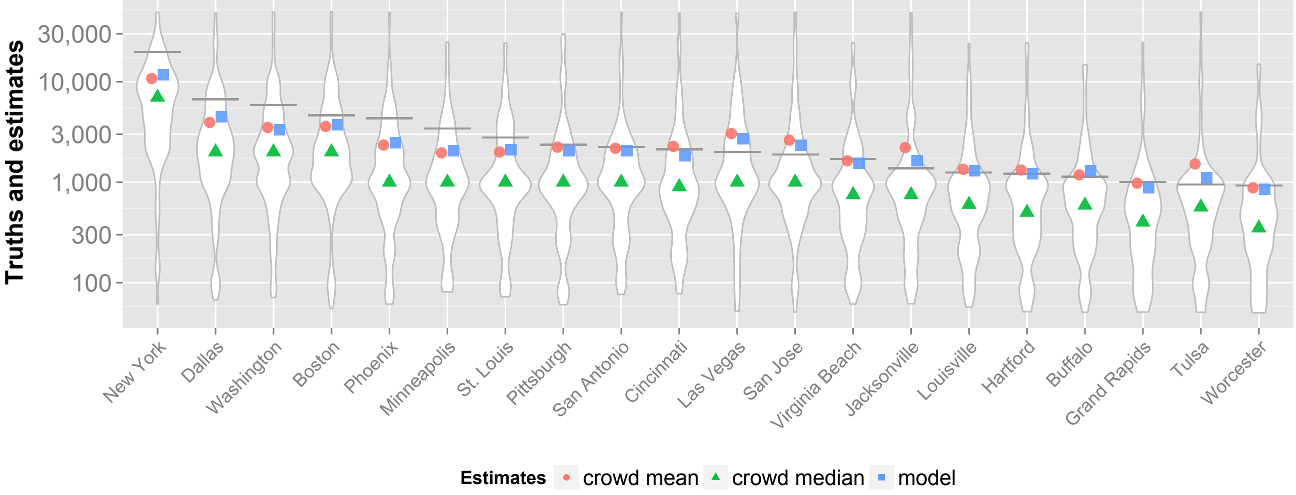

Figure 2: This figure shows the stimuli, responses, and outputs from various wisdom of the crowd methods. The 20 metropolitan areas are listed in the X-axis in descending order of size. For each question, the density of the estimates from the experiment is represented by the violin plot. The horizontal bar indicates the true values. The salmon colored circle represents the WoC mean estimate, the green triangle the median, and the blue square the full model predictions. The Y-axis is log-transformed.

question (11th largest metro); whereas participant B had a 5% error for the New York question and a 10% one for Las Vegas. The total MAD of the two participants in these two questions would be 2083 and 1192, for A and B, respectively. However, it can be argued that these two participants should be rated as equally good, as each made a 10% error and a 5% one.

For this reason, we also used two other metrics that are based on the size of the error relative to the truth (Makridakis, Wheelwright, & Hyndman, 2008): mean absolute percentage error, $MAPE = \frac{1}{N}\sum|100 \cdot \frac{\text{estimate}-\text{truth}}{\text{truth}}|$ and root mean square percentage error, $RMSPE = \sqrt{\frac{1}{N}\sum(100 \cdot \frac{\text{estimate}-\text{truth}}{\text{truth}})^2}$. Like the difference between MAD and RMSE, the main difference between these two metrics is that RMSPE uses squared loss and therefore gives more weight to avoiding large errors. These metrics address the problem of overweighing items with large truth values by evaluating all items on the same scale. For example, an estimate that is 30% higher than the truth for the New York item and one that is 30% higher than the truth for the Las Vegas item would receive equal scores in this metric, regardless of their differences in absolute magnitude. Also, it gives a more generally applicable performance index by giving the error in terms of a percentage of the truth.

Lastly, we compared the model predictions with the truths using Pearson's correlation coefficient, in order to focus of the performance in terms of relative size between the items.

These metrics were computed using the figures taken from the Wikipedia page specified earlier, using the figures from the "2012 Estimate" column. Although we consider the two percentage-based metrics to be more generally applicable, as we will see, the performance numbers from all metrics largely agree with each other. Because of space limit in this paper we will focus on comparisons based on MAPE and MAD. The full result is shown in Table 1. The average MAPE of all participants was 102.1 while the best performing individual had a MAPE of 21.43. This means, on average, the best participant made an error that was 21.43% times of the truth. In terms of MAD, the overall mean was 3013, with a noticeable right skew (best: 582.2; worst: 33,112).

A more interesting analysis is to compare individuals' performance against the wisdom of the crowd. We will focus on two standard ways of aggregation — taking the mean or the median of all individuals' responses. We first look at the performance based on MAPE. The MAPE of WoC$_{mean}$ was 26.23 and was better than all but 2 participants (98.0 percentile). WoC$_{median}$ performed much worse than WoC$_{mean}$, with an MAPE of 58.0, better than only 62.4% of the participants. Performance based on MAD reflects a similar pattern. The WoC$_{mean}$ MAD was 1169, better than all but 4 participants (96.0 percentile), while WoC$_{median}$ MAD was 2118 and was better than only 63.4% of participants.[2] This result was somewhat surprising as we had expected the crowd median to have performed better than the crowd mean, because the median is a more robust statistic than the mean, and we presumed that the variability among estimates would be an issue.

Self-rated expertise was 4.93 at the beginning of the experiment, compared to 4.53 afterwards. A paired t-test showed that the difference was significant ($t(100) = 2.09$, $p = 0.04$). However, neither measure correlates with actual performance. Pearson's correlation coefficient between participants' MAPE and the two elicitations of self-rated expertise were 0.002 and 0.123 respectively (both *n.s.*).

We also looked at the correlation between the performance at the trivia questions and at population estimates. The par-

[2] We also computed the performance for estimates using the geometric mean of all participants. Its performance was worse than all other aggregation methods in all metrics, except for r, where it was worse than only the full model and WoC$_{median}$ at 0.986.

ticipants performed well overall in the trivia questions, getting an average 6.84 ($s.d.$ = 1.00) out of 9 questions correct. However, their trivia scores were almost independent of their MAPE ($r = 0.096$, *n.s.*). The independence of trivia knowledge and self-rated level of knowledge to actual performance agrees with similar results previously reported (Lee, Steyvers, de Young, & Miller, 2012; Lee & Danileiko, submitted).

A computational model

Although the simple WoC procedure of averaging produced estimates that were quite good, it processes each item separately and therefore fails to take advantage of the regularities of the estimates at the individual level. In particular, it ignores the fact that there were high degrees of correlation between estimates made by the same individual. We suggest that models that taking advantage of this characteristic of people's responses might be able to generate more accurate estimates.

Our model focuses on one type of individual level regularity — the overall magnitude of the estimates. We have seen that estimates made by different participants varied to a large degree. We formalize the concept of the overall magnitude using the ratio between an individual's mean estimates to the overall crowd mean. We label this ratio the *scale bias*, and represent it using β. For example, a β of 0.75 means that, on average, this participants' responses are 75% that of the crowd average. Figure 3 shows the distribution of scale bias.

Estimates for any particular metro are positively and highly correlated with all other estimates by the same individual. We computed the Pearson's correlation coefficient between all possible pairs of estimates in the experiment. For these $\binom{20}{2} = 190$ pairs, the smallest correlation was 0.197 and the mean was 0.756. This result demonstrates the highly positive correlation between an individual's estimates. Furthermore, it supports the psychological construct of an individual level scale bias that applies to a closely related set of estimates.

In order to investigate the impact of β on estimate aggregation, we instantiated β in three different ways in our model. In the first variation, one that we call *the full model*, β_i, the scale bias of participant i, is sampled from a prior distribution. The distribution we use is $gamma(2, \frac{2}{\lambda_i})$, in which λ_i is the empirical value of the scale bias of each individual, as calculated from the experimental data. This gamma distribution was chosen as the prior for β because it captures all features that correspond to our prior beliefs about the value of β — it has a mean of λ_i, is unimodal roughly in the middle of the distribution, and is slightly right skewed.[3]

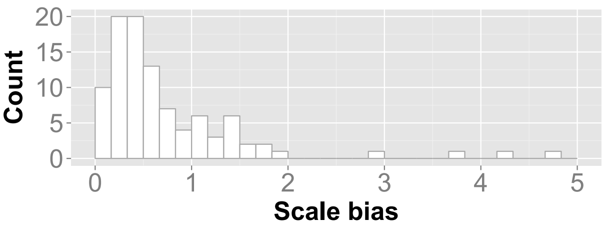

Figure 3: This figure shows a histogram of the scale bias, the ratio between the mean estimate of each participant and the overall mean estimate of all participants.

Table 1: Results of the experiment in different metrics. Performance is measured in MAPE (mean absolute percentage error), RMSPE (root mean square percentage error), MAD (mean absolute distance), RMSE (root mean square error), and r (Pearson's correlation coefficient). For all metrics but r, smaller values indicate better performance. Best performance in each category is in bold.

	Full model	Fixed-β model	Const.-β model	WoC (mean)	WoC (median)
MAPE	**21.07**	23.16	69.36	26.23	58.03
RMSPE	**24.95**	28.52	70.07	33.42	58.73
MAD	**1043**	1144	2587	1169	2118
RMSE	**2060**	2532	4394	2301	3456
r	0.988	0.974	0.965	0.974	**0.989**

An alternate model, *the fixed-*β *model*, is constructed by setting β_i to λ_i. On one hand, this construction of the model is slightly simpler and has a more straightforward modeling interpretation. On the other hand, it prevents the model from incorporating the stochastic noise of each individual's β.

In order to assess the contribution of individualized β_i to model performance, we also implemented a version of the model with the component set to a constant. In this *constant-*β *model*, we simply set all β to 1. This represents a version of the model in which we do not account for the scale bias, and assign all individual differences to individuals' random noise.

Other than the definition of β, the rest of the model is identical between the three variations. The full model is represented graphically in Figure 4. Graphical models are probabilistic models in which graphs are used to express the relationships between variables (Shiffrin, Lee, Kim, & Wagenmakers, 2008). In graphs, nodes represent variables and edges represent the relationship between these variables. Moreover, variables that are observed are shaded; nodes with double border indicate that the values of the variables are deterministically computed based on their parent nodes; and plates group variables that form repeated sub-units. Formally, the model is defined as follows:

$$\begin{aligned}
\psi_j &\sim gamma(10^{-9}, 10^{-9}) \\
\mu_{ij} &\leftarrow \beta_i \psi_j \\
\phi_i &\sim gamma(10^{-9}, 10^{-9}) \\
\sigma_i &\sim gamma(1, \frac{1}{\beta_i \phi_i}) \\
x_{ij} &\sim \mathcal{N}(\mu_{ij}, \sigma_i)
\end{aligned} \tag{1}$$

The main objectives of the model's inference are the latent ground truths, represented by ψ_j for question j. In the model ψ_j is sampled from a weakly informative prior. ψ_j of each item is multiplied with each individual's scale bias β_i to produce μ_{ij}, which can be thought of as the estimate that i-th participant would have produced if the only error were due to the scale bias.

[3] We have conducted a sensitivity analysis concerning the parameters for the gamma prior. We found that the performance of the model does not fluctuate much ($< \pm 0.01$) using the gamma prior $gamma(x, x/\lambda)$ in which $1 \leqslant x \leqslant 10$.

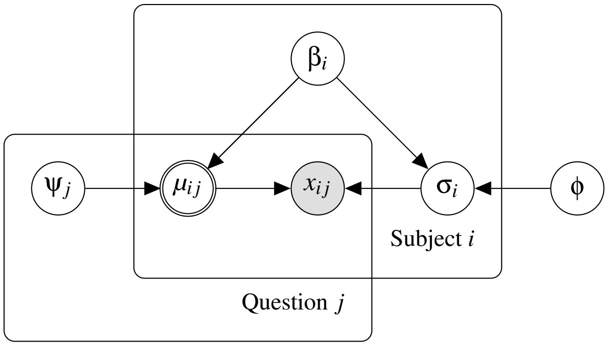

Figure 4: Graphical model for the full model.

The error and random noise for estimates produced by the i-th participant is represented by σ_i, and is dependent on β_i and ϕ. An intuition about the relationship between β_i and errors is that, if an individual's estimates tend to be bigger in absolute terms, the magnitude of the errors are likely to be bigger to the same degree as well, and vice versa if β_i is small. The other component of the estimate error, ϕ, represents both random noise and the normalizing constant for bringing the variability of the estimates to the right magnitude.

The final component of the model is the empirical estimates given by participant i for item j, represented by x_{ij}. It is the only node that is observed, and is produced by sampling a Gaussian distribution with a mean of μ_{ij} and a standard deviation of σ_i.

We implemented the model using a Markov Chain Monte Carlo procedure using JAGS (Plummer, 2004). 1,000 adaptive samples and 3,000 burn-ins were used. The actual samples contain three chains of 30,000 steps each with no thinning. Finally, we emphasize that the model is defined and implemented without the need of knowing the ground truth.

Modeling results

The model predictions are displayed graphically in Figure 2. Although ultimately, the model predictions were generated based on estimates by the crowd, the model produced the estimates through a process much different from simple aggregation. Therefore, in terms of the directions of error from the truth, the model predictions and estimates produced by standard methods do not necessarily agree.

We tabulated the performance of various methods of WoC aggregations in Table 1. These include the full model (with β sampled from a distribution), the fixed-β model (with β_i set to λ_i), the constant-β model (with β set to 1), the crowd mean, and the crowd median. The full model had the best performance in all distance and ratio based metrics. For example, MAPE for the full model is 21.1, versus the next best of the fixed-β model at 23.2, and WoC$_{mean}$ at 26.2. This represents an average error for the full model that is 19% smaller, relative to that of WoC$_{mean}$. Remarkably, the full model outperformed every individual in the experiment.

Similar pattern of results is found based on comparisons using MAD and RMSE. The full model has the best MAD at 1043, and is better than all but 3 of all participants. In terms of RMSE, the full model also outperformed other methods at 2060 (91.1 percentile). The slightly lower performance in RMSE with respect to the crowd was due to a combination of the overweighting of the New York item in an absolute distance based metric, and the fact that the estimates of a few participants for this question were significantly better than those of the crowd. In fact, 51.3% of the full model RMSE was due to this single item. This supports the notion that absolute distance based metrics like MAD and RMSE should be complemented by other metrics for more a comprehensive evaluation of performance.

Finally, we calculated the Pearson's correlation coefficient between the truth and various predictions. Overall the predictions of all models correlated highly with the truth. WoC$_{median}$ performed the best at 0.989, although the performance of the full model was almost the same at 0.988. Both methods outperformed all participants.

To assess the impact of different instantiations of β on performance, we compared the full model to the two variations. The constant-β model performed very poorly. This is not surprising as the model does not incorporate the greatly different magnitudes of the responses, and therefore σ_i needs to be much bigger to account for the huge disparity in all subjects' estimates. As a result, the differences between better and worse subjects were washed out. In contrast, the fixed-β model performed quite well, although it is also outperformed quite clearly by the full model. This suggests that allowing for stochasticities in individuals' scale bias can improve the performance of the model.

Overall, based on all metrics evaluated, the full model compared favorably against the standard WoC methods of mean and median, and against the other two variations of the computational model.

General Discussion

In this paper we investigated the problem of aggregating estimates from informants about quantities that vary by multiple orders of magnitude, both between-items and between-informants. Although this is not a very well understood question, it is an extremely important one because many real life estimation tasks belong to this category.

We proposed a family of hierarchical Bayesian models constructed based on a graphical model, with the three variations differ in terms of how they account for the overall magnitude of an individual's responses. The model relies on psychological assumptions about the structure of responses within individuals. We found that the full model outperformed standard wisdom of the crowd aggregation techniques as well as variations that do not incorporate stochasticity in individuals' overall magnitude.

This work contributes to the existing bodies of work in wisdom on the crowd research in three different ways. First, we showed that estimates from the same individuals have a high degree of correlation, especially in terms of their biases from the truths. Second, we demonstrated that a hierarchical

Bayesian model leveraging the first point can produce better performing estimates for the kind of questions found to be problematic for traditional methods of aggregation (Müller-Trede, 2011; Yeung, 2013). Third, it suggests that utilizing the structure of the responses, in addition to using the responses as isolated data points, can bring upon further improvements in estimating about the unknown ground truth.

The key component of the proposed model is each individual's scale bias, a systematic tendency to over- or under-estimate quantities for a particular set of questions. We found supporting evidence in the high correlations between estimates made by the same individuals. Moreover, its significance can be seen in the huge difference in performance between variations of the model that incorporate it and one that does not. However, we speculate that an individual might not have the same scale bias for all kinds of tasks, but might rather like the Person × Situation interaction theory (Diener, Larsen, & Emmons, 1984), in that different circumstances might bring out different scale biases for the any individual.

The distribution of the scale bias in a crowd is an intriguing topic in of itself and has raised a few interesting theoretical questions. First, why are the mean estimates of individuals self-organize into the distribution shown in Figure 3? Second, how stable is this distribution and would similar distributions be found in estimation of quantities in other domains? Although in the current paper we have not explored these questions, future research might be able to shed light on them.

The distribution of the β parameter also has great implication on the application of the model, as it gives rise to the great variability of the estimated quantities. The results showed that while completely negating this variable produced pretty bad estimates, simply setting β to the ratio between an individual's mean estimate and the overall crowd mean produced reasonable performance gain over traditional WoC methods. However, we speculate that knowing how this variable is distributed in the population will help us further improve WoC techniques. More specifically, a model that makes the correct assumptions about the distribution of the crowd's scale bias might be able to produce better results with smaller sample of informants, or even with a single informant.

Over- and under-estimation have previously been studied extensively. However, in most of these cases, the target of these mis-estimation were self-relevant quantities such as one's own levels of confidence, abilities, or performance (e.g., Moore & Healy, 2008; Klayman, Soll, González-Vallejo, & Barlas, 1999), or in the judgment of probability (e.g., Kahneman & Tversky, 1979). The scale bias variable studied in the current paper manifested in quantities that were neither self-relevant nor about probability. Hence, systematic errors at the individual level signal a potentially exciting new research direction, especially because prescriptive procedures such as those suggested above might be able to alleviate such bias. Moreover, although the experiment used questions with knowable ground truth, we expect that this technique will be useful in improving predictions and other estimates in which the ground truth in not known in advance.

Conclusion

In this paper, we have highlighted the problem of aggregating estimates for quantities that are not naturally bounded. The hierarchical Bayesian model we proposed makes assumptions about the structure within individual's estimates to produce better estimates than those by the standard aggregation methods. Applying this idea to other wisdom of the crowd problems might similarly improve estimates and bring insights to our knowledge about how individuals and crowds produce judgments and make decisions.

Acknowledgments. We acknowledge the support of the Basic Research Funds from the Beijing Institute of Technology, and Grant No. 71373020 from the National Natural Science Foundation of China.

References

Barabási, A. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*, 207-211.

Diener, E., Larsen, R. J., & Emmons, R. A. (1984). Person × situation interactions: Choice of situations and congruence response models. *Journal of Personality and Social Psychology*, *47*(3), 580–592.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review*, 1429–1451.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 263–291.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational behavior and human decision processes*, *79*(3), 216–247.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7.

Lee, M. D., & Danileiko, I. (submitted). Using cognitive models to combine probability estimates.

Lee, M. D., Steyvers, M., de Young, M., & Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science*, *4*(1), 151–163.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting methods and applications*. Hoboken, NJ: Wiley.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. *Judgment and Decision Making*, *6*(4), 283–294.

Plummer, M. (2004). *JAGS: Just another Gibbs sampler.*

Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of mathematical Psychology*, *55*(2), 191–197.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science*, *32*(8), 1248–1284.

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor.

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

Yeung, S. (2013). Wisdom of the crowd can improve confidence interval estimates, but a systematic bias could lead to underperformance. In *The society for judgment and decision making annual conference.*