Using Ground Truths to Improve Wisdom of the Crowd Estimates

Andrew Whalen (aczw@st-andrews.ac.uk)
School of Biology, University of St Andrews, St Andrews, Fife KY16 9TH UK

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

Abstract

In this paper we explore a cognitive modeling approach to aggregating individuals' estimates of unknown quantities without natural bounds. We carried out two experiments that elicited individuals' estimates of the population of US metropolitan areas, and domestic box office returns for movies. We found that the means of individuals' responses correlate well with the true sizes, but participants systematically underestimated these values. We formulated a cognitive model that uses the true values of known items to correct for individuals' biases, and demonstrated that this model can drastically improve predictive accuracy. Because our model quantitatively infers individual's biases on the estimation tasks we were able to examine the distribution of individual biases, and found that there were substantial between-individual differences in the magnitude of the responses. This work demonstrates how individuals' biases, whether over- or underestimation, can be corrected using a cognitive model together with known ground truths.

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

Introduction

Past research has found that the aggregate estimates or predictions of multiple untrained individuals can outperform experts in a wide variety of domains (e.g., Galton, 1907; Clemen, 1989; Surowiecki, 2004). This effect has been called the "wisdom of the crowd" and was shown to be an efficient way to provide estimates for unknown quantities (Clemen, 1989) or solutions to technically challenging problems (Yi, Steyvers, Lee, & Dry, 2012). One of the advantages of using a wisdom of the crowd procedure over taking the advice of a single individual is the reduction in individual biases: by averaging over multiple individuals, their biases are likely to cancel out each other, leaving the aggregate estimates less biased (Surowiecki, 2004).

Early work on the wisdom of the crowd effect found that the mean or median of individuals' responses could often produce good forecast for the likelihood of future events or good estimates for unknown quantities (Armstrong, 2001; Larrick, Mannes, & Soll, 2012). However, recent research has also revealed a number of situations in which the simple aggregations underperform. First, aggregate estimates can be distorted by systematic group bias—if the majority of the crowd over- or underestimate on a task, then the aggregate will tend to be biased in the same direction as well (Simmons, Nelson, Galak, & Frederick, 2011).

Second, a substantial body of psychological research has found that individuals are biased in how they process numbers (Kahneman & Tversky, 1979; Dehaene, 2003). In particular, the standard aggregation techniques of mean or median perform less well in unbounded estimation tasks—here the values being estimated are not constrained between 0 and 1, and may lie across several different orders of magnitude. Take the size of cities as an example. The smallest cities have only dozens of people, but the largest cities can have over 10 million. Recent research has found that crowd means and medians performed less well for quantities with the large variation in magnitude (Yeung, 2013).

Various techniques have been proposed to mitigate these two problems. Budescu and Chen (2015) demonstrated a statistical approach to reduce systematic biases by using known ground truths to identify experts in the crowd and overweight their judgments. They showed that this approach can significantly improve the aggregate estimates. With respect to the psychological biases, Lee and colleagues created cognitive models that explicitly take into account individuals' differences in calibration of probability. They showed that this approach can create aggregate estimates better than taking the crowd mean or median (Lee & Danileiko, 2014; Lee, Steyvers, de Young, & Miller, 2012).

Building upon these findings, we explore in this paper how to improve the aggregation of the responses of a crowd about unbounded quantities using a Bayesian approach. The goal of this work is twofold. On the one hand, we expand on previous techniques for leveraging known ground truths in a wisdom of the crowd framework, and provide a novel technique for accounting for individuals' biases in estimation of unbounded quantities. On the other hand, we also seek insights into individual biases in unbounded estimations. This can shine light on the underlying psychological processes underlying estimation of unknown quantities.

In the rest of this paper, we first present two experiments in which individuals estimated unbounded quantities in two different domains, and examine the predictive accuracy based on the standard wisdom of the crowd technique. Next we present two Bayesian cognitive models that takes into account known truth values and individual biases. To examine the performance of the model predictions we carry out a simulation study and compared the accuracy of our models against those several other aggregation techniques. Finally, we explore the distribution of biases among individuals, and conclude by providing recommendations for future wisdom of the crowd aggregation techniques.

[†] Corresponding author

Experiments 1 and 2: Estimating Metropolitan Populations and Box Office Returns

In these two experiments we collected data on how individuals make estimates about unknown quantities and analyzed the performance of various aggregation techniques. In Experiment 1, participants estimated the sizes of 20 US metropolitan areas, whereas in Experiment 2, the domestic box office returns for the 20 highest-grossing movies of 2013.

Each experiment offers a place to better understand individuals' judgments about unbounded quantities. Although the underlying distributions of metropolitan populations and movie returns are both log-normal (Eeckhout, 2004; Griffiths & Tenenbaum, 2006), the distributions have different tail behaviors. In the case of box office returns figures, the values tend to be within the same order of magnitudes. For example, the top grossing movie of 2013, The Hunger Games, grossed $424 million while the third highest grossing movie, Frozen, grossed a comparable figure of $401 million. In contrast, there is far more weight on the tail end of metropolitan population sizes. The largest metropolitan area in the US, New York, has 20 million inhabitants, while the third largest metropolitan area, Chicago, has half the number of people at 10 million inhabitants. The contrast is similar for other members of each list as well.

The difference in tails allows us to test the effectiveness of wisdom of the crowd aggregation techniques for unbounded estimation tasks, both when the quantities tend to be within the same order of magnitude, and when they tend to be on different orders of magnitude. The metropolitan population data set was previously reported in Yeung (2014). We present both experiments together.

Participants

Participants were recruited from Amazon's Mechanical Turk (`http://mturk.com`). There were 101 participants in the Experiment 1, and 100 participants in Experiment 2. Because Experiment 1 took participants slightly longer they were compensated US$0.40 for their time, whereas participants in the Experiment 2 were compensated US$0.30 for their time. Participants were required to be 18 years or older, be residing in the U.S., and have a lifetime acceptance rate on Mechanical Turk of at least 95%.

Methods

The experiments were web-based and were administered using the Qualtrics survey service. Participants were instructed to not use any external resources during the task. At the beginning of the experiment, we asked participants to self-rate on a seven-point scale (from "Very Good" to "Very Poor") their level of knowledge about geography or about movies from 2013. In Experiment 1 participants were also given a general geographic knowledge questionnaire. Performance on this questionnaire did not correlate with the participants' performance, similar to what was found by Lee and Danileiko (2014), and so the questionnaire was dropped for Experiment 2.

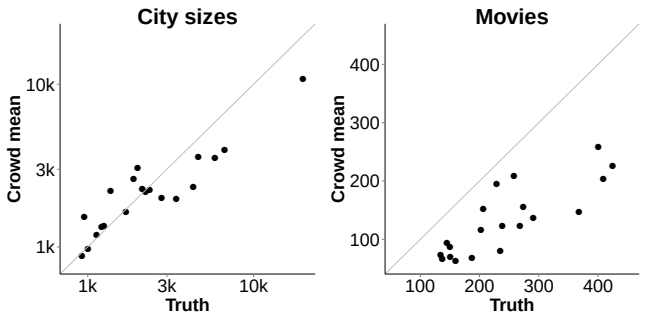

Figure 1: The means individual responses compared to the true values.

Next participants were asked to estimate the size of 20 metropolitan areas in the year 2012, or the domestic box office returns for the top 20 grossing movies of 2013[1], by giving their median estimates for either quantity. The language used follows that of Soll and Klayman (2004). For example, in Experiment 2, the estimates were elicited using: "I think it is equally likely that the domestic ticket sales of this movie is above or below (in millions): ____". In the metropolitan size experiment, participants had the option to give their estimates in either millions (i.e., "1.2m") or thousands (e.g., "1,200k"). In the box office returns experiment, participants gave estimates in millions. In Experiment 2, after giving the estimations, the participants were asked whether they had heard of each movie before. At the end of each experiment, participants completed an optional short demographic survey.

The true data for the US metropolitan population were collected from the U.S. Census (United States Census Bureau, 2013), and those for the box office returns were collected from the web site Box Office Mojo (Box Office Mojo, 2014).

Basic Results and Discussion

We analyzed the performance of the mean and median estimates of all individuals' responses as compared to the true population or box office figures. We found a high correlation between the mean of individuals' responses and the true values in both the metropolitan population, $r = .97$, and box office returns, $r = .83$ (Figure 1). The mean estimates were generally lower than the true values in both experiments (particularly Experiment 2), suggesting that on average individuals underestimate the population of metropolitan areas and box office returns. This effect was not significant in Experiment 1, based on a two-tailed paired t-test, $t(19) = 1.61$, $p = 0.12$, although it was in Experiment 2, $t(19) = 8.89$, $p < 0.01$.

We also analyzed two measures of accuracy: the root mean squared error (RMSE) and the root mean squared percent error (RMSPE) (Makridakis, Wheelwright, & Hyndman, 2008). These metrics provide insights into the absolute and proportional errors of these estimates with respect to the

[1]Experiment 1 was run in 2013 and Experiment 2 was run in 2014, so both sets of values were for the preceding year.

Table 1: Performance of the crowd means and medians. For RMSPE and RMSE, smaller values indicate better performance; for correlation, higher is better.

Experiment 1: Metropolitan Population	Mean	Median
RMSPE	33.42	58.73
RMSE	2300.85	3455.93
cor	0.97	0.99
Experiment 2: Box Office Returns	**Mean**	**Median**
RMSPE	47.66	58.53
RMSE	123.89	147.70
cor	0.83	0.86

truths, respectively. The two measures are calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - e_i)^2}$$

$$RMSPE = 100\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{t_i - e_i}{t_i})^2}$$

where i is the index for the n questions, t_i is the true value for value i, and e_i is the estimated value for e_i. As these two metrics represent the amount of error in the estimates, lower values represent better performance. The results based on these metrics for the crowd means and medians are given in Table 1. We found that in both experiments the crowd mean outperformed the crowd median. This is likely due to the fact that, while most of the individuals underestimated the true values, a small number of individuals drastically overestimated. These outliers increased the value of the means and brought them closer to the truths, but had little impact on the medians. We found that other performance metrics, including the mean absolute distance (MAD), give similar pattern of results—the mean estimates outperformed the median estimates.

The high correlation between mean responses and the true values suggests that the crowd, as an aggregate, was highly accurate in terms of judging the relative sizes of these quantities. It is possible that individuals have a poor sense of what scale these quantities lie on: participants may not know if box office returns are on the order of tens of millions or hundreds of millions, or if high grossing movies make $200 million, or $400 million. Because of this, they may over- or underestimate these quantities. Although in general it may be hard to determine a priori if a group of individuals will over- or underestimate a quantity, in both experiments we found the overall estimates to be low.

The above finding suggests that correcting for individuals' scaling biases might be useful in improving aggregate estimates. If some of the true values are known before the elicitation, we can use the technique of *seeding*—giving individuals some ground truth data (e.g., the true box office figures for a handful of movies), so that they can update their knowledge concerning the distributional properties of the entities in that category and produce more accurate judgments in subsequent trials. Brown and Siegler (1993) investigated this technique (outside of the wisdom of the crowd context) and found that knowledge of ground truths improved participants' accuracy. However, seeding often required a large number of known truth values (in many cases over 20). Yeung (2013) found that providing up to three seeds had a limited impact on improving estimates. Often times we may not have access to a large number of ground truths, either because they are costly to gather, or because they involve events that will happen in the future. We instead take a different approach by constructing a cognitive model that can take into account various amounts of ground truths and estimate both individuals' biases and the quantities of interest.

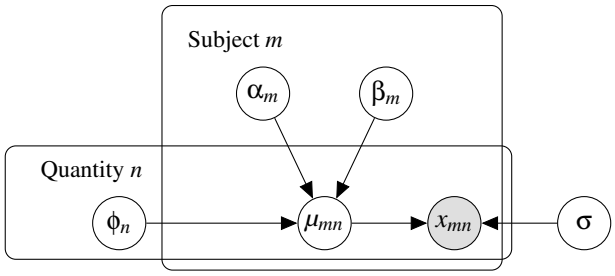

Figure 2: A graphical representation of the Bayesian cognitive model used to aggregate participant responses. In the individual level model, α and β were assumed to vary between individuals, while in the group level model α and β were the same for all.

Aggregating Estimates Using a Cognitive Model

Our model follows from previous Bayesian cognitive models used to correct for individual biases in creating wisdom of the crowd aggregates (Lee & Danileiko, 2014; Yeung, 2014). These previous approaches relied on implicitly correcting for individual differences in expertise and their probability weighting function. In contrast, our model focuses on the assumption that individuals have good knowledge about the relative magnitudes of each item, but also have individual-level biases with respect to the scales of these values.

Informally, this model takes as input the judgments of each individual and the partial ground truth. It then uses the known ground truths to correct for the individual level biases by implicitly estimating the degree of over- or underestimation of each individual. Finally, it produces estimates on the items without known ground truth based on each individual's estimates and the computed biases.

The model is formalized using a graphical model (Figure 2). The true values are represented by ϕ_n, where n indexes a specific question. In the model we assume that individuals are influenced by ϕ_n but their responses are subject to scaling bias that they are not aware of. In the case of the box office data set where each of the values are within an order of magnitude of each other, we model the scaling using a linear function, so that the estimate of participant m on question n is given by $x_{m,n} = \alpha_m + \beta_m\phi_n + \varepsilon$, where ϕ_n is the true value for question n, α_m and β_m are individual parameters that account

for the scaling biases of individual m, and ε is a normally distributed error term. Because the values for metropolitan populations lie across multiple orders of magnitude, we use an exponential function instead, given by $x_{m,n} = \alpha_m \phi_n^{\beta_m} + \varepsilon$, which is equivalent to linear scaling on a log scale.

To evaluate the degree to which the scaling biases of individuals varied, we considered two forms of the model: in the group level model, all α_m and β_m were assumed to be equal; in the individual level model, α_m and β_m varied between individuals. The individual level model can be considered as a general case of the group level model.

The goal of the model is to infer the true value of each quantity ϕ_n, using participant responses $x_{m,n}$ while simultaneously estimating ϕ_n, α_m, and β_m. To do this we define the following dependencies for the metropolitan population estimate:

$$
\begin{aligned}
x_{m,n} &\sim lognormal(log\mu_{m,n}, \sigma) \\
\mu_{m,n} &= \alpha_m \phi_n^{\beta_m} \\
\sigma &\sim lognormal(0.37, 0.12) \\
\phi_n &\sim lognormal(6.73, 3) \\
\beta_m &\sim lognormal(0, 1) \\
\alpha_m &\sim lognormal(0, 1)
\end{aligned}
$$

And similarly for the box office returns:

$$
\begin{aligned}
x_{m,n} &\sim normal(\mu_{m,n}, \sigma) \\
\mu_{m,n} &= \alpha_m + \beta_m \phi_n \\
\sigma &\sim normal(118.6, 49.8) \\
\phi_n &\sim lognormal(4.33, 1.2) \\
\beta_m &\sim lognormal(0, 1) \\
\alpha_m &\sim normal(0, 100)
\end{aligned}
$$

The prior distributions were either chosen to be uninformative (α and β), or computed using the empirical data (σ and ϕ). In the case of ϕ, the prior distribution was created by examining the distribution of all participants' responses. We found this distribution to be approximately log-normal in both the metropolitan size and the box office returns tasks, and used the mean and variance of those distributions to create the priors. The prior on σ was constructed by taking the mean and standard deviation of the standard deviation of participants' responses.

One of the main advantages of our model is that it can naturally incorporate ground truth data, making it straightforward to improve the aggregate judgments and to provide estimates of each individual's scaling biases. Here the target values ϕ's, regardless of whether they are known, are represented as nodes in the graphical model. To represent the known values in our model, we fixed the values of the known ϕ_n's to the known values, whereas the ϕ_n's without known values remained in the model as the unknown latent variables whose values are to be inferred.

Bayesian inference of our models was performed using Stan (Stan Development Team, 2014), which uses a No U-Turn Sampling algorithm to estimate the posterior distribution (Hoffman & Gelman, 2011). α and β were initialized to 1, and ϕ_n to the means of all participants' estimates. For each model considered, we ran a single chain for 4,000 samples, after a burn-in period of 4,000 samples. Pilot simulations suggested that this was sufficient time to reach convergence.

For each model we used the mean value of ϕ_n in the posterior distribution as the model's estimate for each quantity. Estimates of each individual's scaling biases, α_m and β_m, were obtained similarly.

Simulation

To evaluate the performance of the models, we ran two sets of simulations, one on the metropolitan population sizes (Experiment 1) and the other on domestic box office returns (Experiment 2). We varied the number of known ground truths from 1 to 19 (out of 20) to evaluate model performance with different numbers of known values. Because performance of the models depend on which questions are chosen to be known, we ran a series of simulations in which the questions with known values were randomly sampled. For each number of known truths we sampled 100 different combinations of questions with known values.[2] We then fitted both the group and individual level models to the data, and additionally, computed the crowd means as our baseline measure. Model performances were compared based on RMSPE and RMSE.

Results

Figure 3 gives the performance (in RMSPE) of both group and individual level models and the crowd mean with 1 to 19 known data points, separately for the two experiments.[3] In Experiment 1, the individual level model had better performance than both the group level model and the crowd mean for all numbers of known truths. The poor performance of the group level model with a small number of known data points is likely due to difficulties in fitting the exponential scaling function, which has the possibility of resulting in drastically worse fits than the mean.

The performance of the crowd mean remained stable over different numbers of known truths. This is expected because it does not take into account of known truths. The performance of both Bayesian models, however, improved as the number of known truths increased. If all but one estimates (19) are known, the individual level (20.60) and group level (22.44) model had fairly similar performance, and both were better than those of the crowd mean (26.20).

A similar pattern of result was found in Experiment 2. The individual level model had the best performance across all

[2] For simulations with 1 or 19 ground truths, we performed only 20 simulations, one for each possible set of known data points.

[3] As the median estimates performed even worse than the mean estimates, their performance figures are not reported here.

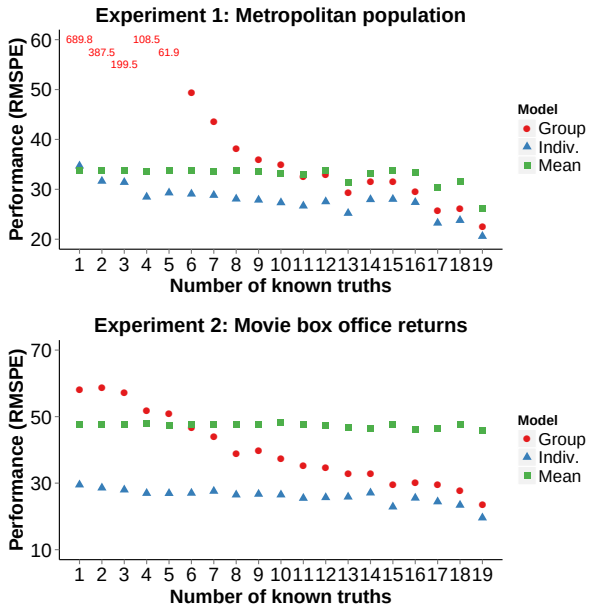

Figure 3: Performance of Experiments 1 (top) and 2 (bottom). In each chart, performances in RMSPE for all three models evaluated are plotted in the order of numbers of known truths.

numbers of known truths. The group level model performed poorly with only one known ground truth, but its performance improved with higher numbers of known truths. The individual level model maintained its performance advantage across different numbers of known ground truths. For example, when all but one data point were known, the individual level model (19.59) had better performance than both the group level model (23.50) and the crowd mean (45.68). For both experiments, analysis based on RMSE yielded similar results.

In both experiments, the individual level model outperformed the group level model, although the relative performance of the group level model did improve substantially as more data points were known. Such differences suggest that there are meaningful individual differences in the scaling biases. At the same time, the fact that the individual level model was able to achieve good performance with relatively few numbers of known data points suggests that the scaling biases were largely consistent within the same individuals and across different questions. Taken together, these results suggest that there may be stable individual differences in the scaling biases, at least within the same domain.

We can specifically analyze these differences by examining the inferred values of α_m and β_m. To do this, we ran an additional model where all 20 of the ground truths were assumed to be known. A histogram of the resulting mean values for α and β is given in Figure 4. Consistent with the above results, we find that there is a large variability among individuals' scaling bias parameters. Moreover, the results suggest that many individuals underestimated these quantities while a

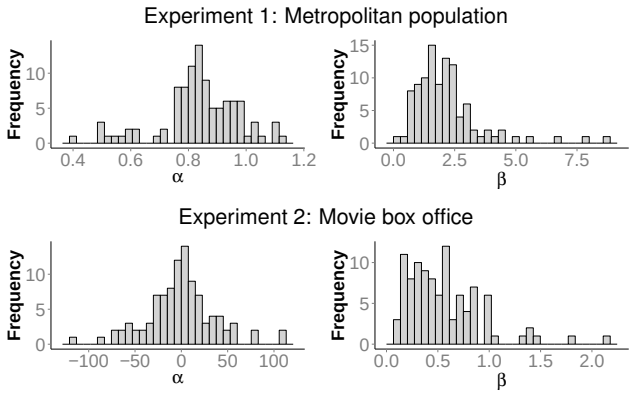

Figure 4: Histogram of the inferred individual level scaling biases.

minority overestimated.[4]

General Discussion

In this paper we explore how to combine individuals' judgments to make accurate estimates for unknown and unbounded quantities. We used two experiments to elicit individuals' estimates in two different domains. We found that individuals systematically underestimate the populations of metropolitan population and domestic box office returns. We constructed an individual level and a group level cognitive model that correct for individuals' scaling biases and make aggregate estimates. We found that the individual level model outperformed the crowd mean, crowd median, and the group level model. We also found substantial individual level differences in terms of their scaling biases.

In our experiments there were no truly unknown values. However, our models can be used in situations in which we know the ground truths for only some of the questions, and need to create accurate estimates for the unknown ones. These situations could be results of a variety of reasons. For example, we might only have the resources to determine the truths for some of the questions; or, the judges were originally asked to make predictions on all events, and later the outcomes for some of these events are known, and we want to use these known outcomes to better estimate the ones not yet finalized. Our findings will be useful in creating estimates in these situations.

This work presents one approach to aggregating known ground truths in a wisdom of the crowd framework. Budescu and Chen (2015) demonstrated a different approach. They used known ground truths to identify experts in the crowd and overweight these experts' judgments in order to create more accurate aggregates. Future work is needed to evaluate whether the current approach can be combined with the identify-the-experts approach to further improve the aggregate estimates.

Another alternative is to provide individuals with ground

[4] In Experiment 1, a value of $\alpha < 1$ or $\beta < 1$ suggests underestimation; in Experiment 2, a value of $\alpha < 0$ or $\beta < 1$ suggests underestimation.

truth data and so that they might recalibrate their own judgments based on these data. Although this approach can work well—Brown and Siegler (1993) found that providing ground truths in one domain lead to increased accuracy in other questions on the same domain—it seems to require a high number of known truths. Our model has the advantage of not requiring the ground truths to be known at elicitation time and requiring fewer known data points.

While a simple wisdom of the crowd setup may allow for the estimation of some quantities even without ground truth data, we demonstrated that in at least two domains, these estimates were biased, agreeing with the results of Simmons et al. (2011). However, we found that, using an individual level cognitive model, only a small number of known truths is needed to correct for individual biases, and significantly boost performance, suggesting that this approach may be useful in cases where collecting ground truths is expensive.

One advantage of using a cognitive model to examine aggregate judgments is that we can better understand the computational principles underlying estimation of unknown quantities. Although we found substantial variabilities among individuals' scaling biases, the within-individual biases across questions were found to be quite consistent. This result is supported also by the high performance of the individual level model with only a small number of known data points. This suggests that scaling parameters may be linked to the underlying psychological processes of how individuals make these estimates.

Overall this paper shines light on how individuals estimate unknown, unbounded quantities, and provides a method for correcting for over- or underestimation in individuals' judgments. We found that a cognitive model was able to account for individuals' scaling biases, and obtained better performance than both the crowd means and the crowd medians. This work demonstrates how to further improve wisdom of the crowd aggregation techniques; this is a particularly important finding as the crowd mean is already quite competitive compared to experts' judgments on a wide variety of tasks (Surowiecki, 2004). By integrating known truth data with a Bayesian cognitive model, we show that the performance of aggregate judgments can be improved even further and may provide an efficient way to obtain expert quality estimates in a broader range of tasks.

Acknowledgments. This research was supported by John Templeton Foundation Grant #40128.

References

Armstrong, J. S. (2001). Combining forecasts. In *Principles of forecasting* (pp. 417–439). Springer.

Box Office Mojo. (2014). *2013 Yearly Box Office Results - Box Office Mojo.* http://www.boxofficemojo.com/yearly/chart/?yr=2013.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*(3), 511–534.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Dehaene, S. (2003). The neural basis of the Weber-Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences*, *7*(4), 145–147.

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

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Hoffman, M. D., & Gelman, A. (2011). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1111.4246*.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers of Social Psychology: Social Psychology and Decision Making* (pp. 227–242). Philadelphia: Psychology Press.

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*.

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.

Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. (2011). Intuitive biases in choice versus estimation: implications for the wisdom of crowds. *Journal of Consumer Research*, *38*(1), 1–15.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning Memory and Cognition*, *30*(2), 299–314.

Stan Development Team. (2014). *Stan: A c++ library for probability and sampling, version 2.5.0.* Retrieved from http://mc-stan.org/

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Anchor.

United States Census Bureau. (2013). *2012 Population Estimates.* http://www.census.gov/popest/data/metro/totals/2012/tables/CBSA-EST2012-01.csv.

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.*

Yeung, S. (2014). A hierarchical bayesian model for improving wisdom of the crowd aggregation of quantities with large between-informant variability. In *36th Annual Conference of the Cognitive Science Society.*

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science*, *36*(3), 452–470.