Collective Wisdom: Methods of Confidence Interval Aggregation[1]

Aidan Lyon[1], Bonnie C. Wintle[2], and Mark Burgman[2]

1: Philosophy Department, University of Maryland, College Park, Maryland, 20742, USA. alyon@umd.edu
2: Quantitative and Applied Ecology Group, Environmental Science, School of Botany, University of Melbourne, Parkville, Victoria, 3010, Australia. bonnie.wintle@unimelb.edu.au
3: Centre of Excellence for Biosecurity Risk Analysis, Environmental Science, School of Botany, University of Melbourne, Parkville, Victoria, 3010, Australia. markab@unimelb.edu.au

Preprint, August 27, 2014.

To appear in *Journal of Business Research: Conditions and Complexity in Forecasting*

Abstract

We report the results of a meta-analysis study of the relative accuracies for a range of methods for aggregating confidence interval estimates of unknown quantities. We found that a simple *trim-and-average* method—that is, remove outliers and then average—produced the most accurate estimates. Our results show that more complicated methods of confidence interval aggregation, which factor in confidence levels and estimate imprecisions, do not produce estimates more accurate than those produced by the simple trim-and-average method.

1 Introduction

The so-called *wisdom of crowds*, or *collective wisdom*, has proven to be an effective method for forming accurate judgements in situations of uncertainty (see e.g., Armstrong 1989, Clemen 1989, Surowiecki 2004, Page 2008, and Nielsen 2011). The essence of the idea is simple: by collecting a large number of judgements and combining those judgements into a single, collective judgement, one will obtain a judgement that will tend to be more accurate than the individual judgements. In the language of folk wisdom: *two heads are often better than one*, and more heads tend to be *even* better.

[1]Aidan Lyon acknowledges support from the Munich Centre for Mathematical Philosophy, the Humboldt Foundation, and the Centre of Excellence for Biosecurity Risk Analysis.

How one decides to combine the individual judgements into a collective judgement is one's *aggregation method*. An aggregation method takes as part of its *input* a set of individual judgements, and produces a single judgement as its *output*. Aggregation methods can be classified in terms of their inputs and outputs, and we can break them up into two broad categories: (i) methods whose input and output judgements are of the same kind, and (ii) methods whose input judgements are different in kind from their output judgements. An example of the first kind of aggregation method is a jury voting method: the inputs are binary "innocent"/"guilty" judgements and the output is also a "innocent"/guilty" judgement (see e.g., Condorcet 1785, Ladha 1992, and List and Pettit 2002). The vast majority of the literature on aggregating judgements tends to focus on the first kind of aggregation method, with a particular focus on point estimates/forecasts of quantities and probability estimates/forecasts of facts or events (e.g., Clemen 1989, Genest and McConway 1990, Cooke 1991, Clemen 1999, Clements and Harvey 2011). This paper is a study of the second kind of aggregation method.

Why aggregate judgements of one kind into a collective judgement of another kind? One simple reason is that the available input judgements may happen to be different from the desired kind of judgement. For example, a probabilistic forecast of an event may be desired but only binary judgements are available, in which case the binary judgements will somehow need to be combined into the desired probabilistic outcome.

The above reason relates to outputs. A second reason relates to process. In translating one kind of judgement to another—or even, complementing one kind of judgement with another—there is an opportunity to use elicited information in the aggregation process. For example, suppose we need a point estimate of how much sales will grow by in the next year and we consult two sales experts. The first expert reports that sales will grow by 5% and the second reports that they will grow by 10%. In this situation, it probably makes sense to take a simple average of the estimates, where the estimates are given equal weight. However, suppose instead that we ask the sales experts for interval estimates that they are 90% confident in. The first expert reports that sales will grow by 4-6% and the second expert reports that sales will grow by 0-30%. All else being equal, it now appears that we might be better off giving more weight to the first expert's estimate, because, given standard conversational norms (Grice 1975), precision in a judgement indicates that the judge believes he or she has relevant knowledge about the issue at hand. Indeed, Yaniv 1997 found that the average of the midpoints of confidence intervals, weighted by the precisions of the confidence intervals, tended to be more accurate (in terms of mean

absolute percentage error) than the simple, unweighted average of the midpoints.

However, individuals are notoriously overconfident when they produce interval estimates for pre-set confidence levels (e.g., Alpert and Raiffa 1982, Lichtenstein, Fishhoff, Phillips 1982, Yaniv and Foster 1997, Soll and Klayman 2004, Teigen and Jørgensen 2005). Although this overconfidence is high and robust, it can nevertheless be mitigated by asking for the lower and upper bounds of the interval estimates separately (Soll and Klayman 2004). And Speirs-Bridge, Fidler, McBride, Flander, Cumming, Burgman 2010 have shown that overconfidence can be further reduced if individuals are allowed to assign their own confidence levels to their interval estimates—rather than producing an interval estimate for a level of confidence that is pre-set by the experimenter. This effect leads to a natural question: Given a set of confidence intervals with varying levels of precision and varying probabilities (i.e., confidence levels), are there aggregation methods, which produce point estimates, that are systematically better than simple unweighted averages? This paper examines this issue by conducting a meta-analysis of the performances of a range of aggregation methods of varying complexity over the results of 15 experiments which elicited such confidence levels.

2 Theory: Aggregation Methods

The baseline of confidence interval aggregation is the method that simply takes the unweighted average of the midpoints of each interval (*cf.* Yaniv 1997). If N confidence intervals have been elicited, and m_i are their midpoints, this aggregation method says:

$$\text{Unweighted Average: } \mathcal{J} = \frac{1}{N}\sum_{i=1}^{N} m_i \tag{1}$$

where $\mathcal{J}$ is the collective judgement. The unweighted average is *guaranteed to be no less accurate*—in terms of absolute error—than the typical individual judgement, and if the individual judgements *bracket* the true value of the quantity in question (i.e., some of the m_i are too low and some are too high), then the unweighted average is *guaranteed be more accurate* than the typical individual judgement (*cf.* Larrick and Soll 2006 and Herzog and Hertwig 2009). Since we take this to be the baseline method, we will use it to assess the performance of the other aggregation methods described below.

The unweighted average has proven to provide surprisingly accurate collective judgements (Armstrong 1989, Surowiecki 2004, Fischer and Harvey 1999; Wallsten, Budescu,

Diederich 1997). However, there are number of ways in which it appears that the unweighted average can be improved upon. These possible improvements fall into kinds: (i) statistical improvements, and (ii) psychological improvements. If we think of each individual human judgement as a *measurement* of the quantity that we're interested in and each human as a *measuring device* with its own bias and distribution of error, then the process of determining a collective judgement is a familiar task in statistics: choosing an appropriate estimator given what is known about the data. Viewed this way, the problem of judgement aggregation is just a problem of statistics, and we can help ourselves to the statistician's sophisticated tools of estimation.

However, we know that the measuring devices are really humans, with their own complicated psychologies. They think about evidence and they form beliefs with uncertainties and biases, and they announce those beliefs and uncertainties with possibly yet another layer of biases (Kahneman, Slovic, Tversky 1982). Given our knowledge about human psychology, it seems that there ought to be ways of improving upon the unweighted average by incorporating more psychological information about the judgements. The Unweighted Average of the midpoints ignores a lot of psychological information that can also be gleaned from the individual confidence intervals, and perhaps by incorporating that information in the average, we can produce a better aggregation method. We will first discuss some psychologically motivated improvements over the Unweighted Average, and then we will turn to the statistical ones.

Given that we are assuming that confidence intervals have been elicited as our inputs, we know that the Unweighted Average of the midpoints of the intervals ignores two potentially important pieces of information contained in the confidence intervals: (i) the confidences assigned to the intervals, (ii) and the precisions—or widths—of the intervals. For example, if someone is asked what the unemployment rate is, and they respond with the interval of 0% to 50%, then that is a sign that they know (or at least think they know) very little about the true value of the unemployment rate. In contrast, someone who gives a very precise interval, say 7.1% to 7.2% seems to have (or at least think they have) a lot of evidence about the unemployment rate. In which case, it may be desirable for former judgement to have less influence on the collective judgement and the latter to have more. Moreover, Yaniv 1997 and Wintle 2013 have found that individuals who produce more precise intervals tend to have lower absolute errors. Therefore, weighting the midpoints by the interval precisions should produce more accurate judgements. If the interval lengths, l_i, have been rescaled to a scale of 0 to 1, with 1 being the length of the most imprecise

interval for the given question, then one way of giving more weight to more precise judgements is the following weighted average:

$$\text{Precision-Weighted Average: } \mathcal{J} = \frac{1}{\mathbf{N}} \sum_{i=1}^{N} m_i(1 - l_i) \tag{2}$$

where $\mathbf{N}$ is the appropriate normalization factor for the weighted average: $\mathbf{N} = \sum_{i=1}^{N}(1 - l_i)$. Here, the precision of an interval with midpoint m_i is measured as $1 - l_i$.

At the same time, it seems that the probability that an individual assigns to their interval is also a relevant factor. Several studies have found small-to-moderate positive correlations between the probability assigned to a judgement and the accuracy of that judgement (e.g. Armstrong 1985, p. 138, Braun & Yaniv 1992, Lichtenstein *et al.* 1982, Wells & Murray 1984, Winkler 1971, Yates 1990). If the probabilities assigned to interval estimates also track the accuracy of the midpoints of the intervals, then the following aggregation method may prove effective:

$$\text{Probability-Weighted Average: } \mathcal{J} = \frac{1}{\mathbf{N}} \sum_{i=1}^{N} m_i p_i \tag{3}$$

where p_i is the probability (i.e., confidence level) assigned to the interval with midpoint m_i. This Probability-Weighted Average assigns more weight to the midpoints that come from intervals with higher probabilities assigned to them.

Ideally, however, it seems that the aggregation method should be sensitive to both pieces of information: it should take into account both the precision of the interval and the probability assigned to it. One way to do this is to combine the Precision-Weighted Average and the Probability-Weighted Average into what we will call an Certainty-Weighted Average:

$$\text{Certainty-Weighted Average: } \mathcal{J} = \frac{1}{\mathbf{N}} \sum_{i=1}^{N} m_i p_i (1 - l_i) \tag{4}$$

where again N is the appropriate normalization factor for the weighted average: $\mathbf{N} = \sum_{i=1}^{N} p_i(1 - l_i)$.

All three weighted averages take into account more of the available psychological information than the unweighted average does. Therefore, in so far as there is reason to think that this additional psychological information tracks the accuracy of the midpoints of the confidence intervals, there is reason to expect the weighted averages will outperform the unweighted average. Moreover, since the Certainty-Weighted Average takes into

account the most psychological information available, there is reason to think it will be the best performing method.

Although the Certainty-Weighted Average takes into account all of the *elicited* psychological information, it may, nevertheless, be possible to use this information to *infer* further psychological information and incorporate that into a more sophisticated weighting scheme. In particular, given a confidence interval that someone has assigned, it should be possible to make a rough approximation of the probabilities they would assign to other intervals if asked to do so. For example, someone who has assigned an 80% confidence to the interval 7% to 8% will probably think that the interval 8% to 9% is much more likely than the interval 20% to 21%. Such an inference will not always be true. For example, it might be that the individual has reason to think a bimodal distribution is appropriate and they assign a higher probability to the 20%-21% range than the 8%-9% range. However, if such cases are rare, then modeling the individuals with unimodal distributions may provide accurate enough inferences about their hypothetical judgements.

The reason why we mention this is that if we can infer a probability distribution from an individual's confidence interval, then we can weight the midpoint of their interval by the *Shannon entropy* of their distribution. By calculating the Shannon entropy, e_i, of their distribution, we may obtain a better assessment of how much information that person thinks they have about the question at hand. This is because the Shannon entropy of a distribution has proven to satisfy key constraints that any measure of information ought to satisfy (Shannon 1948).

There are many ways to infer a unimodal distribution from a confidence interval. However, to keep our analysis manageable, we will assume that confidence intervals are drawn from Gaussian distributions. Although we have little evidence for this assumption, it seems to be the simplest assumption that can be made. Putting these ideas together, we can weight the midpoints by the entropies inferred from the confidence intervals, with the most weight going to the lowest entropy and 0 weight going to the maximum entropy:

$$\text{Entropy-Weighted Average: } \mathcal{J} = \frac{1}{\mathbf{N}} \sum_{i=1}^{N} m_i(\max\{e_j\} - e_i) \tag{5}$$

where $\mathbf{N} = \sum_{i=1}^{N} \max\{e_j\} - e_i$, and

$$e_i = -\sum q_j ln(q_j)$$

and $\max\{e_j\}$ is the largest entropy for the judgement set, and q_j are the probabilities of the normal distribution inferred from the confidence interval with midpoint m_i. Since we

have reason to think e_i is a better measure of information than $m_i p_i (1 - l_i)$, we have reason to think the Entropy-Weighted Average will outperform the Certainty-Weighted Average. (For an alternative entropy-based weighting scheme, which relies on "intrinsic ranges" of the quantities to be estimated, see Cooke 1991 pp. 190–2.)

We now turn to some potential statistical improvements on the Unweighted Average. One potential problem with the Unweighted Average is that it can be a poor representation of the general location of a data set. For example, consider the two sets of judgements concerning GDP growth:

(i) $-0.1\%, 0.1\%, 0.2\%, -0.3\%, 0.1\%, 0.3\%, -0.3\%, 0.2\%, -0.1\%, -0.1\%$

(ii) $-19.1\%, 5.1\%, 5.2\%, 4.7\%, -20.5\%, 5.4\%, 4.7\%, 4.6\%, 4.8\%, 5.1\%$

In both cases, the average of the estimates is 0%, but the distributions of the guesses differ in an important way. The first set of estimates cluster around 0%, but the second tend to cluster more around 5% than they do around 0%. The only reason why the average of the second set of estimates is 0% is because of the two extremely negative estimates. In the first case, the individuals could agree to 0% as the collective judgement as a compromise—perhaps because 0% is so close to each individual estimate. However, in the second set, no one believes that the effect will be about 0%, so to take 0% as the collective judgement seems like a rather odd thing to do.

One way to obtain a better representation of the value that the data clusters around is to take the median of the data:

$$\text{Median Judgement: } \mathcal{J} = \text{median}(\{m_i\}) \tag{6}$$

The median of the second judgement set is 4.75%, and this seems to be a better measure of the location around which the judgements cluster.

Yet another way to improve on the mean is to simply first remove any outliers from the data set, and then take the unweighted average of the remainders (*cf.* Armstrong 2001, Jose and Winkler 2008). There are many methods for removing outliers, each with its distinct statistical properties. We will focus on one common method of outlier trimming: the Median Absolute Deviation (MAD) filtering method. Each midpoint m_i is determined to be an outlier and trimmed from the data set if:

$$m_i - \text{median}(\{m_j\}) / mdev > t \tag{7}$$

where

$$mdev = \text{median}(\{|m_k - \text{median}(\{m_h\})|\})$$

and t is a parameter that controls the sensitivity of the trimming (k and h are dummy variables). Smaller values of t tend to result in more of the m_i being counted as outliers than larger values of t. (From our empirical results (Section 4), we found the optimal value for t to be 2.) Let m_i^* be the remaining judgements. We then define the MAD Unweighted Average as:

$$\text{MAD Unweighted Average: } \mathcal{J} = \frac{1}{N}\sum_{i=1}^{N} m_i^* \tag{8}$$

The MAD Unweighted Average tends to be better than the Unweighted Average at measuring the location around which data sets cluster. For example, the MAD Unweighted Average (with $t = 2$) of the second set of GDP growth judgements is 4.95%. One potential drawback to this method—and any method that removes outliers—is that it has the potential to ignore strong dissenting voices. This can be a problem in situations of groupthink, where the process of forming a collective judgement neglects well-justified outlier opinions and is biased towards a consensus judgement, irrespective of the evidence for that judgement. (For more on the problem of groupthink, its causes, and methods for mitigating it, see Janis 1982, Taleb 2007, and Woodside 2012).

Statistics affords us many more sophisticated estimators (e.g., instead of the MAD filter, we could use the Dixon Outlier Q-Test, or we could a fit a distribution to the midpoint data and use a maximum likelihood estimator), however, we will leave our discussion here, for as we will show in the next section, the methods discussed above already serve our main point: the more complicated—and in this case—psychologically motivated aggregation methods do not outperform the simpler and statistically motivated methods. In fact, we will show that even combining the statistically motivated methods with the psychologically motivated ones also doesn't result in any method that is superior to the simple statistical methods. There are many such methods and, for simplicity, we restrict our focus to the following methods:

- MAD-Entropy-Filtered Entropy-Weighted Average
- MAD-Certainty-Filtered Certainty-Weighted Average
- MAD-Precision-Filtered Precision-Weighted Average

- MAD-Probability-Filtered Probability-Weighted Average
- MAD-Entropy-Filtered Unweighted Average
- MAD-Certainty-Filtered Unweighted Average
- MAD-Precision-Filtered Unweighted Average
- MAD-Probability-Filtered Unweighted Average
- MAD-Filtered Median Judgement
- MAD-Filtered Entropy-Weighted Average
- MAD-Filtered Certainty-Weighted Average
- MAD-Filtered Probability-Weighted Average
- MAD-Filtered Length-Weighted Average

For example, the MAD-Probability-Filtered Probability-Weighted Average takes the midpoints multiplied by their probabilities, applies a MAD outlier filter to those points, and then takes a probability weighted average.

The reason why we consider the psychologically motivated methods of this paper to be more complicated than the statistically motivated methods, is because, here, the former take more variables as input. We do not wish to imply that psychologically motivated methods are inherently more complex. Written functionally, and where $\mathcal{J}$ is the collective judgement, the statistically motivated methods have the form:

$$f(\{M_i\}) = \mathcal{J}$$

whereas the psychologically motivated methods have the form:

$$\begin{aligned} f(\{M_i\},\{l_i\}) &= \mathcal{J} \qquad \text{(E.g., the Precision-Weighted Average)} \\ f(\{M_i\},\{p_i\}) &= \mathcal{J} \qquad \text{(E.g., the Probability-Weighted Average)} \\ f(\{M_i\},\{l_i\},\{p_i\}) &= \mathcal{J} \qquad \text{(E.g., the Certainty-Weighted Average)} \end{aligned}$$

There are more dimensions to the complexity of functions, but this is at least one of them. Another aspect to the complexity of a function is its structural details—for example, in model selection theory, a linear model is considered more simple than a parabolic one, and so on (Akaike 1973, Forster and Sober 1994). For this reason, we consider the aggregation methods that are combinations of the psychologically and statistically motivated methods (listed in above) to be even more complicated.

3 Methods

To assess the relative performances of the aggregation methods, we conducted a random-effects meta-analysis (Cumming 2012, ch. 8) of the data collected from 15 experiments.

Each experiment consisted of 10 to 30 participants who were asked 5 to 30 questions about the true values of a range of different quantities. In each experiment, the set of questions was chosen so that all participants were either experts or they had at least a basic familiarity with the questions. Table 1 specifies the precise details of each experiment. In total, 311 questions were asked and 264 participants were involved (experiments 12 and 13, and 14 and 15 shared the same participants). All answers were in the form of confidence intervals elicited using either the 3- or 4-step method described in Section 2.

Inputs were elicited using a variant of the 4-step elicitation method developed by Speirs-Bridge *et al.* 2010, which reduces the overconfidence effect for confidence intervals (see also Soll and Klayman 2004 and Teigen and Jorgensen 2005 for similar methods). The Speirs-Bridge *et al.* 4-step method asks for a confidence interval in the following way:

> Given the evidence you have,
>
> (i) Realistically, what do you think is the lowest value that the unemployment rate could be?
>
> (ii) Realistically, what do you think is the highest value that the unemployment rate could be?
>
> (iii) Realistically, what is your best guess (i.e., most likely estimate)?
>
> (iv) How confident are you that the actual unemployment rate is between your lower and upper estimates?

Some of the experiments reported in the next section used a variant of this method, in which a best-guess estimate was not elicited. For convenience we will call the method that skips question (iii) the *3-step method* and the method described above the *4-step method*. (Note that this differs from the 3-step method as described by Speirs-Bridge *et al.* 2010 and Soll and Klayman 2004, which asks for (i) the lower bound, (ii) the upper bound, and (iii) the best guess; at a fixed level of confidence.)

The 15 experiments were originally conducted to test other hypotheses. Experiments 1–6 were conducted to see if participant expertise correlates with judgement accuracy. See Burgman, McBride, Ashton, Speirs-Bridge, Flander, Wintle, Fidler, Rumpff, Twardy 2011 for further details. Experiments 7–15 were conducted to see if overconfidence in confidence interval judgements can be reduced by having participants assign probabilities to each other's confidence intervals. Initial results from experiment 7 (the first experiment in the series) suggest that this is in fact the case (Lyon, Fidler, Burgman 2012). Our purpose

Experiment	N	Q	Elicitation	Question Topic(s)
1	21	10	4-step	Animal and plant biosecurity and quarantine
2	24	10	4-step	Animal and plant biosecurity
3	13	8	4-step	Ecology, frog biology
4	25	5	4-step	Public health, medicine
5	20	6	4-step	Risk analysis, biosecurity
6	14	8	4-step	Weed Ecology
7	22	30	3-step	General knowledge and the history of philosophy
8	24	24	4-step	Conservation Biology
9	30	30	3-step	Birth years of famous figures in probability and statistics
10	13	30	3-step	Birth year of famous historical figures
11	24	30	3-step	General knowledge and the history of philosophy
12	17	30	3-step	Birth year of famous historical figures
13	17	30	3-step	Birth year of famous historical figures
14	17	30	3-step	Birth year of famous historical figures
15	17	30	3-step	Birth year of famous historical figures

Table 1: Details of the 15 experiments, listed in chronological order. N = number of participants, Q = number of questions.

here, however, is to present a meta-analysis of the data collected in these 15 experiments to study the relative performances of the aggregation methods described in the previous section.

Since the experiment questions covered a wide range of very different quantities, we chose to rescale the participants responses to a $[0, 1]$ scale, so that that the performances of the aggregation methods can be compared across the questions. For a given question, each participant response, j, was rescaled using the following formula:

$$j^* = \frac{j - x_{\min}}{x_{\max} - x_{\min}}$$

where $x_{\min}$ is the smallest lower bound of the confidence intervals given in response to the question, and $x_{\max}$ is their largest upper bound. This range coding ensures that the answers to each question contribute roughly equally to the overall assessment of accuracy, which was measured in terms of the Absolute Error.

Since the Unweighted Average is the baseline by which we compare the other aggregation methods, we chose our effects to be the differences between the aggregation methods and the Unweighted Average. For example, where $J^*_{Unweighted}$ is the range-coded judgement given by the Unweighted Average, and $J^*_{Precision}$ is the range-coded judgement given by the Precision-Weighted Average, and T^* is the range-coded true value of the quantity in question, then the effect for the Precision-Weighted Average is:

$$\text{Effect for Precision-Weighted Average } = |T^* - J^*_{Unweighted}| - |T^* - J^*_{Precision}|$$

where a positive value means that the Precision-Weighted Average was more accurate than the Unweighted Average, and a negative value means it was less accurate.

4 Results

As expected, the Precision-Weighted average tended to be more accurate than the Unweighted Average (Figure 1). This result is similar to that of Yaniv 1997 who found that the average weighted by the *inverse* of the confidence interval lengths was more accurate than the Unweighted Average. (We also looked to see which weighting scheme was more accurate and found that the Precision-Weighted Average was more accurate than the average weighted with the inverses of the confidence interval lengths.)

There was also a slight improvement in the Probability-Weighted Average over the Unweighted Average. However, by combining the probability weights with the precision

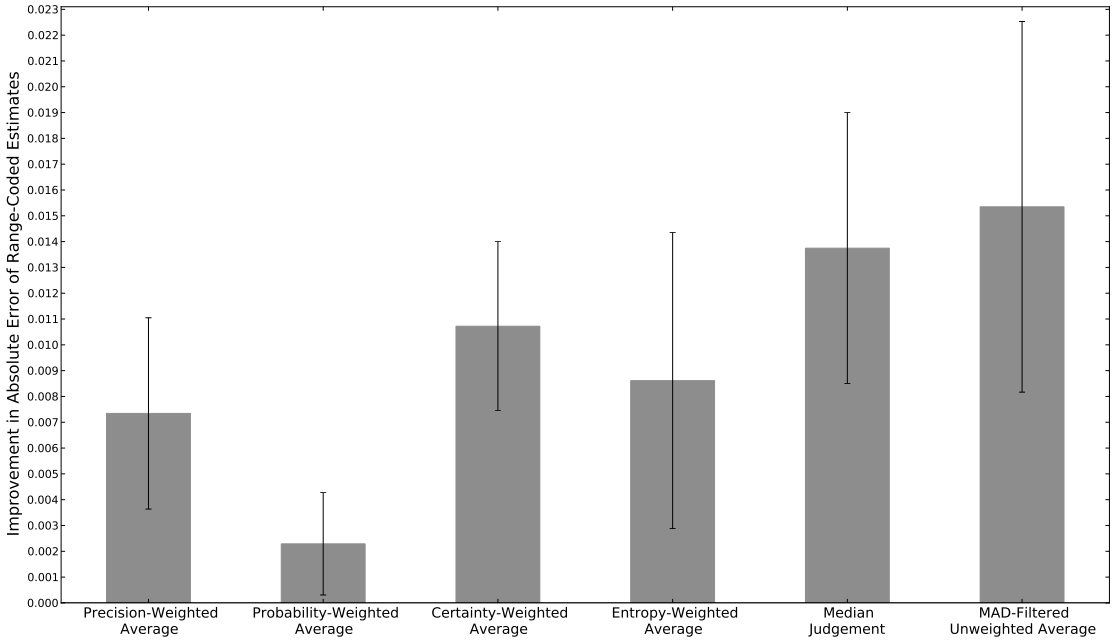

Figure 1: Improvements of the aggregation methods over the Unweighted Average. Error bars are 95% CIs.

weights to form the Certainty-Weighted Average, we found there was an improvement over the Unweighted Average, Precision-Weighted Average, and the Probability-Weighted Average. (Pairwise comparisons produced no overlapping 95% CIs.) We expected this result because the Certainty-Weighted Average takes into account the most psychologically relevant information, and there is reason to think this information tracks accuracy (*cf.* Section 2). As explained in Section 2, there is reason to expect that the Entropy-Weighted Average might be an improvement over the Certainty-Weighted Average. However, this proved not to be the case. The Entropy-Weighted Average proved to be slightly less accurate than the Certainty-Weighted Average (although this wasn't statistically significant), and its improvements over the Unweighted Average were substantially more variable (Figure 1).

Although the methods motivated by psychological considerations were more accurate then the Unweighted Average, they were not more accurate than the methods motivated by statistical considerations. The Median Judgement and the MAD-Filtered Unweighted Averaged were both slightly more accurate than the Certainty-Weighted Average (although this wasn't statistically significant). As we explained earlier, the MAD-Filtered Unweighted

Average first trims any outliers from the data set before the midpoints are averaged. The amount of trimming is controlled by the parameter t (in Equation 7) and we found the optimal value to be $t = 2$, which resulted in about 25% of the judgements being ignored by the method for each question (mean = 25.5, 95% CI = [24.3, 26.7]). Although the optimal value was $t = 2$, we found that the result that none of the psychologically motivated methods were more accurate than the MAD-Filtered Unweighted Average to be quite robust—from $t = 0.51$ (approximately 70% of judgements being trimmed) to $t = 5.5$ (approximately 8% judgements being trimmed). (This is perhaps not surprising given that the psychologically motivated methods also didn't beat the Median Judgement, which is also robust against outliers.)

Figure 2 shows two forest plots comparing the MAD-Filtered Unweighted Average and the Certainty-Weighted Average (the best-performing statistically-motivated method and the best-performing psychologically-motivated method). Both effects were positive in almost every experiment. There were only two exceptions. In experiment 4, both effects were negative; however, given that the experiment consisted of only 5 questions, we think it is likely that this anomaly is due to sample noise. In experiment 6, the Certainty-Weighted Average effect was negative; and again, we think this probably due to sample noise, since there were only 8 questions. The MAD-Filtered Unweighted Average was slightly more accurate than the Certainty-Weighted Average in the meta analysis results; however, this was not a consistent trend across the experiments and the difference was not statistically significant.

Because our results are formulated in terms of errors of the range-coded responses, it can be difficult to get a sense of what these effects mean in practical terms. However, experiments 9, 10, 12, 13, 14, and 15 all asked questions of the same form: the birth year of a historical figure. So for these experiments, we can report the performances of the aggregation methods in terms of how many years they missed the true birth year by. The graph in Figure 3 shows the mean improvements in absolute year error over the Unweighted Average for the Certainty-Weighted Average and the MAD-Filtered Unweighted Average. The MAD-Filtered Unweighted Average outperformed the Unweighted Average in every experiment (although there was a lack of statistical significance in experiments 9 and 14), and it resulted in an approximate 10 to 60 year mean improvement over the Unweighted Average, whose mean errors ranged from approximately 30 to 135 years.

We found that it wasn't possible to improve upon the MAD-Filtered Unweighted Average with statistical significance by incorporating the precisions and probabilities and

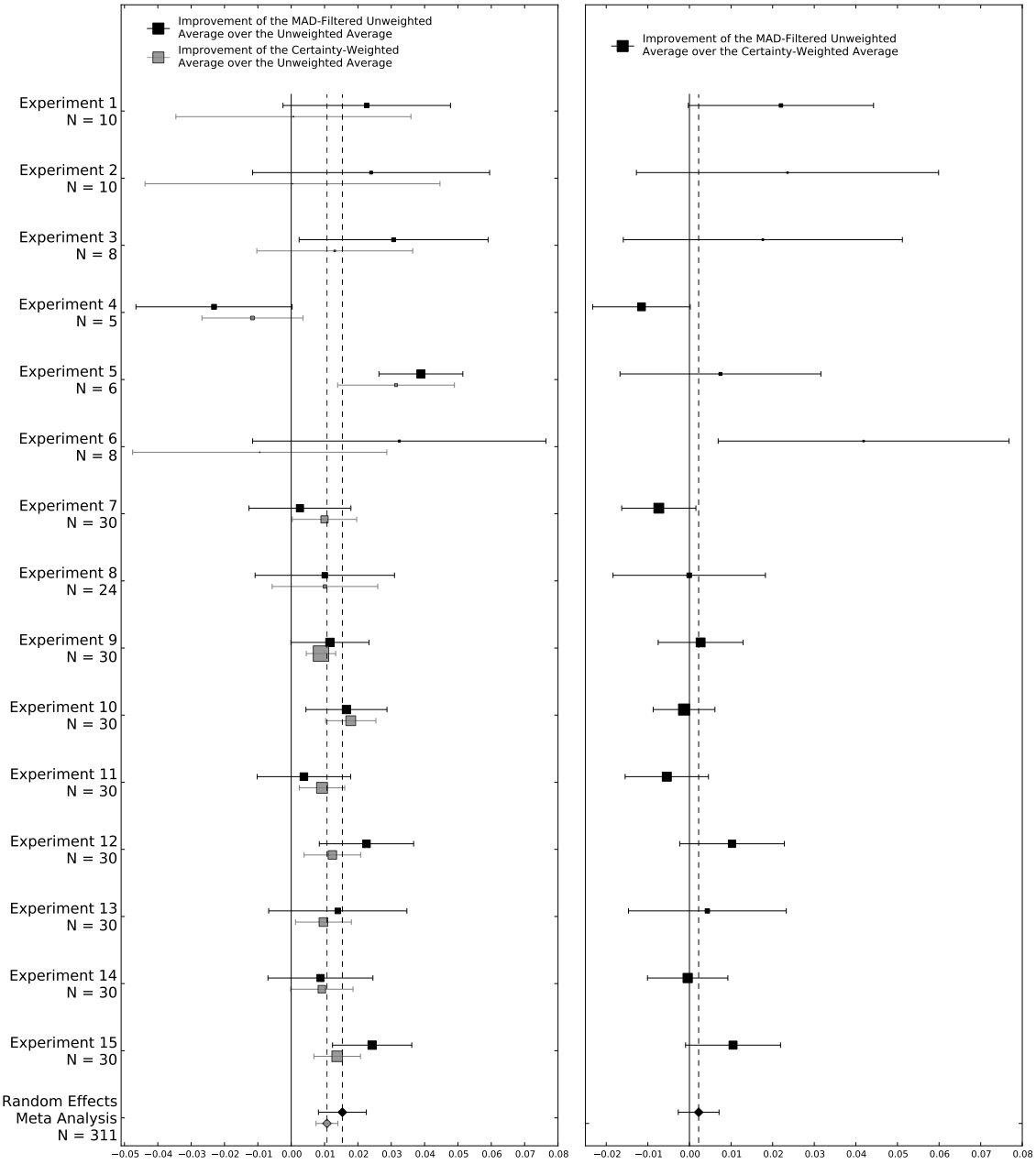

Figure 2: Two forest plots comparing the Certainty-Weighted and the MAD-Filtered Unweighted Average. The plot on the left shows the improvements of the two aggregation methods over the Unweighted Average. The plot on the right shows the improvement of the MAD-Filtered Unweighted Average over the Certainty-Weighted Average. The marker size of an effect corresponds (proportionally) to the weight assigned to that effect by the random effects meta analysis (*cf.* Cumming 2012, Ch. 8.). Error bars are 95% CIs.

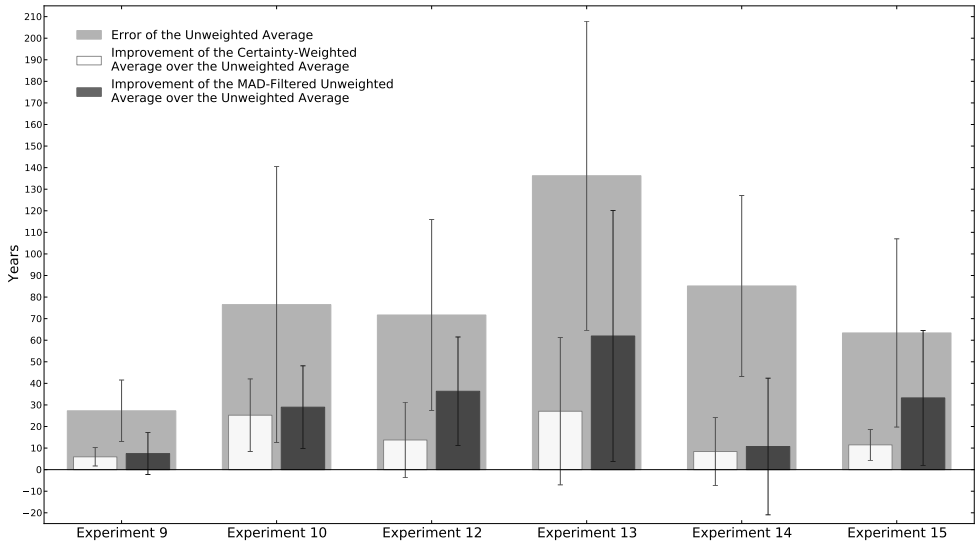

Figure 3: Improvements of the Certainty-Weighted Average and MAD Filtered Unweighted Average over the Unweighted Average for experiments that only asked "Year of Birth" questions. Effect sizes are measured in years, and error bars are 95% CIs. Results are reported in terms of the *errors* of the Unweighted Average and the *improvements* of the Certainty-Weighted Average and the MAD-Filtered Unweighted Average over the Unweighted Average.

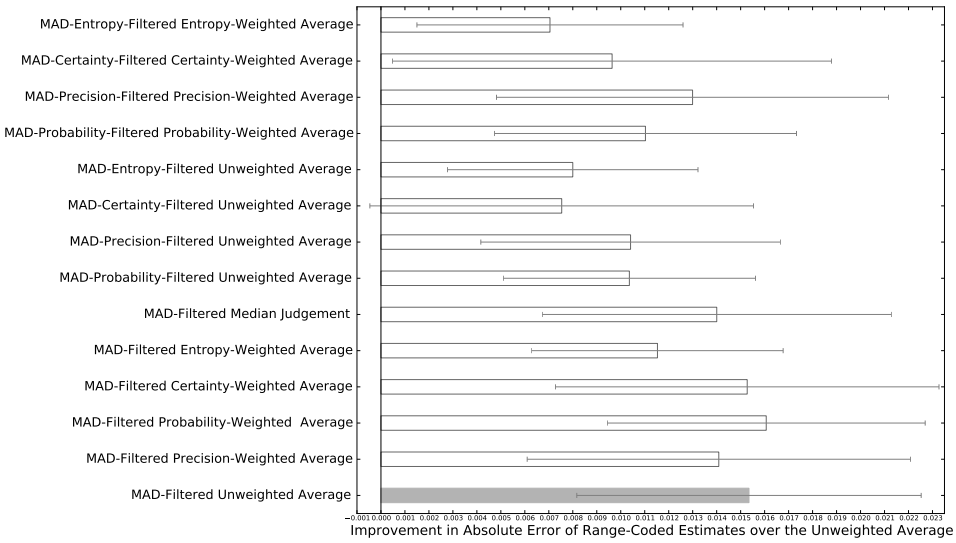

Figure 4: The more complicated methods, which combined MAD filtering with the median, weighted averages, and MAD filtering of weighted midpoints, did not prove to be more accurate than the MAD-Filtered Unweighted Average. Example: The method *MAD-Entropy-Filtered Entropy-Weighted Average* first applies the MAD filter (with $t = 2$) of the $m_i(\max\{e_j\} - e_i)$ and then takes the Entropy-Weighted Average of the remaining midpoints.

making even more complicated weighted averages. For example, we expected that if after filtering out the outliers, we then took a Certainty-Weighted Average—thus producing a new aggregation method which we call the *MAD-Filtered Certainty-Weighted Average*—we should be able to improve upon the MAD-Filtered Unweighted Average. However, this turned out not to be case. All of the more complicated methods that we considered (listed in Figure 4) were not more accurate than the the MAD-Filtered Unweighted Average, with the exception of the MAD-Filtered Probability-Weighted Average, which was slightly more accurate, but the difference was not statistically significant.

5 Discussion

It is somewhat surprising that the psychologically motivated methods didn't outperform the statistically motivated methods. The statistically motivated methods dismiss seemingly important information—*viz.*, the probabilities and precisions of the confidence intervals.

That this is important information is confirmed by the result that both the precision and probability weighted averages were better than the unweighted average, and that the confidence weighted average (which takes into account both the precisions and probabilities) was better yet again. (Pairwise comparisons yielded no overlapping 95% CIs.)

Even more surprising was that the Certainty-Weighted Average still did not beat the MAD-Filtered Unweighted Average even when it was applied to the MAD-Filtered judgement set. It appears that once outliers have been removed, the extra information used by the Certainty-Weighted Average—that is, the precisions and probabilities—is of little value (the MAD-Filtered Probability-Weighted Average was ever-so slightly more accurate than the MAD-Filtered Unweighted Average, and this difference was not statistically significant). This extra information appears to be only of value when the outliers have not been removed from the judgement set.

It is important to stress that results such as these can be sensitive to how accuracy is measured. We chose to measure accuracy in terms of the absolute error of the range-coded estimates. However, there are many other ways to measure accuracy and each has its pros and cons (see e.g., Armstrong and Collopy 1992, Armstrong and Fildes 1995, Hyndman and Koehler 2006). We lack the space to systematically explore the results in terms of different error measures. However, we can briefly report two main findings with respect to these issues. First, the main results were very similar when accuracy was measured in terms of absolute percentage error and log-ratio error. Second, the results were quite different when accuracy was measured in terms of squared error (of the range-coded estimates). This resulted in the Certainty-Weighted Average being the only method that was more accurate than the Unweighted Average with statistical significance. (A major difference between the squared error measure and the absolute error, absolute percentage error, and the log-ratio error measures is that the former penalizes a method heavily for the occasional judgement that is far from the observed value.)

All of the studied methods average the midpoints of the elicited confidence intervals. Averaging the midpoints of confidence intevals and looking at the effects of weighting and trimming has been studied also by Yaniv 1997; however, in many situations best guess estimates are also elicited and it makes sense to average those instead of the midpoints. Seven of the experiments described above elicited best guess estimate—that is, those that used the 4-step elicitation method. A meta-analysis on those experiments that replaced the midpoints with the best guess estimates produced results similar to those reported above. However, since there were only seven experiments and these experiments had

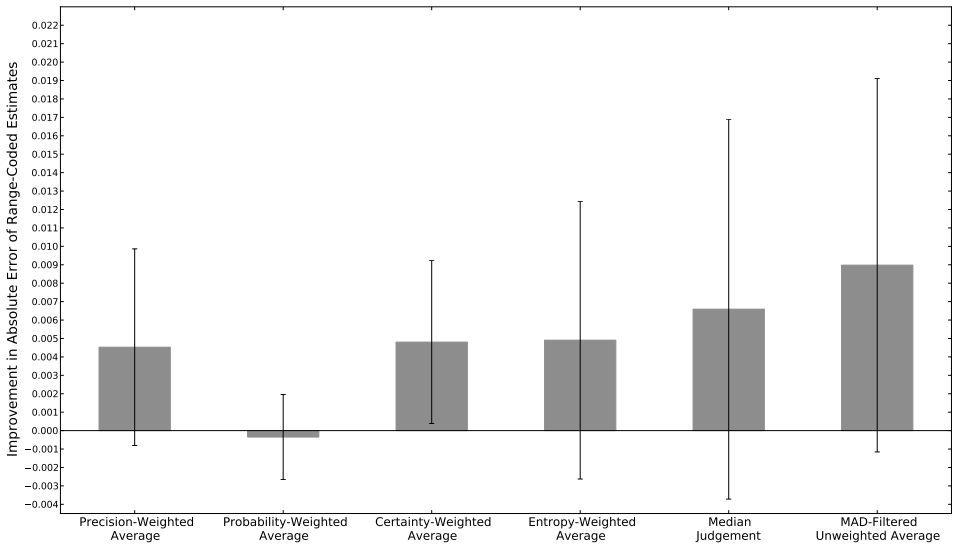

Figure 5: Improvements of the aggregation methods over the Unweighted Average using the best guess estimates from experiments 1, 2, 3, 4, 5, 6, and 8. Error bars are 95% CIs.

few questions (Table 1), the 95% CIs surrounding the effect sizes are wide (see Figure 5). Further work needs to be done to see if the main results of this paper transfer to the aggregation of best guess estimates.

Our results have implications for choosing an appropriate method for eliciting judgements. Much work has been done in developing elicitation methods that reduce the overconfidence effect for confidence intervals (see e.g., Soll and Klayman 2004, Teigen and Jorgensen 2005, and Speirs-Bridge *et al.* 2010). Naturally, these methods are appropriate when eliciting confidence intervals for a collective wisdom task, and one might expect that the information contained in such intervals could be used to improve the aggregated 'best estimate'. However, our results suggest that if one's goal is to simply obtain accurate point estimates, then efforts might be better spent on elicitation methods that improve the accuracy of estimates, rather than those that reduce the overconfidence effect. For example, it has been shown that by repeatedly asking for a single number estimate from an individual and then taking the average of those estimates, one obtains a more accurate estimate. This has been referred to as the *crowd-within* effect (see e.g., Vul and Pashler 2008, Herzog and Hertwig 2009, Hourihan and Benjamin 2010). If accuracy of point estimates is the priority, our results suggest that it might be worthwhile to use such an

elicitation method—rather than one of the 3- or 4-step elicitation methods—and combine the estimates with a statistical aggregation technique that trims outliers. However, to make this conclusion, another study that directly compares the different elicitation methods is required.

6 Conclusion

We studied the accuracies of a range of methods for combining confidence interval inputs into point-estimate outputs. The aggregation methods we studied fell into two kinds: those motivated by psychological considerations and those motivated by statistical considerations. We found that a simple trim-and-average method—that is, average interval midpoints after outliers have been removed—produced estimates that were more accurate than the unweighted average. Our results show that other, more complicated methods, which factor in psychologically relevant information such as confidence levels and estimate imprecision, do not produce estimates that are reliably more accurate than those produced by the simple trim-and-average method.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *2nd International symposium on information theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971*, pp. 267. Akademiai Kiadó.

Alpert, M. and H. Raiffa (1982). A progress report on the training of probability assessors. *Judgment under uncertainty: Heuristics and biases*, 294–305.

Armstrong, J. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (1989). Combining forecasts: The end of the beginning or the beginning of the end? *International Journal of Forecasting 5*(4), 585–588.

Armstrong, J. S. and L.-R. Forecasting (1985). From crystal ball to computer. *Wiley, New York*.

Armstrong, S. and F. Collopy (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting 8*(1), 69–80.

Armstrong, S. and R. Fildes (1995). Correspondence on the selection of error measures for comparisons among forecasting methods. *Journal of Forecasting 14*(1), 67–71.

Braun, P. A. and I. Yaniv (1992). A case study of expert judgment: Economists' probabilities versus base-rate model forecasts. *Journal of Behavioral Decision Making 5*(3), 217–231.

Burgman, M., M. McBride, R. Ashton, A. Speirs-Bridge, L. Flander, B. Wintle, F. Fidler, L. Rumpff, and C. Twardy (2011). Expert Status and Performance. *PLoS One 6*(7), e22998.

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

Clemen, R. T. and R. L. Winkler (1999). Combining probability distributions from experts in risk analysis. *Risk analysis 19*(2), 187–203.

Clements, M. P. and D. I. Harvey (2011). Combining probability forecasts. *International Journal of Forecasting 27*(2), 208–223.

Condorcet, M. (1785). *Essai sur l'application de l'analyse á la probabilité des décisions rendues á la pluralité des voix.* Paris: l'Imprimerie Royale. Reprint. New York: Chelsea, 1972.

Cooke, R. M. (1991). Experts in uncertainty: opinion and subjective probability in science.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge New York.

Fischer, I. and N. Harvey (1999). Combining forecasts: What information do judges need to outperform the simple average? *International journal of forecasting 15*(3), 227–246.

Forster, M. and E. Sober (1994). How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science 45*(1), 1–35.

Genest, C. and K. J. McConway (1990). Allocating the weights in the linear opinion pool. *Journal of Forecasting 9*(1), 53–73.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

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

Hourihan, K. L. and A. S. Benjamin (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition 36*(4), 1068.

Hyndman, R. J. and A. B. Koehler (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting 22*(4), 679–688.

Janis, I. L. (1982). Groupthink: A psychological study of foreign policy decisions and fiascoes.

Jose, V. R. R. and R. L. Winkler (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting 24*(1), 163–169.

Kahneman, D., P. Slovic, and A. Tversky (1982). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Ladha, K. K. (1992). The Condorcet Jury Theorem, Free Speech, and Correlated Votes. *American Journal of Political Science*, 617–634.

Larrick, R. P. and J. B. Soll (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science 52*(1), 111–127.

Lichtenstein, S., B. Fishhoff, and L. D. Phillips (1982). Calibration of Probabilities: The State of the Art to 1980. In *Judgment under Uncertainty: Heuristics and Biases*. University of Cambridge.

List, C. and R. E. Goodin (2002). Epistemic Democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy 9*(3), 277–306.

Lyon, A., F. Fidler, and M. Burgman (2012). Judgement swapping and aggregation. In *2012 AAAI Fall Symposium Series*.

Nielsen, M. (2011). *Reinventing Discovery: The New Era of Networked Science*. Princeton University Press.

Page, S. (2008). *The Difference*. Princeton University Press.

Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal 27*, 379–423.

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

Speirs-Bridge, A., F. Fidler, M. McBride, L. Flander, G. Cumming, and M. Burgman (2010). Reducing Overconfidence in the Interval Judgments of Experts. *Risk Analysis 30*(3), 512–523.

Surowiecki, J. (2004). *The Wisdom of Crowds*. Doubleday.

Taleb, N. N. (2007). *The Black Swan:: The Impact of the Highly Improbable Fragility*. Random House Digital, Inc.

Teigen, K. and M. Jørgensen (2005). When 90% Confidence Intervals are 50% Certain: On the Credibility of Credible Intervals. *Applied Cognitive Psychology 19*(4), 455–475.

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

Wallsten, T. S., D. V. Budescu, I. Erev, A. Diederich, et al. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making 10*(3), 243–268.

Wells, G. L. and D. M. Murray (1984). Eyewitness confidence. *Eyewitness testimony: Psychological perspectives*, 155–170.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association 66*(336), 675–685.

Wintle, B. C. (2013). *Making Decisions When Estimates Conflict: Improving Judgements in Environmental Science*. Ph. D. thesis, School of Botany, University of Melbourne.

Woodside, A. G. (2012). Incompetency Training: Theory, Practice, and Remedies. *Journal of Business Research 65*(3), 279–293.

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes 69*(3), 237–249.

Yaniv, I. and D. P. Foster (1997). Precision and accuracy of judgmental estimation. *Journal of behavioral decision making 10*(1), 21–32.

Yates, J. F. (1990). *Judgment and decision making*. Prentice-Hall, Inc.