AAAI Technical Report FS-12-06
Machine Aggregation of Human Judgment

Judgement Swapping and Aggregation

Aidan Lyon
Department of Philosophy
University of Maryland
College Park, MD, 20842, USA
alyon@umd.edu

Fiona Fidler
Australian Centre of Excellence for Risk Analysis
University of Melbourne
Melbourne, VIC, 3010, Australia

Mark Burgman
Australian Centre of Excellence for Risk Analysis
University of Melbourne
Melbourne, VIC, 3010, Australia

Abstract

We present the results of an initial experiment that indicates that people are less overconfident and better calibrated when they assign confidence levels to someone else's interval judgements (evaluator confidences) compared to assigning confidence levels to their own interval judgements (judge confidences). We studied what impact this had on a number of judgement aggregation methods, including linear aggregation and maximum confidence slating (MCS). Using evaluator confidences as inputs to the aggregation methods improved calibration, and it improved hit rate in the case of MCS.

Introduction

It's well known that groups can outperform individuals on a variety of judgement tasks. For example, for a quantity estimation task (e.g., guess the number of jellybeans in a jar), the straight average of a set of individual estimates can outperform most, or even all, of the individual estimates (Surowiecki (2004)). But there are many methods of determining a group judgement from the group's individual judgements, and it is natural to ask which methods perform best, and under which circumstances. Koriat (2012b) has recently shown that the method of choosing the individual judgement with highest confidence—known as *maximum confidence slating* (MCS)—can perform better than individual judgements, for binary judgement tasks. Can the same method work well for interval judgement tasks?

One reason to think that MCS won't work as well for interval judgement tasks is that the overconfidence effect can be much stronger for interval judgements tasks than it is for binary judgement tasks (e.g., Klayman *et al.*, (1999)). A plausible explanation for the success of MCS for binary judgement tasks is that high confidence in a binary judgement by an individual is a signal of relevant knowledge possessed by the individual (Koriat (2012a)). However, this would mean that high confidence by an individual in an interval judgement would be less of a signal of the individual possessing relevant knowledge, because of the strong overconfidence effect for interval judgement tasks.

Fortunately, the overconfidence effect for interval judgement tasks can be reduced if certain judgement elicitation methods are used. One way in which overconfidence in interval judgements can be substantially reduced is if individuals are asked to give confidences to preset intervals, rather than produce interval judgements for preset confidence levels. Teigen and Jørgensen (2005) found that individuals appeared to be massively overconfident (44%–67%) when they performed the former kind of task, and much less overconfident (12%–39%) when they performed the latter kind of task. Winman *et al.* (2004) also found a similar effect. In other words, the task of *evaluating* an interval judgement appears to result in substantially less overconfidence than the task of *producing* an interval judgement for a given level of evaluation.[1]

In the context of eliciting expert opinions, however, it can be undesirable to give experts preset intervals for them to assign confidence levels to. For example, experts are likely to be asked opinions about highly uncertain outcomes and one needs to be an expert to even know a reasonable range within which to set the intervals, or there may be widely varying opinions about what a reasonable range is. In many situations, it would be better if the experts produced the interval judgements themselves. One way in which the advantages of using experts to produce interval judgements can be obtained while also reducing the overconfidence effect is if one group of experts are asked to generate interval judgements and another group is asked to evaluate those judgements. In addition to the results from Teigen and Jørgensen (2005) and Winman *et al.* (2004), one reason to think that this would be effective is that when one expert is assigning a level of confidence to another's interval judgement, that expert can be in

[1] This is certainly not the only method of elicitation by which overconfidence can be reduced. See e.g., Speirs-Bridge *et al.* (2010).

a better position to objectively evaluate the judgement than the expert who produced the judgement (Talyor and Brown (1988)). This method of splitting the experts up into producers and evaluators should therefore produce better calibrated judgements.

However, Koehler (1994) showed that evaluators ("observers" in his terminology) can be more overconfident than judges ("actors"). This effect disappeared, though, when it was ensured that the judges and evaluators both considered a closed set of alternative hypothesis (*ibid.*). Koehler and Harvey (1997) argue that in producing their judgements, judges consider alternative hypotheses and these remain in their minds when they assign their levels of confidences to their judgements—and such consideration of alternatives seems to reduce overconfidence. Evaluators on the other hand are simply presented with the judgements and asked to assign confidences to them. Evaluators were therefore not primed to consider alternative hypothesis, whereas the judges were. In support of this explanation, Koehler and Harvey (1997) report that the judges were more overconfident than the evaluators when a distractor task was inserted between the production of the judgement and the judge's evaluation of their own judgment.

In the context of expert judgement elicitation, specifying a closed set of alternatives is just as undesirable as specifying pre-set intervals for the experts to evaluate (and distracting the experts to make them more overconfident is even less desirable!). However, a plausible mechanism by which the evaluators can be prompted to consider alternative hypotheses is to make them produce their own judgements before evaluating someone else's. Instead of splitting experts up into judges and evaluators, it seems it would be better to get every expert to make a judgement and then have them swap their judgement with someone else for evaluation (i.e., confidence assignment). In other words, swapping judgments may reduce overconfidence and improve calibration.

One way to test this is to elicit interval judgements and confidences from a group of individuals and have them swap their interval judgements with each other and assign new levels of confidences of them. Call the confidence assigned to a judgement by the individual who made the judgement a *judge confidence*, and call the confidence assigned to a judgement by an individual who only evaluates the judgement an *evaluator confidence*. The experiment reported in this paper was conducted to see if evaluator confidences are less overconfident and better calibrated than judge confidences.

Eliciting two confidence levels for every judgement opens up the possibility of two variants of MCS. MCS, as used by Koriat (2012b), involves selecting the judgement with maximum confidence assigned by the person who made the judgement, i.e., the *judge*. (Koriat doesn't explain what the method does if there is more than one judgement with maximum confidence. For our purposes, we'll take MCS for interval judgements to be the method of selecting the judgement with maximum confidence if there is only one such judgement, and if there are ties, the method will select the linear average of the ties.) However, we could instead select the judgement that is assigned maximum confidence by the person who only evaluated the judgement, i.e., the *evaluator*. Call these two versions of MCS, *maximum judge confidence slating* (MJCS) and *maximum evaluator confidence slating* (MECS). If evaluator confidences are better calibrated than judge confidences, then it seems that MECS should perform better than MJCS.

There is an important difference between binary judgements and interval judgements when it comes to measuring uncertainty. A 90% confidence level in a binary judgement is quite a low level of uncertainty, but a 90% confidence level in an interval judgement is not necessarily a low level of uncertainty. This is because an interval judgement, unlike a binary judgement, allows two expressions of uncertainty—a confidence level and the interval width, or precision, of the estimate. Even when the confidence level is high (e.g., 90%) the interval may be wide (imprecise), suggesting a high level of uncertainty about the estimate—e.g., consider the judgement that Barack Obama was born somewhere between 2000 BC and 2000 AD, and a 90% confidence level in this judgement. The width of an interval judgement should therefore be taken into account. Call the method that chooses the judgement with the highest value of confidence level divided by judgement width, *maximum information slating* (MIS). (In the case of ties, the method takes the linear average of the ties.) There are (at least) two variants of MIS in the context of judgement swapping: *maximum judge information slating* (MJIS), and *maximum evaluator information slating* (MEIS). For reasons similar to those given above, if evaluator confidences are better calibrated than judge confidences, then MEIS should produce better judgements than MJIS.

By eliciting two confidence levels for every judgment, it may also be possible to detect signals of knowledge of knowledge and knowledge of the lack of knowledge, and thereby improve calibration. Consider the method that chooses the judgement for which the judge and the evaluator's confidences are in most *agreement* (i.e., have minimal absolute difference); call this method *maximum consensus slating* (MConS). (In the case of ties, the method takes the linear average of the ties.) If the judge and the evaluator agree on a confidence level for a given judgement, then that's a sign that the confidence is appropriate for the judgement. Therefore, MConS may tend to produce well calibrated judgements.

In this paper, we report the results of an initial experiment to test these hypotheses. We examine if evaluator confidences tend to be less overconfident and better calibrated than judge confidences, and we examined whether this holds at the group level, by comparing the *linear aggregation of judgements paired with average judge confidences* (LinAgg-J) with the *linear aggregation of judgements with average evaluator confidences* (LinAgg-E). We also compared the performances of MJCS, MECS, MJIS, MEIS, and MConS. As explained above, these are all methods for determining group *judgements*, and so each group judgement determined by each method can be paired with the corresponding judge confidence or the evaluator confidence. There are, therefore, at least two variants of each slating method: one that pairs the method's judgements with judge confidences and one

that pairs the judgements with the evaluators' confidences—e.g., for MJCS, there is MJCS-J and MJCS-E. We examine how each aggregation method performs with respect to hit rate, overconfidence, calibration, and accuracy.

Methods

Experiment

Thirty individuals were invited to participate in the experiment and 8 didn't respond, declined, or didn't show up. The remaining 22 participants were a mixture of undergraduate and graduate students in philosophy, and philosophy faculty. Participants were mostly male (approximately 80%) and North American.

They were asked to give 30 interval judgements about the true values of a range of quantities. Some questions involved facts about the history of philosophy that the participants were expected to have some knowledge about (e.g., how many books a famous philosopher had published; in what year a famous philosopher was born), and some questions were general knowledge type questions (e.g., how many tracks are there on the Beatles "1" album; what was Apple Inc.'s closing share price the day before the experiment).

Judgements were elicited by asking separately for a (i) lower estimate, and (ii) an upper estimate. Individuals were also asked to express confidences in these interval judgements—i.e., the probability the true value is between the values elicited from questions (i) and (ii). Confidences were allowed to be expressed on the full 0% to 100% scale. These confidences are the *judge confidences*, because they were assigned by those who gave the judgements. After this judgement and confidence elicitation process was completed, the judge confidences were removed from the judgements, and everyone was given someone else's 30 judgements (randomly and anonymously). Everyone then assigned their own confidences to the 30 judgments that they received. These confidences are the *evaluator confidences*.

(For convenience we will talk as though there were two groups: judges and evaluators, even though everyone acted both as a judge and as an evaluator.)

Aggregation Methods

For question q, let A_q be the set of answers to that question that were elicited from the group. Each answer, $a_n (n = 1, ..., 22)$, has three components: the interval judgement $a_n^{(i)}$, the judge confidence in that interval judgement $a_n^{(Jc)}$, and the evaluator confidence in the interval judgement $a_n^{(Ec)}$. Each aggregation method results in an answer with three components ($\chi \in \{i, Jc, Ec\}$) and are defined as follows:

Linear Aggregation:

$$\text{LingAgg}(A_q)^{(\chi)} = \frac{1}{|A_q|} \sum_{a_n \in A_q} a_n^{(\chi)}$$

Maximum Judge Confidence Slating:

$$A_q^{Jc*} = \{a_n \in A_q : a_n^{(Jc)} = \max_{a_n^{(Jc)}}(A_q)\}$$

$$\text{MJCS}(A_q)^{(\chi)} = \frac{1}{|A_q^{Jc*}|} \sum_{a_n \in A_q^{Jc*}} a_n^{(\chi)}$$

Maximum Evaluator Confidence Slating:

$$A_q^{Ec*} = \{a_n \in A_q : a_n^{(Ec)} = \max_{a_n^{(Ec)}}(A_q)\}$$

$$\text{MJCS}(A_q)^{(\chi)} = \frac{1}{|A_q^{Ec*}|} \sum_{a_n \in A_q^{Ec*}} a_n^{(\chi)}$$

Maximum Judge Information Slating:

$$A_q^{\text{JI}*} = \{a_n \in A_q : a_n^{(Jc)}/||a_n^{(i)}|| = \max_{a_n^{(Jc)}/||a_n^{(Ji)}||}(A_q)\}$$

$$\text{MJIS}(A_q)^{(\chi)} = \frac{1}{|A_q^{JI*}|} \sum_{a_n \in A_q^{JI*}} a_n^{(\chi)}$$

Maximum Evaluator Information Slating:

$$A_q^{\text{EI}*} = \{a_n \in A_q : a_n^{(Ec)}/||a_n^{(i)}|| = \max_{a_n^{(Ec)}/||a_n^{(Ei)}||}(A_q)\}$$

$$\text{MEIS}(A_q)^{(\chi)} = \frac{1}{|A_q^{EI*}|} \sum_{a_n \in A_q^{EI*}} a_n^{(\chi)}$$

Maximum Consensus Slating:

$$A_q^{\text{Con}} = \{a_n \in A_q : |a_n^{(Jc)} - a_n^{(Ec)}| = \min_{|a_n^{(Jc)} - a_n^{(Ec)}|}(A_q)\}$$

$$\text{MConS}(A_q)^{(\chi)} = \frac{1}{|A_q^{Con}|} \sum_{a_n \in A_q^{Con}} a_n^{(\chi)}$$

Performance Measures

The hit rate of a set of interval judgements is defined as the percentage of occasions the interval judgements include the true values of the quantities in question. Overconfidence is measured as the difference between the average of the confidence level and hit rate—a positive difference indicates overconfidence, a negative difference indicates underconfidence. For a given set of judgements, there are two under/overconfidences: the under/overconfidence determined of the judge confidences and the under/overconfidence of the evaluator confidences.

Since the questions ranged over a variety of quantities, a standardised measure of accuracy was used. First, each answer for a given question q was range coded:

$$\overline{a}_{n,q}^{(i)} = \frac{\text{mid}(a_n^{(i)}) - a_n^{\min}}{a_n^{\max} - a_n^{\min}}$$

where $\text{mid}(a_n^{(i)})$ is the mid-point of the interval $a_n^{(i)}$, $a_n^{\min}$ is the minimum value of $a_n^{(i)}$ in A_q or the truth if the truth was less than this minimum value, and $a_n^{\max}$ is the maximum value of $a_n^{(i)}$ in A_q or the truth if the truth was greater than this maximum value. This range coding ensured that the answers to each question contributed roughly equally to

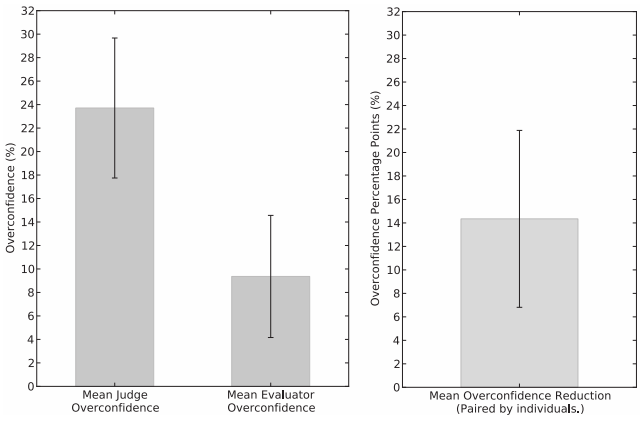

Figure 1: Evaluators were less overconfident than judges by 14.35 percentage points on average; 95% CI = ± 7.53. (Error bars are 95% CIs.)

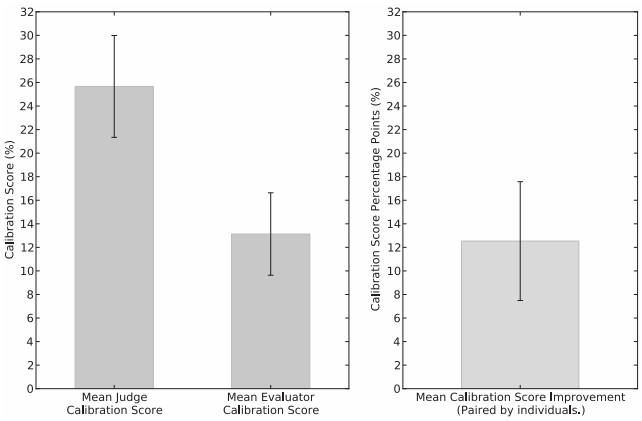

Figure 2: Evaluators were better calibrated than judges by 12.53 percentage points on average; 95% CI = ± 5.05.

the overall assessment of accuracy, which was measured in terms of the *average log-ratio error* (ALRE):

$$ALRE_m = \frac{1}{Q}\sum_{q=1}^{Q}\left|\log_{10}\left(\frac{\overline{T}_q + 1}{\overline{a}_{n,q}^{(i)} + 1}\right)\right|$$

where m is an individual or aggregation method, and Q is the number of questions (in this paper, $Q = 30$), and $\overline{T}_q$ is the range coded true answer to question q. The smaller an ALRE score for an individual or aggregation method, the more accurate the individual's or aggregation method's answers were—a perfect ALRE is 0.

Results

Individual hit rates ranged from a minimum of 13.33% to a maximum of 76.67%, with an average of 43.33% (SD = 15.37). Individual ALRE scores ranged from a minimum of 0.0250 to a maximum of 0.0787, with an average of 0.0392 ($SD = 0.0128$).

On average, judges were overconfident by 23.71% (SD = 14.27), and evaluators were also overconfident, but less so: 9.36% (SD = 12.45). Judgement swapping reduced the overconfidence effect on average by 14.35 percentage points (a 60.5% reduction); 95% CI = ± 7.53. See Fig. 1.

Calibration (the average of the absolute values of the over/underconfidence scores) was also improved. On average, judges were miscalibrated by 25.66% (SD = 10.35), and evaluators were miscalibrated by 13.13% (SD = 8.37). Judgement swapping improved calibration on average by 12.53 percentage points (a 48.8% improvement); 95% CI = ± 5.05. See Fig. 2.

The different aggregation functions displayed different strengths with respect to the different measures of performance. As expected, LinAgg-E resulted in less overconfidence than LinAgg-J (2.70% and 17.04% respectively). MECS resulted in a much higher hit rate than MJCS (86.67% and 70.00% respectively) and was slightly more accurate. MJIS and MEIS did substantially better than MJCS and MECS in terms of accuracy, but also substantially worse in terms of hit rates and overconfidence levels. MConS had an accuracy and hit rate between the accuracies and hit rates of the MIS and MCS methods, and MConS-J and MConS-E both had low overconfidence levels (6.99% and 6.71% respectively). No aggregation function had an ALRE that was better than the best individual ALRE. The aggregation function with the best ALRE was MJIS with an ALRE of 0.0305, while the best individual ALRE was 0.0250. MECS was the only aggregation function to have a hit rate that was better than the best individual hit rate (86.67% and 76.67% respectively). The best calibrated method was LinAgg-E which had an overconfidence of 2.70%. The full results for the different aggregation methods are in Table 1.

Discussion

Judgement swapping reduced overconfidence and improved calibration. We take these results to support the hypothesis that people are better at evaluating other people's judgements rather than their own (Teigen and Jørgensen (2005), Winman *et al.* (2004), Talyor and Brown (1988)), but only so long as everyone considers a space of alternative hypotheses (Koehler (1994), Koehler and Harvey (1997)).

This effect was maintained at the group level, when the group's judgment was determined by linear aggregation (LinAgg). Linear aggregation paired with average evaluator confidences had the best calibrated confidences and was a close second best for judgement accuracy (out of the aggregation methods studied).

Maximum confidence slating (MCS) for interval judgements (adapted from MCS for binary judgments in Koriat (2012b)) produced the highest hit rates, but performed poorly in terms of accuracy. Paired with evaluator confidences (i.e., MECS), the method substantially improved the hit rate again (compared to MJCS), but only slightly improved accuracy.

Maximum information slating (MIS) produced very good accuracies, but very low hit rates. Using evaluator confidences (MEIS), the method improved in hit rate (compared

Group Judgement	ALRE	Hit Rate	Judge Overconfidence	Evaluator Overconfidence
Mean Individual	0.0392 (SD =0.0128)	43.33% (SD =15.37)	23.71% (SD =14.27)	9.36% (SD =12.45)
LinAgg	0.0306	50.00%	17.04%	2.70%
MJCS	0.0649	70.00%	29.40%	5.58%
MECS	0.0627	86.67%	-6.02%	11.10%
MJIS	0.0305	16.67%	53.90%	19.70%
MEIS	0.0322	26.67%	44.77%	42.60%
MConS	0.0550	60.00%	6.99%	6.71%

Table 1: Aggregation Results.

to MJIS), but reduced accuracy. MIS also resulted in a very high overconfidence levels—with both judge confidences and evaluator confidences.

As expected, maximum consensus slating (MConS) produced well calibrated judgements and a reasonable hit rate (better than LinAgg but worse than MJCS). However, MConS resulted in low accuracies (but not as bad as the MJCS and MEIS accuracies).

The different aggregation functions clearly have different strengths, and so one's choice of aggregation function will depend on how much one cares about the different aspects of judgement quality (accuracy, hit rate, calibration). It's therefore natural to ask if the aggregation methods can be combined in some way to produce an aggregation of the aggregation methods that has the best features of the individual aggregation method. Results from an initial investigation into this question look promising. For example, by simply taking the average of the judgements resulting from LingAgg-E and MEIS-E, an ALRE of 0.0294 was achieved with a hit rate of 56.67%, judge overconfidence of 12.57% and evaluator overconfidence of 4.31%. This ALRE is better than any other aggregation method's ALRE and the hit rate is better than the individual hit rates of LinAgg-E and MEIS-E. The evaluator overconfidence is also quite low—substantially lower than MEIS-E's overconfidence.

It should be stressed that these results are from just one experiment, and clearly need replication. However, the statistically significant and strong improvement in overconfidence and calibration that resulted from judgement swapping is very promising.

References

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

Koehler, D., and Harvey, N. 1997. Confidence Judgments by Actors and Observers. *Journal of Behavioral Decision Making* 10(3):221–242.

Koehler, D. 1994. Hypothesis Generation and Confidence in Judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(2):461.

Koriat, A. 2012a. The Self-Consistency Model of Subjective Confidence. *Psychological Review* 119(1):80–113.

Koriat, A. 2012b. When are Two Heads Better than One and Why? *Science* 336:360–2.

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

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many are Smarter than the Few and how Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday.

Taylor, S., and Brown, J. 1988. Illusion and Well-Being: A Social Psychological Perspective on Mental Health. *Psychological Bulletin; Psychological Bulletin* 103(2):193.

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

Winman, A.; Hansson, P.; and Juslin, P. 2004. Subjective Probability Intervals: How to Reduce Overconfidence by Interval Evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30(6):1167.