The Impact of the Number of Experts on Prediction Accuracy

Ali Mosleh[a*], Ellis Feldman[b]
[a] The B. John Garrick Institute for the Risk Sciences UCLA, USA
[b]Self, Maryland, USA

Decision-makers solicit and use expert judgment opinion when data are unavailable, or it is impracticable to calculate a quantity analytically. Estimates may have life-or-death consequences. Accuracy can be increased by combining individual estimates. This research investigates the sensitivity of aggregated estimate accuracy to number of experts using a large, diverse meta-database. A multiplicative metric for accuracy is defined, and a rationale for its use is provided. Sensitivity analysis is conducted separately for physical and probabilistic data, based on separate research which found different expert judgment reliabilities for the two types over the meta-database. For both data types, increasing the number of experts from one to two reduces multiplicative error by a factor of ten. The largest improvements occur between one and four experts. Results for probabilistic data are more dispersed than for physical data; larger numbers of experts should be employed for the former, to help avert large estimation errors. A plateau is reached at about six or seven experts, with little consistent improvement for larger numbers of experts. Finally, the extent of improvement with increased numbers of experts is highly dependent on the problem domain.

KEY WORDS: Expert Judgment; Numbers of Experts; Meta-database, Multiplicative error

1. INTRODUCTION

It is said that "two heads are better than one". For the purposes of this paper, "better" will be defined in terms of a lower expected value of a ratio-based error metric, in the context of expert judgment predictions drawn from a large, diverse meta-database of point estimates, aggregated into estimates compared against realized values. The sensitivity of the error metric to "number of heads" or experts, n, is explored, for different data types.

Expert judgment is used generally when there is uncertainty regarding insufficient data or data cannot be obtained, e.g., [1], [2]. It is used where data are not available, or it is impracticable to calculate a quantity analytically. Notwithstanding the potential significant ramifications of an erroneous expert prediction, actual and predicted values can diverge by orders of magnitude.
Similar discrepancies have been observed between predictions made by different experts. These discrepancies need not arise from "incompetence, venality and ideology", but rather "may be attributable to the character and fallibilities of human judgement itself" [3]. Although point estimates are less useful than interval estimates, the former continue to be observed in U.S. Government (USG) expert judgment elicitation studies, e.g. [4], [5]. Because significant

* mosleh@ucla.edu

economic consequences such as nuclear reactor core melt or volcanic events, can follow inaccurate predictions. Therefore, their reliability needs to be scrutinized.

Let r be the ratio of realized value to predicted median, e/e'. A derived metric, Maximum Multiplicative Error, or MME is the maximum of this ratio and its inverse. The use of MME as a metric penalizes multiplicative excursions of estimates on either side of the realized value equally. Such excursions could reflect a cost overrun, or an effectiveness underrun, with equal impacts on a cost-benefit ratio.

Research by the authors, discussed in a paper currently in preparation, found a significant distinction in the reliability of estimates between *physical* and *probabilistic* data: elicited point estimates for a probabilistic variable, such as "probability of a runway overrun during takeoff at Schiphol Airport", tended to over- or under-estimate the actual value by larger factors than corresponding estimates for a physical variable such as "Shiveluch volcano AD1430 3 cu km avalanche area, in sq. km". Over all of the meta-database variables, the fraction of predictions for which the MME exceeded a given factor, such as 2,5,10,100, or 1000, was approximately twice as large for probabilistic data as for physical data. The distinction between the errors associated with the two types was especially pronounced when the quantity being estimated had a small realized value.

Research by the authors found that the differences in prediction accuracy between the two data types, observed for individual elicited e', persist when the latter have been mathematically aggregated into a point estimate, ê. The measure of accuracy is again MME, defined in a manner analogous to that used for individual e': MME = max(e/ê, ê/e). This research will be documented in a third paper.

[6] states the motivation "behind using multiple experts and/or multiple methods is simply to get additional information that can lead to more accurate forecasts or estimates and, ultimately, to better decisions". [7] states multiple experts are preferred to a single expert since the reliability of an aggregated opinion is generally more reliable than a single opinion, and the error or bias in the individual opinions may be lowered. It will be shown in the third paper that MMEs associated with aggregated estimates are indeed reduced from those associated with individual predictions. Nine aggregation techniques will be discussed in that paper, ranging from simple arithmetic averaging to a more complex Bayesian method incorporating a regression equation. Although there continue to be differences in accuracy between the two data types, these differences are not as great as in the individual e' prediction case. Of the nine aggregation methods, two of the simplest—the *median* and the *geometric mean*—were, respectively, the most accurate and in the "middle of the pack" for accuracy[1]. (The arithmetic and harmonic mean, while simple methods of aggregation, were sensitive to outliers, and had

[1] Winkler notes "while still being open to more complex combining models/methods [he] had gravitated over the years to a feeling that there are considerable advantages to simple models when combining subjective probability forecasts" [11, p. 17]. The geometric mean was generally within twenty percent of other methods, excluding the poorly performing arithmetic and harmonic mean.

poor performance in terms of average MME). It is these two methods that will be considered in the present, sensitivity analysis paper.

Because of the large possible economic consequences attendant on inaccurate expert judgment estimates, there is a need for practical guidelines and observations concerning the behavior of aggregated judgment versus number of experts. The current paper makes a contribution by presenting results concerning the rate at which aggregated point estimates become more accurate, as the number of experts is increased. The sensitivity analyses of MME versus n is performed using a large, diverse meta-database containing thousands of individual predictions which can be aggregated and compared to realized values. The analyses are performed separately for physical and probabilistic data types.

2. BACKGROUND

According to [6], the "motivation behind using multiple experts and/or multiple methods is simply to get additional information that can lead to more accurate forecasts or estimates and, ultimately, to better decisions" (p. 167). There is no best practice or framework that indicates how many experts are required for decision support in any particular domain; however, [7] recommend choosing the experts so that their "combined knowledge and expertise reflects the full scope of the problem domain". [8] argued that multiple experts are required so that a problem will be addressed from different perspectives. Additionally, use of "a single expert will slant results towards the content and functioning of his or her memory" ([8], p. 87). [9] observed that in "general the group judgment will be more accurate than the individual judgments primarily due to a decrease in error variance" (p. 25). [10] concluded that "Aggregation of expert opinion using group medians does give some improvement in accuracy".

The number of experts required for such aggregation is an open question. [12] observed that personal "experience with more than 20 panels suggests that 8–15 experts is a viable number — getting more together will not change findings significantly, but will incur extra expense and time. However, this has not been rigorously tested" (p. 295). [13] noted that "there may be diminishing returns on the number of experts used in an elicitation" (p. 159). [14] reported that based on panelists interviewed during an expert judgment policy symposium and workshop, the number of experts for most studies they conducted was "targeted to lie between 6 and 12".
These considerations coupled with resource factors such as budget and time, render the response to the question "how many experts are required?" more of a judgment call than a scientific determination.

This paper aims to inform that judgment by reporting on the variation in the accuracy of the aggregated estimate, MME, versus the number of experts used, n, where the latter ranges from 1 to 45. The data sets used are extracts from Delft University of Technology (TUD) expert

judgment meta-databases, described in [15] and University of Maryland Center for Reliability and Risk Analysis (UMD) expert judgment meta-data sources. The TUD source is cited in the peer-reviewed literature. The UMD source was derived from two dissertations addressing expert judgment [16], [17], and related graduate coursework, conducted in the Department of Mechanical Engineering and thus provided an opportunity to introduce other similar data. All data consist of named variables with known realized values, against which experts predicted median values. (The TUD data also contained 5th and 95th percentile values, but these were used by only one aggregation method, the so-called "classical" method of Cooke [18]; with respect to probabilistic data, it was the least accurate of the aggregation methods, in terms of average MME; it was not considered further for this analysis.) The data are referred to collectively as the Expert Judgment Extracts, EJE.

Each variable along with realized values and elicited predicted medians is called a record. Sets of variables related in terms of subject-matter were assigned to themes where each theme is unique to either TUD or UMD. The number of records from TUD and UMD are 606 and 1,182 respectively, for a total of 1,788 records. Each EJE record was assigned a Sequence ID number (SeqID) for data management purposes. Examples of EJE physical and probabilistic records are shown in Table 1 below. In the table, "Nobs" refers to number of observations, and the e'i are elicited predicted values. Only the first four or "Nobs", whichever is smaller, predictions, are shown. The "Variable" column contains a compressed description[2] of the variable against which the estimates were elicited; the "Value" column contains the corresponding realized value. SeqIDs 27, 39, and 11 are examples of probabilistic variables; while IDs 937, 909, and 1711 are examples of physical variables.

Table 1: Examples of EJE Records

SeqID	Theme	Variable	Nobs	Value	e'_1	e'_2	e'_3	e'_4
937	Dams	Teton first	11	2.5	4	10	10	15
909	Crop Yield	Wheat Trilby KS 76	1	19.2	32.1			
1711	Volcanoes	Hudson clast d	45	60	500	800	1	100
27	INFOSEC	QI4 (How likely ...)	13	0.01	0.05	0.54	0.1	0.4
39	PM25	L97 > 50	6	0.0603	0.0329	0.0548	0.1233	0.0548

[2] SeqID 11: Inflight shutdown rate per flight hour for CFM56-7B engine in 2004
SeqID 27: How likely is a successful attack on a system to be detected and reported to authorities by managers responsible for the security of the system?
SeqID 39: Number of days in London in 1997, that the density of inhalable particles 2.5 microns or less across, exceeds 50 μg per cubic meter of air; as a fraction of the days in that year.
SeqID 909: 1976 forecast crop yield (bushels per acre) in Tribune, KS area.
SeqID 937: Teton dam (collapse of June 5, 1976) time in hrs from mud seen to whirlpooling.
SeqID 1711: Mount Hudson volcanic eruption, maximum pumice clast dimension at 50km isopleth, in mm

SeqID	Theme	Variable	Nobs	Value	e'_1	e'_2	e'_3	e'_4
11	Aviation	ShtDwnRate	5	0.000002	0.00025	1.00E-06	5.00E-06	1.00E-04

Table 2 below provides summary information on records, predictions and themes by data category (type) in the EJE database.

Table 2: Record, Predictions and Theme Counts by Data Category in the EJE Database

Record, Predictions and Theme Counts by Data Category	TUD	UMD	EJE Total
Number of records in Physical Category	540	1,181	1,721
Number of records in Probabilistic Category	66	1	67
Number of predictions in Physical Category	4,661	1,445	6,106
Number of predictions in Probabilistic Category	516	13	529
Number of themes in Physical Category	27	16	43
Number of themes in Probabilistic Category	8	1	9

The number of experts associated with each of the 1788 records in the EJE meta-database ranges from 1 to 45; the median and average numbers of such experts are 6 and 6.7, respectively3. 1141 records (64%) are associated with a single observation each; 143 (8%) are associated with seven observations each (n=7); and smaller percentages are associated with other values of n. Table 3 below gives the distribution of the number of predictions by number of EJE records.

Table 3: Distribution of Number of Predictions by Number of Records in EJE

Number of Observations	Number of Records: Physical Data	Number of Records: Probabilistic Data
1	1141	
2	26	
3	5	1
4	65	11
5	64	11
6	84	20
7	137	6
8	27	

[3] Sets of variables related in terms of subject matter were assigned to themes where each themTable e is unique to either TUD or UMD. Records are weighted so that each theme has the same total weight. Statistics related to numbers of experts incorporate the weights.

Number of Observations	Number of Records: Physical Data	Number of Records: Probabilistic Data
9	13	
10	28	
11	49	9
12	8	2
13	8	5
17	47	
20	1	
31	9	1
45	9	1
Total Number of Records:	*1,721*	*67*

The variation in the quality of aggregated expert judgment, defined in this paper as expected MME, versus number of experts, n is explored with respect to this database, using the median and the geometric mean as aggregation methods.

3. MATHEMATICAL FORMALISM

The following explication of the method of computation of average MME for a given data type and aggregation method assumes, without loss of generality, that only records representing physical variables are under consideration, and aggregation is via the median. To compute average MME corresponding to a given number of experts, call it s, we first compute average MME for each record having n≥s observations, hereinafter, "obs". Before presenting the formalism associated with the calculations, an example is given:

Let the number of experts equal two (s=2). This immediately excludes 1141 records having n=1 obs. Of the remaining 580 physical records, there are 26 records having n=2 obs. For each of these records, the MME is simply the average of its two observations, e'1 and e'2. Five records have n=3 obs. For this example, consider one of the five: SeqID 1486, belonging to theme "PhD Surveys 2005"[4]. This variable is "Precip-Detroit Airport-D3", representing a 48-hour forecast of precipitation (in inches) at Detroit Metropolitan Wayne County Airport (DTW) for the third of three forecast days. The realized value was 0.08; and the three predictions were e'1=0.10, e'2=0.04, and e'3=0.20 inches of precipitation. Since we are interested in the average MME when there are s=2 experts, we take all possible combinations of the observations taken

[4] This EJE record incorporates data taken from one of a number of surveys of expert judgment forecasts compared to outcomes, contained in "THE QUALITY OF EXPERT JUDGMENT: AN INTERDISCIPLINARY INVESTIGATION", a dissertation by Yashika Forrester, UMD, 2005.

two at a time, without regard to order: e'1 and e'2—average=0.07; e'1 and e'3—average=0.15; and e'2 and e'3—average=0.12. Averaging these averages to get expected MME yields (0.07+0.15+0.12)/3≈0.113.

In general, there are $M = \binom{n}{s}$ distinct subsets of size s≤n obs which can be drawn, without regard to order. For a given method of aggregation, data type (physical or probabilistic), value of s, and record containing n≥s obs, the average, A of the M MMEs is computed by direct enumeration. Once the averages Ai have been computed for each record i having at least s observations, they must be combined in order to obtain a final expected value of MME associated with s. Simply averaging the Ai would not be correct, as this would overweight themes having many variables compared to those having few variables. The following weighting scheme is applied:

Without loss of generality, let the record with average MME=A, belong to a theme containing *nvar* records for which nobs≥s ; and let there be a total of *nthemes* containing at least one record with n≥s. Then the weight w applied to A is given by $[nthemes \cdot nvar]^{-1}$. This scheme ensures that each "participating record" is given equal weight within its theme; and that each theme is given equal weight in calculating the expected value of the error metric, MME, associated with *s* experts.

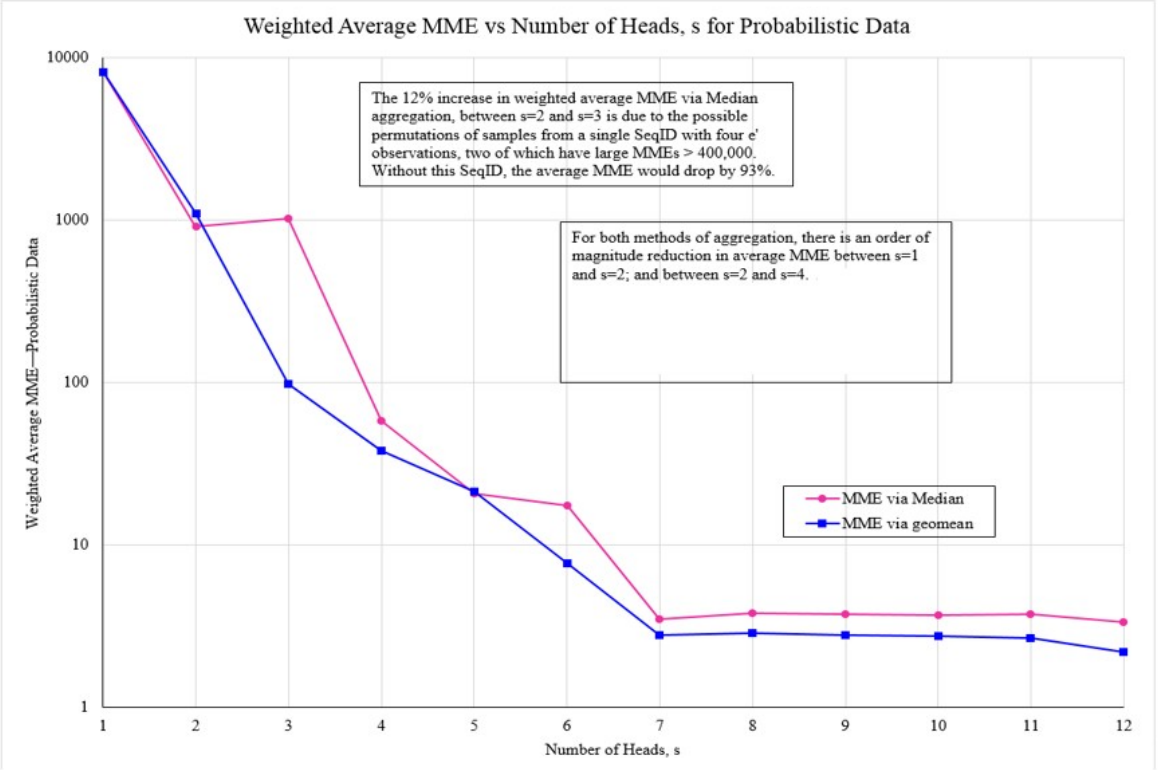
Figure 1, Weighted Average MME vs Number of Heads, s for Probabilistic Data shows results for aggregation of subsets of probabilistic EJE data for s=1 through 12, using the two aggregation methods. As shown in the figure, for both methods of aggregation, there is an order of magnitude reduction in average MME between s=1 and s=2, that is, when increasing the number of experts from one to two. There is an additional order of magnitude reduction between s=2 and s=4. The rate of reduction slows as s is further increased. It should be noted that these reductions occur over the entire set of probabilistic meta-data; a concluding note will discuss briefly what may occur when attention is confined to a particular theme or domain.

The figure shows a seemingly anomalous increase in average MME versus s between s=2 and s=3, using median aggregation. Such increases can occur because of the different methods of computation of the median depending on whether s is even or odd. For odd s, the "middle" value of the sorted e' is used as the median; for even s, the average of the two "middle values" is used. The following notional example will show how these different methods of computation can cause average MME versus subset size to decrease from s=1 to s=2, then increase from s=2 to s=3.

Consider a record containing n=3 obs e' = 0.001, 0.01 and 1, with a true value of e equal to one. At s=1, there are three possible values of MME depending on which e' is drawn: 1000, 100 or 1. Therefore, the average MME at s=1 will be 367. At s=3, the middle e' value must be chosen: 0.01, yielding an MME of 100. However, at s=2, the median is found by taking the average of two values, drawn repeatedly from the three observations. There are three possibilities: 0.001 and 0.01; 0.001 and 1, and 0.01 and 1. These give rise to medians of 0.0055, 0.5005, and 0.505, respectively. The corresponding MMEs are approximately 181.82, 2.00, and 1.98, respectively.

The average MME is, therefore, 61.93. Thus for this example, there is a factor of six reduction in average MME between s=1 and s=2, followed by a 60% increase between s=2 and s=3.

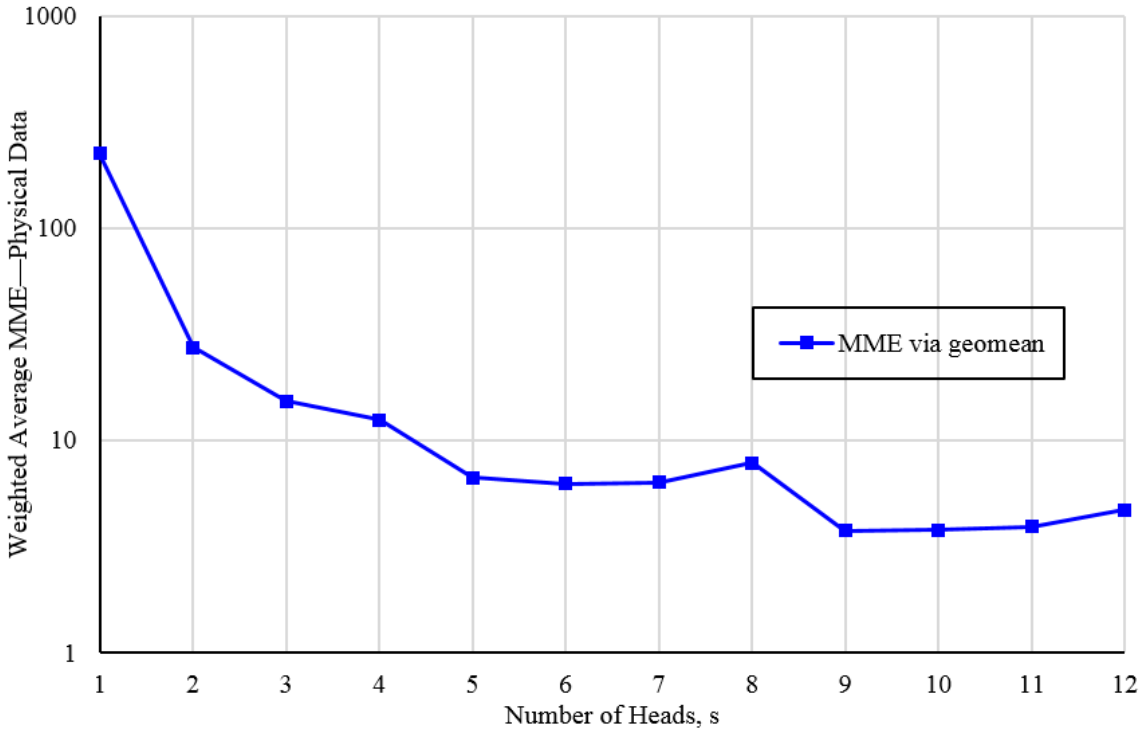
Figure 1: Weighted Average MME vs Number of Heads, s for Probabilistic Data



For physical data, a similar phenomenon caused a 37% increase in average weighted MME, from 7.04 to 9.67, between s=6 and s=7. The increase was associated with the different methods of computation of the median, and with permutations of e' values selected in the subsets for two SeqIDs, SeqID 1131 and 1134 (both having n=8 obs, with all e' values less than e, in some cases, much less: by factors of 30,000).

To avert this behavior, aggregation via the median is not further considered in this paper. Even when aggregating via the geometric mean only, anomalous increases can occur as s is increased: see Figure 2, Weighted Average MME vs Number of Heads, s for Physical Data.

Figure 2: Weighted Average MME vs Number of Heads, s for Physical Data



An increase of 23% in average weighted MME was observed between s=7 and s=8. This was due to 137 records with n=7 and somewhat lower MMEs for individual SeqIDs, dropping out of the mix at s=8. (See Table 3 above for numbers of records having given numbers of obs.)

As was the case for probabilistic data, a large drop in MME occurs betwen s=1 and s=2. The rate of reduction generally slows as s increases; and fluctuations increasing MME are possible. In order to avert these anomalous increases arising from the changing mix of participating records as s is varied, a second approach was taken.

For each given value of s, the weight wi applicable to the average MME, Ai associated with each record, was determined according to the method discussed previously. The number of heads was then varied between 1 and s, without changing the mix of participating records. The only change was that for number of heads k≤s, Ai for each participating record was set equal to the average of the MMEs associated with the geomeans of the $M = \binom{s}{k}$ distinct subsets of size k drawn from the record's observations. It can be shown that averages constructed in this manner monotonically decrease as k increases. Figure 3 below shows the decline in expected MME for probabilistic data as the number of experts increases.

Figure 3, Weighted Average MME—Probabilistic data vs. Number of Heads, using twelve sets of weights.

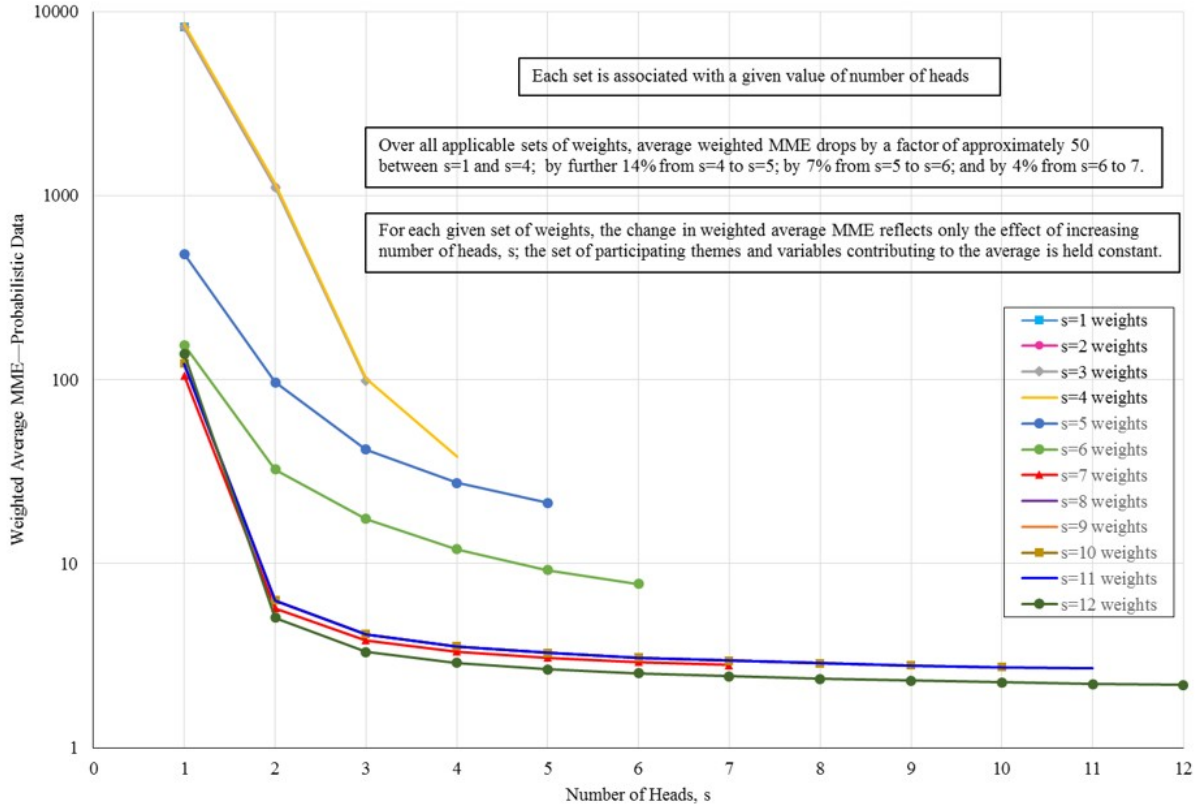
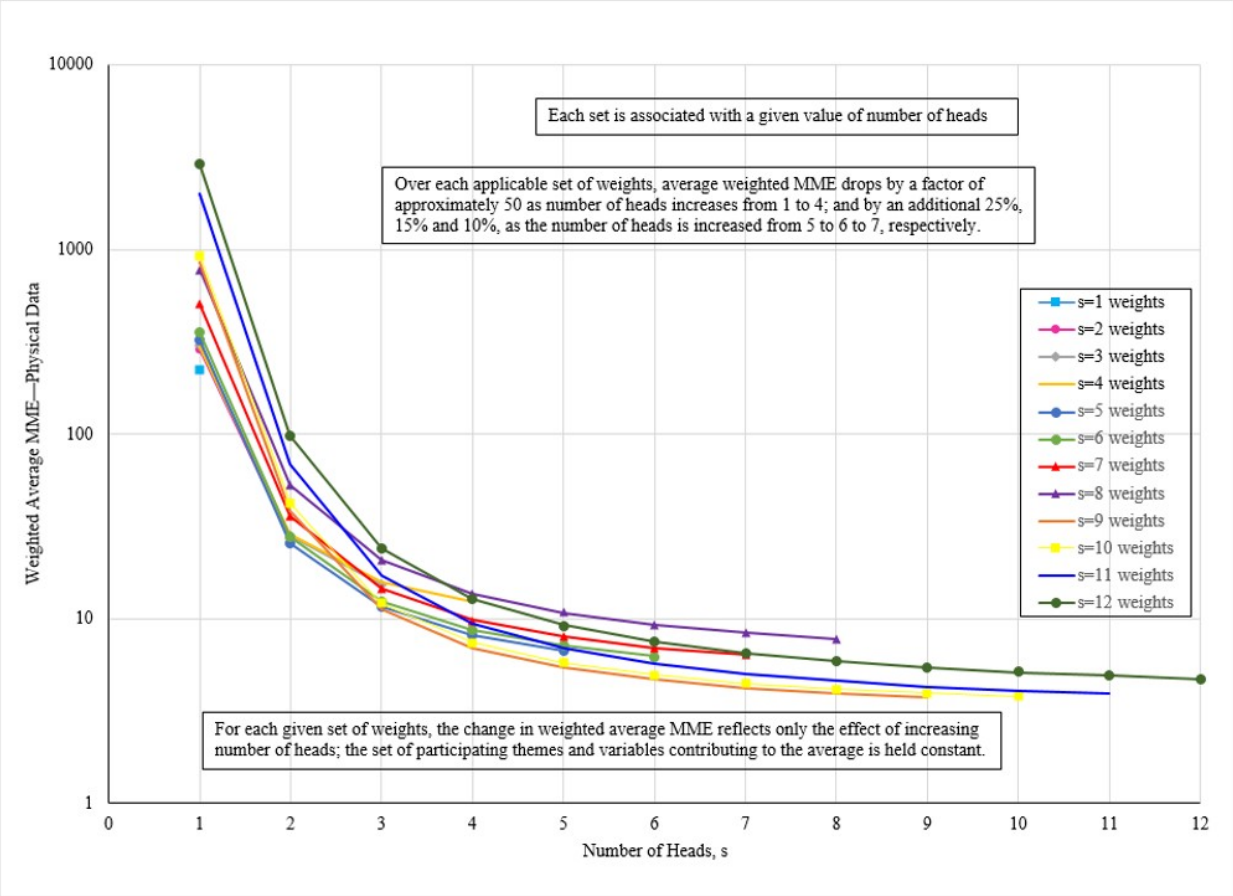


Figure 4 below shows the analogous decline with increasing numbers of experts, for physical data. Note that the "MME versus s" curves lie closer together for physical data than for probabilistic data. For both data types, average weighted MME decreases by an approximate factor of 50 as the number of experts is increased from one to four, over all sets of weights shown, with smaller declines for further increases in s. However, the graphs suggest that using larger numbers of experts, e.g., six, when aggregating individual estimates for probabilistic data, may help avert incurring errors exceeding a factor of ten. Note that if only three experts are used, an average MME of one hundred may be incurred. By contrast, for physical data, the corresponding average MME is less than thirty.

Note that although the percentage changes in average MME decline as the curves are traversed from left to right with increasing s, the levels of the curves (their ordinates at the largest value of s) can fluctuate. This reflects the changing mix of participating variables. For example, the weighted average MME over the 74 records participating at s=13, is 3.95 (not shown in the

figure)5. Half of this value arises from a single record6, for which one of the 13 individual predictions was in error by a factor of 400,000. This record would drop out of the mix at s=14, and the overall MME would be reduced to approximately two. At s=45, only the nine records associated with the "Volcanoes" theme can participate. Because the experts were less accurate against the variables in this theme, even with all 45 participating, aggregated MME ranged from 1.04 to 6.55, averaging 3.1 —larger than the approximate value of two corresponding to s=14.

Figure 4: Weighted Average MME—Physical data vs. Number of Heads, using twelve sets of weights



The larger error corresponding to a particular theme or field reflects an important point: the order-of-magnitude reduction in average MME when number of experts was increased from one to two, involved all applicable records of a given data type contained in the meta-database. However, if attention had been confined to a particular subset of the data, much more modest gains from increasing the numbers of experts might have been observed. Shanteau (2015) noted that for certain domains, such as those involving physical systems, or where stimuli were "relatively

[5] Results were computed only through s=13 due to computational processing limitations (number of enumerated subsets of size s=13 which can be drawn from n=45 is equal to 73,006,209,045); however, the fact that MME cannot increase permits bounding of the results. Results for s=45 can be at most approximately twenty percent below those obtained for s=13.

[6] SeqID 1655: Theme: UMD Campus; Variable: Volumes in UMD Library

constant, ... judges were faced with stationary targets. In contrast, domains with poor performance involved dynamic stimuli, generally involving human behavior." (p. 172). Judges performed more poorly in the latter types of domain. This phenomenon is present in the EJE database. For example, for the theme pm25, which involves particulate emissions, the ratio of maximum to minimum e' against each of the variables in this theme, averages less than 1.14. The average MME when only a single prediction from this theme is chosen, is 1.12. Increasing nheads to six decreases the average MME over this theme by one percent, to 1.11. By contrast, the theme INFOSEC (information security) represents a newer, dynamic field involving human behavior. For this theme, the ratio of maximum to minimum e' against each of its probabilistic variables averages 30,000. The corresponding average MME is 684 when nheads equals one; it declines to 6.5 when nheads equals six. In conclusion, the applicable domain is highly significant as to the extent of improvement associated with increases in numbers of experts.

4. CONCLUSIONS

1. Increasing the number of experts from one to two reduces error by a factor of ten
2. Based on this broad base of meta-data, largest improvements in average MME are observed between n=1 and n=4 heads.
3. Results for probabilistic data are more dispersed than for physical data; therefore, larger numbers of experts should be employed for the former, to help avert large estimation errors.
4. A plateau is reached at about n=6 or n=7, with little consistent improvement for larger n.
5. The extent of improvement associated with increases in numbers of experts is highly dependent on the domain.

5. REFERENCES

[1] Board of Governors of the Federal Reserve System. (2013). *Capital planning at large bank holding companies: Supervisory expectations and range of current practice*. Retrieved on August 14, 2018 from http://www.federalreserve.gov/bankinforeg/stress-tests/ccar/August-2013-Estimation-Methodologies-for-Losses-Revenues-and-Expenses.htm#subsection-1934-08640AF5.

[2] Lewis, H. L., Budnitz, H. J., Couts, C., von Hippel, F., Lowenstein, W. B., & F. Zachariasen, F. (1978). *Risk Assessment Review Group Report to the U. S. Nuclear Regulatory Commission* (NUREG/CR-0400).

[3] Mumpower, J. L., & Stewart, T. R. (1996). *Expert judgement and expert disagreement*. Thinking and Reasoning, 2(2/3), 191-211.

[4] U.S. Department of Agriculture, Food Safety and Inspection Service. (2007). *Results of an additional expert elicitation on the relative risks of meat and poultry products* (Draft Report Contract No. 53-3A94-03-12, Task Order 27). Retrieved on August 14, 2018 from https://www.fsis.usda.gov/wps/wcm/connect/2d081ff1-cdc3-4975-94d4-948930b6e141/RBI_Elicitation_Report.pdf?MOD=AJPERES.

[5] U.S. Department of Agriculture, Food Safety and Inspection Service. (2012). *Expert elicitation on the market shares for raw meat and poultry products containing added solutions and mechanically tenderized raw meat and poultry products*. Retrieved on August 14, 2018 from

https://www.fsis.usda.gov/wps/wcm/connect/3a97f0b5-b523-4225-8387-c56a1eeee189/Market_Shares_MTB_0212.pdf?MOD=AJPERES.
[6] Winkler, R. L., & Clemen, R. T. (2004) *Multiple experts vs. multiple methods: Combining correlation assessments*. Decision Analysis, 1(3), 167–176. doi 10.1287/deca.1030.0008.
[7] Rowe, G. & Wright, G. (2001). *Expert Opinions in Forecasting: The Role of the Delphi Technique*. In J. Scott Armstrong (Ed.) Principles of forecasting: A handbook for researchers and practitioners. https://link.springer.com/chapter/10.1007%2F978-0-306-47630-3_7. Retrieved on August 14, 2018 from https://www.gwern.net/docs/predictions/2001-rowe.pdf
[8] Meyer, M. A. & Booker, J. M. (2001). *Eliciting and analyzing expert judgment: a practical guide*. American Statistical Association and the Society for Industrial and Applied Mathematics, Alexandria, Virginia, USA. Retrieved on November 30, 2014 from http://books.google.com/books?hl=en&lr=&id=ZLt_y-patXYC&oi=fnd&pg=PR2&dq=Meyer+MA+and+Booker+JM,+2001.+Eliciting+and+analyzing+expert+judgment&ots=clvta_OkAQ&sig=cOagmRrq8gQyO1yLuvbNdXiEqjg#v=onepage&q=payment&f=false.
[9] Seaver, D. A. (1976). *Assessment of group preferences and group uncertainty for decision making*. (Defense Technical Information Center Technical Report. Jul 75-Sep 76). Retrieved on August 14, 2018 from http://www.dtic.mil/dtic/tr/fulltext/u2/a033246.pdf.
[10] Martz, H. F., Bryson, M. C., & Waller, R. A. (1985). *Eliciting and aggregating subjective judgements – some experimental results*. Los Alamos National Laboratory LU-UR—84-3193.
[11] Winkler, R. L. (2015) *Equal Versus Differential Weighting in Combining Forecasts*. Risk Analysis 35 (1), 16–18. doi 10.1111/risa.12302.
[12] Aspinall, W. (2010). *A route to more tractable expert advice*. Nature, 463, 294-295. doi:10.1038/463294a.
[13] European Food Safety Authority (2014). *Guidance on expert knowledge elicitation in food and feed safety risk assessment.* EFSA Journal 2014;12(6):3734. Retrieved on November 30, 2014 from http://www.efsa.europa.eu/en/efsajournal/doc/3734.pdf.
[14] Cooke, R. M., & Probst, K. N. (2006). "*Highlights of the expert judgment policy symposium and technical workshop*" Retrieved on October 19, 2014 from http://www.rff.org/Documents/Conference-Summary.pdf.
[15] Cooke, R. M., &. Goossens L. H. J. (2008). *TU Delft expert judgment data base*. Reliability Engineering and System Safety, 93(5), 657–674. doi:10.1016/j.ress.2007.03.005.
[16] Forrester, Y. (2005). *The quality of expert judgment: An interdisciplinary investigation* (Doctoral Dissertation). Retrieved on August 14, 2018 from http://hdl.handle.net/1903/3267.
[17] Shirazi, C. H. (2009). *Data-informed calibration and aggregation of expert judgment in a Bayesian framework* (Doctoral Dissertation). Retrieved on August 14, 2018 from http://hdl.handle.net/1903/9883.
[18] Cooke, R. M., *Experts in Uncertainty*, Oxford University Press, 1991, New York.