Decision Analysis

Two Reasons to Make Aggregated Probability Forecasts More Extreme

Jonathan Baron, Barbara A. Mellers, Philip E. Tetlock, Eric Stone, Lyle H. Ungar

Please scroll down for article—it is on subsequent pages

INFORMS is the largest professional society in the world for professionals in the fields of operations research, management
science, and analytics.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

Two Reasons to Make Aggregated Probability Forecasts More Extreme

Jonathan Baron, Barbara A. Mellers, Philip E. Tetlock, Eric Stone
Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania 19104
{baron@psych.upenn.edu, mellers@psych.upenn.edu, tetlock@wharton.upenn.edu, estone@sas.upenn.edu}

Lyle H. Ungar
Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104,
ungar@cis.upenn.edu

When aggregating the probability estimates of many individuals to form a consensus probability estimate of an uncertain future event, it is common to combine them using a simple weighted average. Such aggregated probabilities correspond more closely to the real world if they are transformed by pushing them closer to 0 or 1. We explain the need for such transformations in terms of two distorting factors: The first factor is the compression of the probability scale at the two ends, so that random error tends to push the average probability toward 0.5. This effect does not occur for the median forecast, or, arguably, for the mean of the log odds of individual forecasts. The second factor—which affects mean, median, and mean of log odds—is the result of forecasters taking into account their individual ignorance of the total body of information available. Individual confidence in the direction of a probability judgment (high/low) thus fails to take into account the wisdom of crowds that results from combining different evidence available to different judges. We show that the same transformation function can approximately eliminate both distorting effects with different parameters for the mean and the median. And we show how, in principle, use of the median can help distinguish the two effects.

Keywords: forecasting; aggregation; wisdom of crowds; confidence
History: Received on October 30, 2012. Accepted by former Editor-in-Chief L. Robin Keller on February 2, 2014, after 3 revisions. Published online in *Articles in Advance* March 19, 2014.

1. Introduction

The standard practice when combining probability estimates from many experts to form a single overall estimate is to take a (perhaps weighted) average of the individual estimates. Several investigators have found that averaged estimates are typically conservative and can be improved by transforming the average so that probabilities become more extreme, closer to 0 or 1 (e.g., Ariely et al. 2000, Turner et al. 2013). We have used the transformation

$$t(p) = \frac{p^a}{p^a + (1-p)^a}. \tag{1}$$

This transformation, equivalent to that used by Erev et al. (1994), Shlomi and Wallsten (2010), and many others (apparently beginning with Karmarkar 1978), is shown in Figure 1, with $a = 2.5$.[1]

In a study of probabilistic forecasting of international events (Mellers et al. 2014), we found empirically that an extremizing transformation in this form, with values of a greater than 1, could improve the quality of aggregated probability judgments, as determined by Brier scores. We were surprised that the optimal values of a were greater than 1. In subsequent research, we have found cases in which they are less than 1, presumably because the judgments are initially much too extreme. We first thought that the need for transformation was the result of compression of the probability scale at the

---

[1] Others (reviewed by Turner et al. 2013) have used a less restrictive transformation in which the level varies as well as the amount of extremization; i.e., $t(p) = \delta p^a/(\delta p^a + (1-p)^a)$. We used the restricted form because, in our data, forecasters provided probabilities for both "yes" and "no" when the target event could be described as happening or not; in other cases, such as which party would win an election, the happen/not-happen distinction was absent. The theoretical points we make would apply to the more general form as well.

Figure 1   Transformation of Probability Judgment with $a = 2.5$
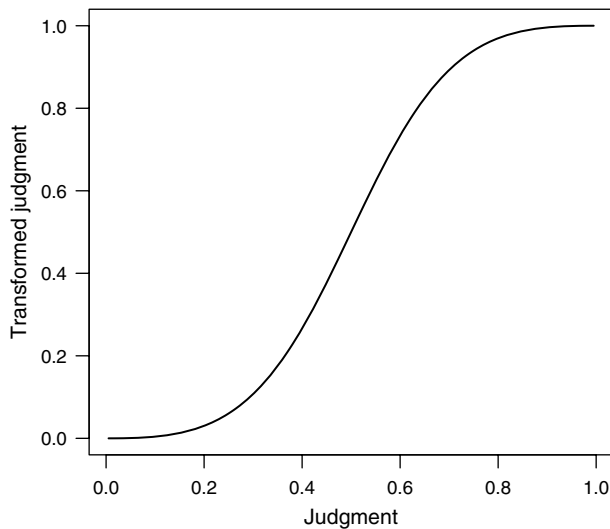
extremes. We supposed that a judgment had a true value for a group of forecasters; a value that is then expressed with error that is equally likely to be positive or negative, but not otherwise symmetric around the true value. Given this assumption, a very high true value, such as 0.99, cannot go much higher if the error is positive, but it can go much lower, so the mean of several observed scores is likely to be less than 0.99, even though 0.99 is the expected median. Juslin et al. (2000) refer to this as an end-of-scale effect. Because of this, transforming the observed mean by pushing it closer to 1 (extremizing, when the average was above 0.5) would then come closer to recovering the true value (as suggested by Erev et al. 1994).

We were, however, puzzled when we found that an extremizing transformation also improved Brier scores when we aggregated by using the median forecast rather than the mean. The argument of the last paragraph implies that the mean of a set of judgments is less extreme than the true value, but the median should be an unbiased estimate of the true value. So why not use the median instead of the mean? With enough forecasts, this is a good idea, but with a few forecasts the median could be more affected by sampling error of the population of forecasters. And, indeed, the median is usually more extreme than the mean, but apparently not as extreme as it should be.

This paper describes our proposed solution to this puzzlement. Specifically, we propose that judges

actually attenuate their own expressions of confidence, to take into account their belief that they are missing useful information that might be available. In other words, stated probabilities are expressions of individual confidence rather than confidence in the best forecast that can be derived from all available information. If you predict an event with 0.6 probability, you are saying that you think it will happen but you are not very confident. You might be more confident if you had more information. Each forecaster may feel this way to some extent, yet the average of their forecasts takes advantage of the fact that the forecasters differ in the information they have (Wallsten and Diederich 2001). If every forecaster said 0.6, and they were using different information, then someone who knew all of this would have a right to much higher confidence. Thus, their average confidence is less than the confidence that anyone should have in the forecast that can be inferred from the group average. The latter is based on something closer to the total set of information available to all the forecasters. The attenuation resulting from the feeling of missing information would affect the median judgment as well as the mean.

Hypothetically, we can imagine that each forecaster could state a probability based on the evidence she has. Sometimes this will only be one or two items, and these will point strongly in one direction. The forecaster might think, "If my evidence were the totality of evidence available, then I would give this a very high probability. But, if I had all the evidence, I might even come out with a probability less than 0.5. So I will reduce the extremity of my judgment." Note that this reduction might increase the calibration of the judge who does it, or improve the Brier score. We are not arguing that any sort of cognitive bias is involved. If anything, all the known biases go in the opposite direction, the direction of overconfidence, and these biases might reduce or reverse the effect we postulate.

The amount of reduction in judgment extremity (which we call Regression) depends on the amount of information that the forecaster feels is missing. This is of course a difficult judgment to make. Moreover, the average of many individual forecasts is usually based on more information than each of the forecasts making it up. Some forecasters might have access to information that says, "Yes, this will happen." Others might have access to conflicting or negative information. Yet, if the

forecasters sample all the information in an unbiased way, the average of their probability judgments should come close to reflecting the best judgment that could be made by a forecaster who had all the information, at least in terms of which side of 0.5 the judgment is on.

We show here that the transform we use can approximately correct for both factors that lead to average probabilities that are not extreme enough: error variation and regression. To do this, we develop a simple model, which we take to be approximately true in the same way in which traditional statistics takes its assumptions about normally distributed error, etc., to be approximately true. It is simple and convenient, and for practical purposes it is probably good enough. We also show how, in principle, the median can be used to distinguish the two different factors that make extremization work.

2. A Sketch of the Model

For a given question at a given time, assume that there is a true correct answer known only to an omniscient arbiter. The arbiter sends a signal $S_1, S_2, \ldots$, with an S_i for each possible answer i. We call this set S. Of course S_i for the true answer will usually be higher than for other possible answers. (Here, for simplicity, we assume that there is only one other possible answer, hence a binary question.) Each signal is bounded by 0 and 1. And the signal is the best possible probability estimate, in the following sense: given the total set of information available to all forecasters, no method of producing probability forecasts on the basis of this information could beat the method used to derive S in terms of maximizing the expected utility of decisions based on the probability. In Appendix A, we argue that we can approximate this criterion by asking how to optimize a strictly proper scoring rule, specifically, the Brier score. We assume that the forecaster does not know anything about what decisions will be made, or their payoffs, so we assume that each decision has a threshold probability for some action and that these thresholds are uniformly distributed between 0 and 1. This assumption leads to the conclusion that a method of judgment that optimizes the Brier score will also optimize the expected utility of these decisions.

2.1. Irreducible Uncertainty

Note that S_i need not be 0 or 1 for each i. The best information we have may leave us with considerable uncertainty, and we cannot reduce this uncertainty by gathering more information, because none exists. We call this *irreducible uncertainty* (IU). Reducible uncertainty exists when information is available to the judge that could improve the accuracy of the probability judgment. For example, if I do not know that Fermat's last theorem is true, I could look in Wikipedia and find out. IU exists for repeated events that are determined in part by conditions that we cannot know, such as the path, starting, point, and rotation speed of the flips of several coins. But IU also exists for unique events, such as whether the Republicans will control the U.S. Senate in the next election.[2]

2.2. The Error Distribution

We have a number of forecasters, indexed by j, who perceive each signal S with error and give probability judgments P_j based on their perception. We assume that error arises from two sources: different forecasters using different information and random noise in the translation of an internal feeling of certainty into a number both within and among forecasters.

Note that the error in the sampling of information may itself have several sources, and this does not affect our argument. Error may arise from judges having different information, some pointing to a positive outcome (i.e., whatever is coded as 1, usually the occurrence of some event) and some against it. Such information could include knowledge of base rates for categories that include the question at issue, analogies with other cases, or specific information about the case at hand. Base rates for different categories that include the target could be higher or lower than the signal for a given outcome S_i. Likewise, each inference from an analogy or from individuating information could, on their own, lead to probabilities above or below S_i, which is based on all sources of information available to anyone.

In sum, we think our assumption that sampling error is equally likely on both sides of S is plausible. It could

[2] IU does not map neatly into the distinction between aleatory and epistemic uncertainty, which is, in any case, largely a psychological distinction (Fox and Ülkümen 2011).

be incorrect in either direction, for specific cases or in general, but we see no reason to think so.

Error in the translation of beliefs into numbers could also have several sources, such as individual differences in the way people use information, or in precision of the numbers they provide, or from judgment-to-judgment variation within a person in the translation of a degree of belief into a number. We do not distinguish any of these sources of error here. We assume that all error is symmetric around the signal: half of it leads to judgments higher than the signal, half lower.

If we had a very large number of forecasters, perceivers of the signal, each of whom had some random sample of the available information, and if we knew the distribution of the error, we could recover S completely. With a smaller number of forecasters, we will have some additional error that results from sampling of the population of possible forecasters. We ignore this source in the following development, which is thus an approximation.

If the error distribution were symmetric, then our best estimate of S would be an average of the P_j's. On a probability scale, it is surely not symmetric in general. For example, if the signal is 0.9, some errors will be positive and others negative. The positive errors cannot be larger than 0.1, but the negative errors could be as large as 0.9. Ideally, we would want to transform individual probability judgments P_j so that their distribution becomes symmetric, average the transformed values, and then reverse the transformation. This approach is explored in a companion paper (Satopää et al. 2014), which uses the log odds transformation to make the error distribution more symmetric (as do Turner et al. 2013). In our application, this approach, using log odds, seems to do slightly better despite the problem of dealing with P_j values near 1 and 0, but this result may depend on the precision of forecasters when they give extreme judgments.

Here we shall stick with the idea of aggregating the probabilities first, and then transforming the aggregate. An advantage of this approach is that we do not need to deal with the problem of continuity in judgments near 0 and 1. We shall assume, however, that a log odds transformation creates the needed symmetry of the error distribution around S, and work through the implications of this assumption for the present approach, even though we do not otherwise use the log odds transformation in this paper. In making this assumption, we are assuming implicitly that nobody actually makes forecasts of 0 or 1, and we are avoiding the problem of interpreting such 0–1 forecasts as stated.

In sum, we assume that the error distribution around S is symmetric only with an appropriate transformation, such as log odds, but without the transformation we can still assume that $p(P_i > S_i) = p(P_i < S_i)$; i.e., positive and negative errors are equally likely.

Another approach would be to apply a transformation to each P_j, such that the transformed value is still in the 0–1 range, and then average the transformed values without any "back transformation." Such an approach might work better than the aggregate-then-transform approach taken here (i.e., average the probability judgments and then transform the average). For example, we could apply a transform such as Equation (1) before aggregating and then aggregate by taking the simple mean. Turner et al. (2013) use this approach with a more general transformation that contains a parameter δ for making all probabilities generally higher or lower, as well as a parameter for making them more or less extreme. Turner et al. found that this model does fairly well in terms of Brier scores, but not as well as the others they test (average then transform; transform with log odds, average, and back transform). We have also found in unpublished data that this method, using Equation (1) without the δ parameter, does worse than the ones we discuss here (and others). Part of the problem with this method may be that it could extremize judgments on the "wrong side" of 0.5, as well as those on the correct side. When judgments are highly variable yet their mean is still high and close to the correct side, extremizing would be insufficient because of these wrong-side probabilities. It isn't clear why we would want to extremize less, holding the untransformed mean constant, when judgments are more variable.

3. Regression of Log Odds Toward Zero

We assume that forecasters regress (or "shrink") their own estimates toward a probability of 0.5 (assuming two options) or toward log odds of 0. The amount of regression corresponds to the amount of information that forecasters individually think they are missing.

We should find less regression in the more expert groups. And we can measure the two variables that affect the transform—error around the signal and regression to zero—separately by looking at the optimal transforms for the mean and median for a given group. The median transform is affected only by regression. The difference between mean and median optimal transforms tells us about the random error component (including all sources of error).

We model the situation by supposing that forecasters develop their forecasts by updating a prior distribution as new information arrives, which we represent with a simple Bayesian model. Suppose we have a normal prior distribution of forecasts in log-odds space, with mean 0 and a large variance.[3] We have a normal likelihood function, so that the probability of an observation x given a true value T, $p(x \mid T)$, is normally distributed with mean T and with a smaller variance. Then the mean of the posterior distribution $p(T \mid x)$ is cx, where $c < 1$. The value of c depends on the relative variances of the prior and likelihood distributions. The normal distribution is "conjugate," so that the resulting distribution is normal. Its mean is a proportion c of its original mean, where the proportion depends only on the relative variances of the likelihood function and the prior. Thus, our simple assumptions lead to the conclusion that individual ignorance leads to a regression of log odds toward 0 by a constant proportion. This is the same as saying that the log odds is multiplied by c.

What effect does this have on the need for transformation? Assuming the transformation function in Equation (1), then (letting t stand for $t(p)$) the expression for the odds of t is

$$\frac{t}{1-t} = \frac{p^a/(p^a + (1-p)^a)}{1 - p^a/(p^a + (1-p)^a)} = \frac{p^a}{(1-p)^a},$$

so the log odds is

$$\log \frac{t}{1-t} = a\log p - a\log(1-p)$$
$$= a[\log p - \log(1-p)]. \quad (2)$$

In other words, the constant a in Equation (1) amounts to multiplication of the log odds by a. Thus, if the log

---

[3] The argument does not depend on it being log odds, just that it is unbounded.

odds have already been multiplied by c, then to undo the effect of this regression and recover what the log odds would have been without it we must let $a = 1/c$.

If this regression were the only reason for distortion, then under the other assumptions stated the optimal value of a should tell us how much regression was done. Specifically, if we start with the median forecast and find that we can improve the estimate optimally with a transformation using a, then $1/a$ is an estimate of c, the amount of regression. If all forecasters know that they have all the available information, then they should not regress at all and a should be 1.

Note that this model assumes that forecasters are regressing optimally. In this case, a transformation of the median that undoes the effect of their regression should recover the signal S_i, provided that the group of forecasters is large and that the errors resulting from incomplete information are unbiased (i.e., equally likely to favor one option or the other, relative to full information). We have no way to know whether regression is optimal, too much, or too little. Importantly, it could be just right. But it does not need to be optimal for the transformation to recover S_i. For example, if forecasters did not regress at all but were responding to information that was on the whole biased against the actual outcome, the bias would have an effect similar to that of regression.

4. Another Way to Think About Regression

In general, we may think of probability judgments as consisting of two parts, a judgment of what will happen and a judgment of confidence. Thus, a probability of 0.40 for some event means, "I think it will not happen, with 60% confidence." Looking at it this way, the median forecast of a group of forecasters with access to different information should be correct (on the correct side of 0.5) more often than the median probability would warrant. This is because the median probability represents the confidence of a single typical forecaster, but the group does better than that by effectively pooling information.

Thus, the regression we propose is not a bias with respect to the question forecasters are asked. We suggest

that forecasters are taking into account the information that they have, which they assume to be less than what is available. Forecasters who do this can be well calibrated individually. But the combined forecast of the group can be more accurate, more likely to be in the correct direction, than that of the average individual. This is why we need to extremize the average of the individuals. Individuals will appear to be under-confident if we suppose that they are providing confidence judgments for the group, assuming that they do not suffer much from individual over-confidence (which would reduce or reverse the apparent under-confidence) and that the average will benefit from the fact that different forecasters have different information. We do not usually ask for judgments of confidence in the group, so this apparent under-confidence is not an error. Thus, rather than talking about individual regression or shrinkage, we can talk about group progression or expansion, starting with the individual forecasts as the baseline. The expansion factor would be a function of the difference between the group's data and the individual's data, and it can be expressed as a ratio. The reciprocal of this ratio is the regression constant we have already described.

The difference between group and individual data would depend on whether the individuals are independent, e.g., whether they share data with each other, how thoroughly each individual samples the available data, and the number of individuals. If we assume independence and random sampling of available data, we could calculate the effect of the number of individuals. (Again, we do not pursue this here.) In the limit, the group would have all available information.

We could also think of irreducible uncertainty within this framework by supposing that some information is not "available"; there is nothing anyone can do to get it. Thus, even all the available information does not allow perfect prediction of which outcome will occur. The signal S_i would be some distance from 0 or 1.

5. A Simulation

The function in Equation (1) can reverse the effect of regression due to uncertainty about missing information. Can the same function also reverse the effect of scale distortion resulting from the asymmetry of the error distribution for probabilities? The answer

seems to be yes, at least to a very close approximation. We demonstrate this with a simulation. The point of the simulation was to generate what mean and median probability judgments would look like under the assumptions we have outlined: judgments result from a signal plus noise that is normally distributed in log odds space, and from a regression toward 0 (in log odds), then transformed to probability.

The simulation gives us sufficient confidence that Equation (1) can serve both purposes: correcting for the regression (which follows from the assumptions of the last section) and approximately correcting for the distortion of the error distribution.

The simulation (presented in detail in the R script in Appendix C) proceeds as follows:

1. Generate a set of 100 "signals" S, which are the best probabilities as defined earlier, ranging from 0.500 to 0.995. (Probabilities below 0.5 would just be the mirror image of these, so would not change any results.)

2. Transform these to log odds. Now the numbers range from 0 to 5.29 (log odds of 0.995).

3. Replicate this 100-item vector 100 times, yielding a 100-by-100 matrix. Each column is the original vector of signals. Each row is one of the numbers in that column. The entire first row is 0; the last row is all 5.29.

4. Add noise to each row. The noise is the same for each row. The basic noise is a vector of 100 normal quantiles of $0.005, 0.015, 0.025, \ldots, 0.995$. This vector thus ranges from -2.58 to 2.58 and is normally distributed. For each run of the model, we multiply this basic noise vector by a constant before adding it. The constant ranges from 0 to 9. It corresponds to the standard deviation of the noise that we add. The entries in the matrix now represent judgments in log-odds space before any regression, including the normally distributed error that we have assumed. Each row corresponds to a different signal. Each row has a mean of 0, and a standard deviation that we have specified for this run (0–9).

5. Multiply the entire matrix by c, the constant that indicates the amount of regression. If c is 1, there is no regression. For different runs of the model, c took values of 0, 0.2, 0.4, 0.6, and 0.8.

6. Transform these aggregates back to probabilities.

7. Aggregate the judgments in each row by averaging. These averages represent the average log odds for each of the signals between 0.5 and 0.995. We also aggregate by taking the median.

8. Find the squared deviation of each aggregate (mean and median) of each row from its corresponding signal probability.

9. Find the optimal transformation constant a that mimimizes the sum of these deviations. Specifically, we apply Equation (1) (which is called *ptrans* in the script) to the mean and median and optimize its constant a so as to optimally recover the signals. We minimize the sum of the squared deviations from the signals.

In sum, we manipulated the noise and regression constant and determined the best fitting constant a in each case, and this tells us how the optimal value of a, the amount of transformation, depends on the two factors of interest, the standard deviation of the noise and the regression constant (c).

To examine the performance of the model for different amounts of regression and noise, we generated 50 cases: 10 different values of noise crossed with five different values of the regression constant c. The noise values ranged from 0 to 9. Specifically, as noted, we started with a normal distribution with mean 0 and standard deviation (s.d.) of 1, then multiplied this by the values 0–9. (Remember that this is now in terms of log odds, not probability.) The values of c were (as noted) 0, 0.2, 0.4, 0.6, and 0.8. The value of 0, of course, corresponds to no regression. We used the median as well, which should (and does) eliminate any effect of the noise. Notice that, for 0 noise, or the median, the optimal transformation is exactly the reciprocal of c.

Figure 2 shows the value of a for the optimal transformation as a function of the two factors, for the mean and median. Note that the optimal transform for the median is unaffected by noise, so the values for the median correspond to the intercepts where the noise is 0. For the median, the optimal transformation is determined by the regression, and it follows $a = 1/c$ perfectly. This is also the result for the means with zero noise. This reciprocal relation is approximately true for the other values of the noise, but the transformation has less of an effect on a as the noise increases.

The fit of the model to the simulated data is very good. All deviations from perfect fit (mostly those in the upper right of Figure 2) are less than 0.005 on the probability scale. Various attempts to improve the fit of the model by using more quantiles (up to 10,000) failed to improve it, so we suspect that the fit is not exact. For practical purposes, the fit seems to be good

Figure 2   Optimal Transformation Constant a as a Function of Noise and Regression, for the Aggregated Mean
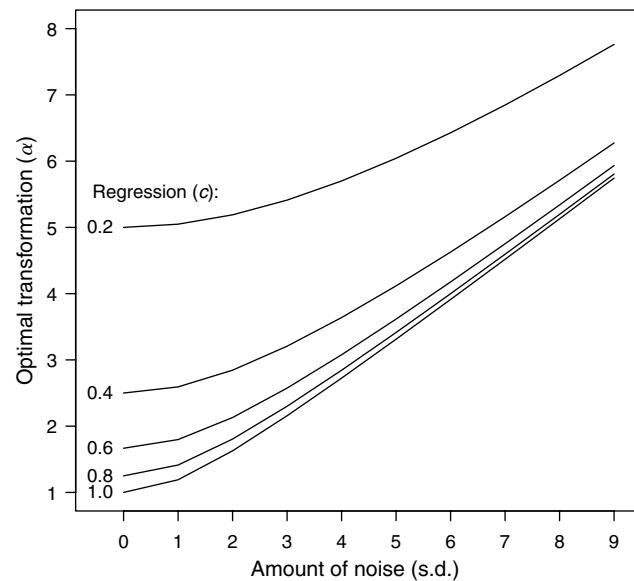
Note. The optimal constants for the median correspond to those for zero noise.

enough, since the assumption of normally distributed error is itself only an approximation.

6. A Demonstration from Data

We conducted a study in which we recruited more than 2,000 people to estimate the probabilities of the outcomes of dozens of international events such as political elections. (The study and its results are described in detail by Mellers et al. 2014.) In the first year, 1,973 forecasters responded to 102 questions; 87 questions were resolved;[4] and the mean number of questions attempted by each forecaster was 47; the mean number of days each question was available was 56. An example of a question was, "Will Italy's Silvio Berlusconi resign, lose re-election/confidence vote, or otherwise vacate office before January 1, 2012?" They provided a probability judgment for the outcome of this question. All forecasters were given some instruction in the idea of calibration, and some forecasters were given more extensive training. Forecasters knew that we would be aggregating their judgments to produce an overall

---

[4] We omitted one question because it was clearly a near miss, leaving 86 questions.

prediction for each event, and that we were competing with other teams who were doing the same.[5]

To get an idea how well the model works in estimating the two parameters for real data, we estimated the optimal transformation constant a for four cases, mean, and median for self-reported expert and nonexpert forecasters.[6] Specifically, we found the value of a that would minimize the mean Brier score across the 80 questions that had two options.[7] (Generalization to more than two options should be straightforward but was not attempted.) Expertise was defined in terms of a self-rating item asked for each of the 80 questions, in which forecasters rated themselves on a five-point scale. These expert ratings were useful: average Brier scores were better if we used them as weights, with the lowest levels of expertise getting no weight at all. For the present demonstration we defined the expert versus nonexpert groups by a median split for each question.

Table 1 shows the optimal transformations (a) and the average Brier scores (that we minimized) for the four conditions, as well as the Brier scores with no transformation. As expected, experts require less transformation, presumably because they think that they have more information and do not regress so much. Likewise, the median requires less transformation than the mean, because it does not require correction for error. The fact that the median requires less transformation, and is thus more extreme, supports the assumption that the error distribution is asymmetric in probability space. Comparison of the right-most two columns shows the benefit of the transformation, which is of course greater when a is greater.

[5] We emphasize that the results reported here are incomplete in several ways even for the first year of a study that is in its third year at the time of writing. The results reported here are intended as illustrative, for the present discussion.

[6] To deal with missing forecasts on a given day, we used a decay constant of 0.6. That is, the mean we report was a weighted mean, in which each day's forecast had its weight multiplied by 0.6 as the day passed. (Hence, on the second passing day, the weight was 0.36.) If the forecaster made a new forecast, all previous weights went to 0. In other analyses of these data, we used other weights as well, but the purpose of the present analysis is just to demonstrate the role of the median in estimating parameters.

[7] We used the optimize( ) function in R to do this. We omitted one additional question that seemed to be misleading (and in which the mean probability was on the wrong side of 0.5).

Table 1 Optimal Transformations (a) and Brier Scores (BS) for Self-Reported Experts and Nonexperts from Year-1 Data (Two Options Only)

Expertise	Method	a (error)	BS, opt.	BS, $a=1$
High	Mean	2.43 (0.05)	0.148	0.187
High	Median	1.78 (0.03)	0.160	0.176
Low	Mean	3.08 (0.09)	0.139	0.196
Low	Median	2.34 (0.06)	0.153	0.184

Note. BS is given for the optimized value of a (shown here) and for $a=1$ (no transformation). Error estimates are shown in parentheses. The error is estimated by leaving out one of the 86 questions at a time and redoing the optimization. The numbers reported are the standard deviations of these values. We did not fit the model to each question because some questions had very surprising resolutions, hence negative optimal values of a, whereas others had unsurprising resolutions, leading to extremely high optimal values of a. The idea was to find a value of a that would minimize the Brier score given the existence of both kinds of cases, in some proportion.

It would appear from Table 1 that the difference between mean and median is consistent across experts and nonexperts. This would imply that, according to our model, the noise is the same for experts and nonexperts. However, Figure 3 tells a different story. The curved lines here are derived from simulations like

Figure 3 Optimal Transformation Constant a as a Function of Noise and Regression, for the Aggregated Mean for Experts and Nonexperts
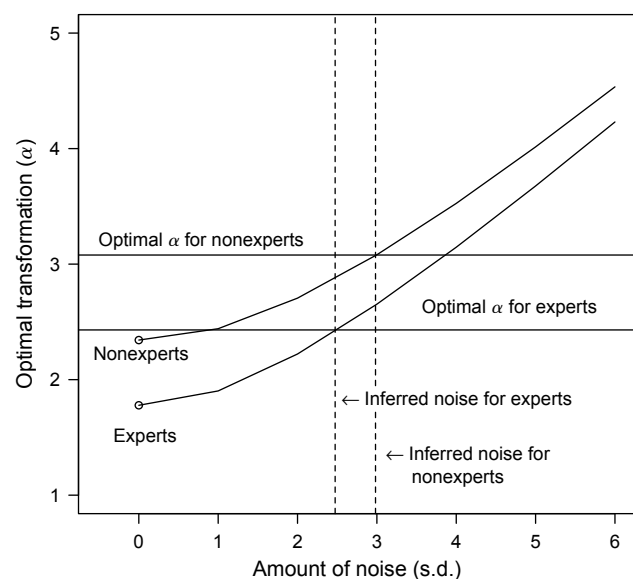
Note. The optimal constants for the median correspond to those for zero noise and are indicated with small circles. The horizontal lines are the optimal transforms for means of experts (lower) and nonexperts (higher). The vertical dashed lines are the inferred amounts of noise (lower for experts).

the lines in Figure 2, but they are generated with the reciprocal of the optimal transformations for the medians (1.78 for experts, 2.34 for nonexperts). The small circles (for zero noise) represent the optimal transformations that result from the simulation for these medians. The two horizontal lines are the optimal transformations for the means, as shown in Figure 2, the higher line corresponding to nonexperts (3.08) and the lower to experts (2.43). To estimate the noise for experts and nonexperts, we can see where these horizontal lines intersect the curved lines. The vertical dashed lines are drawn through the intersections. It is apparent that the noise for the nonexperts (2.98) is greater than that for the experts (2.47). This is a reasonable conclusion. The experts presumably were basing their judgments on a larger proportion of the available evidence, so the nonexperts might have differed more in what evidence they had, and this would show up as noise in our model.

7. Discussion

The main suggestion here is that we can decompose the need for transformation of aggregated probability judgments into two factors. One factor is variation among forecasters, which may be seen as random error producing probability judgments. Because this error is asymmetric near the ends of the probability scale, it will distort the mean forecast toward 0.5. A transformation that extremizes the average of these judgments can approximately undo the effect of this error on the mean probability forecast. The approximation is very close under reasonable assumptions. This factor is not relevant for the median, because we assume that errors are equally likely on both sides of the signal, the optimal probability.

The other factor is the effect of forecasters knowing that they lack information that is potentially available. They thus start with the judgment based on the information they have and then regress it toward 0.5. If a group of forecasters as a whole has some of the missing Information (i.e., if different forecasters are missing different information) then the group's aggregated forecast will be more likely to be correct (on the correct side of 0.5) than the forecast of the average forecaster. We compensate for this regression, again, by extremizing the average forecast. We must do this for the median too, because the regression affects both

mean and median. We thus explain why the median, as well as the mean, can sometimes benefit from an extremizing transformation.

Although the amount of transformation does not seem to be a simple function of the two factors, it is possible to use graphical interpolation to infer the amount of noise, as done in Figure 3.

Note that we can undo the effect of the first factor (error) in other ways. In particular, we can transform judgments before averaging them, e.g., by transforming probability to log odds (as done by Satopää et al. 2014).

We might also try to remove the second factor (regression) by asking different questions. First, it might be useful to change our basic question into a semantically equivalent form (but one that is not necessarily equivalent psychologically); i.e., instead of "What is the probability for X?" we ask "Do you think X will happen or not?" and "What is the probability that you are correct?" Then we could ask an additional question, "What is the probability that you would be correct if you had all the available information?" Note that this question does not assume that the answer to the first question (whether X would happen) is the same.[8] Note that such a question would very likely have more error, because it requires an estimate of what information is available.

In the results shown, a single transformation constant was used for each condition. However, it is unlikely that this is optimal. We have already noted that expertise can affect the optimal transformation, and we did not even break expertise down as far as we could. We also expect that some problems have higher IU (irreducible uncertainty) than others and thus should require less transformation (because forecasters know that they are not missing information that might be available). Estimating the IU of a problem is difficult, in part because the intrinsic uncertainty changes over time, generally decreasing as a question comes closer to being resolved.

Future research should seek more direct measures of the two components we have postulated: error and

---

[8] We could also ask something like, "What is the probability that the group average would be on the correct side of 0.5?" This asks the forecaster to consider not only how much information is missing but also how much the group, together, has available in a way that can affect the group average. Alternatively, we could ask for a direct judgment of the proportion of the information potentially available to others that is available to the forecaster.

the benefit of aggregating that stems from differences in information. Error might be measured by variance among judges. The benefit of aggregation will be greater when judges are less correlated with each other (e.g., over time). Another line of possible future research is to find ways to test the crucial assumption that error is equally likely on both sides of the signal; such a test might involve an explicit and testable hypothesis about why this assumption is false.

Acknowledgments

Appendix A. The Best Probability Judgment

This appendix discusses what it might mean to say that a probability judgment is optimal given a set of information. It may help to clarify our assumptions and distinguish them from other concepts in common use.

In standard accounts (e.g., Yates 1990), probability judgments are often evaluated for correspondence (with reality, as distinct from internal coherence) by two different criteria: calibration and discrimination. Calibration is (roughly) a measure of whether probability judgments are in the correct position in the 0–1 interval: judgments of 0.8 should correspond to true propositions 80% of the time. Discrimination is a measure of how well a judge discriminates true from false propositions. Discrimination can be perfect even when calibration is terrible (e.g., saying 0.02 for all true propositions and 0.01 for all false ones), and vice versa (e.g., predicting rain with probability 0.25, every day, in a place where it rains on 25% of all days). Both of these criteria are obviously relevant, but they represent two measures of goodness, and we need one measure to define the best judgment.

Scoring rules, such as the Brier score, do provide such a measure, but they are not defined contingently on a set of available information. The way to get the best score (the lowest score, for the Brier score) is to report a probability of 1 for true propositions and 0 for false ones. Yet the information required to do this is typically unavailable.

The usual rationale for scoring rules is to provide incentives against distorting well-calibrated judgments and in favor of using all relevant and available information. Given this rationale, it would make sense to define the best probability judgment as the one that would get the best score. We cannot know this for a single judgment, so let us generalize this

statement slightly to say that the best judgment is the one made using (mental and possibly computational) processes that get the best *expected* score, compared to all other possible processes. Note that this criterion is difficult to apply in practice, but here we are trying to define a concept philosophically. The point is to show that the concept has meaning. With real data, we can approximate it.

The next problem is that there are many scoring methods, and they need not give the same answer to this question (although most of them will surely agree closely). Here we provide an approach to finding a good score. We assume that a probability judgment is an input to a decision made by someone other than the judge. A good score should be a function of the expected utility of the decision. Johnstone et al. (2011) take this approach, but their interest is in cases in which the utility function of the decision maker is known. We assume that the judge does not know the decision maker's utilities. Similar approaches, reaching similar conclusions, are those of Hernández-Orallo et al. (2012) and Murphy (1966).

We illustrate this approach by making some assumptions that we think are reasonable approximations in most cases. These assumptions lead to the Brier score as optimal. This is illustrative, and we do not explore the effects of other assumptions, or whether other commonly used scores could be justified by this approach at all.

Consider a choice of the sort that occurs frequently in medicine (Pauker and Kassirer 1980). A doctor has two options: treat or do not treat a particular condition. The proposition of interest is that the condition is present. The doctor asks a specialist for a probability judgment P about this proposition. In making her decision, she considers two disutilities (negative utilities): the disutility of treating when the condition is absent, D_f (f for false alarm) and that of not treating when it is present, D_m (m for miss). Correct decisions are assumed to have a disutility of 0. These two utilities allow the doctor to define a probability threshold T, so that her rule is to treat if $P > T$. It is easy to see that $T = D_f/(D_m + D_f)$. (With probability T, the expected disutilities of treating and not treating are equal.) Each decision has its own threshold.

Now let us make two simplifying assumptions for the class of decisions at issue. The first is that their thresholds are uniformly distributed in the 0–1 interval. This uniformity assumption is surely an approximation, but it seems reasonable to us. Moreover, it is consistent with some intuitive assumptions. Specifically, imagine that some decisions are "lopsided" in terms of differential disutilities and others are not. The lopsidedness results from the fact that utilities of outcomes are highly variable. Most are very small, but a few are huge. A distribution with this property is the exponential, which has a peak at zero. If we assume that D_m and D_f are exponentially distributed across all decisions, then the ratio $D_f/(D_m + D_f)$ is uniform, as we show in Appendix B. The idea that utilities are exponentially distributed might be further justified by the assumption that utilities are the result of the successive addition of small utility increments

of the same size, each added with the same probability at each step regardless of how many increments have been added already, and the addition process stops when a step is reached at which no increment is added. The exponential distribution is "memoryless." It is usually applied in this way to waiting times, although it has also been applied to incomes (Drăgulescu and Yakovenko 2001), which are bounded at 0, like utilities.

The second simplifying assumption is that for each decision $D_m + D_f = 1$. For present purposes, this is equivalent to assuming that the total of both disutilities is independent of the threshold T; if this is true, then we can rescale all disutilities so that their sum is 1 without changing any conclusions. This assumption seems reasonable because it says, roughly, that the overall importance of a decision is independent of T. This situation could arise if there is some mechanism that brings decisions to our attention, and it has some criterion of importance. It is reasonable to think that this criterion would depend on the sum $D_m + D_f$.

These two assumptions imply that $D_f = T$ and $D_m = 1 - T$.

Let us go one step further and ask about the expected disutility as a function of P. To find the expected disutility of say P, hold P constant and suppose (as assumed) that T varies uniformly between 0 and 1. And let us assume (without loss of generality) that the proposition at issue is true, e.g., the patient has the condition that should be treated. If $T < P$, there is no loss, because the correct option is chosen. But, if $T > P$, the incorrect option will be chosen. We thus need to consider values of T between P and 1.

The disutility of an error (D_m, the disutility of missing a true case) at threshold T will be $1 - T$. Thus, when T is close to 1, the cost will be very small, but when T is close to P, the cost will be close to $1 - P$. To get the expected disutility, we integrate over all the different values of T between P and 1. In the general case, the expected disutility would be proportional to $\int_P^1 D_m f(T)\,dT$, where $f(T)$ is the probability density function of T, which we have assumed is uniform, so that this term is not needed. But, even without the assumption of uniformity, $D_m = 1 - T$, so this expression becomes $\int_P^1 (1 - T) f(T)\,dT$. We have, in essence, a right triangle, with base and height of $1 - P$. The total area of this triangle, the expected loss, is $(1 - P)^2/2$. This is proportional to the Brier score for this case. The Brier score can thus serve as an estimate of the expected utility loss that results from giving the wrong probability: in this example, giving P when it should be 1. In other words, the quadratic feature arises because there are two effects: as P gets farther from 1, more mistakes are made, and the cost of each additional mistake increases linearly with the distance from 1 (by assumption).

Under these assumptions, we can think of the best probability judgment as the one made with the processes that would minimize the Brier score for the class of judgments that are of interest. Surely, the processes will depend somewhat on the domain. For practical purposes, we can find a lower bound on the best judgment by finding the best method

for eliciting and aggregating probability judgments from relevant experts in the domain of interest, so as to minimize the Brier score. And, with good methods and good judges, the lower bound should not be much lower than that for the "signal" as we have described it. Any remaining uncertainty is irreducible.

Appendix B. The Uniform Distribution of Thresholds

Given two independent, identically-distributed exponential distributions, X and Y, we will show that $X/(X + Y) \sim$ Uniform$(0, 1)$.

Given $X \sim \lambda e^{-\lambda x}$ and $Y \sim \lambda e^{-\lambda y}$, we note that X, $Y \sim$ Gamma$(1, 1/\lambda)$.

Now we define

$$u = X + Y \quad \text{and} \quad v = \frac{X}{X + Y},$$

where $u \in (0, \infty)$ and $v \in (0, 1)$, which means we have $x = uv$ and $y = u - uv$.

We calculate the determinant of the Jacobian to be $-u$, and then determine the joint density of X and Y:

$$f(x, y) = \lambda e^{-\lambda x} \lambda e^{-\lambda y} = \lambda^2 e^{-\lambda(x+y)},$$

and which, upon substituting in the above values for x and y, gives us

$$f(u, v) = \lambda^2 e^{-\lambda(uv+(u-uv))} |-u| = \lambda^2 e^{-\lambda(x+y)} (x + y).$$

Now we integrate out u over its domain to find our desired distribution:

$$f\left(v = \frac{x}{x + y}\right) = \int_0^\infty \lambda^2 u e^{-\lambda u}\,du.$$

We can see that this is a complete Gamma$(2, 1/\lambda)$ kernel, so $f(v) = 1$, and is defined only on the interval $(0, 1)$. Therefore

$$\frac{X}{X + Y} = \text{Beta}(1, 1) \equiv \text{Uniform}(0, 1).$$

If Beta$(1, 1)$,

$$f(v) = \frac{\Gamma(1)\Gamma(1)}{\Gamma(1+1)x^{(1-1)}(1-x)^{(1-1)}} = 1.$$

Appendix C. R Script for Simulation of Optimal Transformations

```
# Signal probabilities Sig are the answers based on
  the best data.
Sig <- (0:99)/200 + 0.5

# convert these to log odds. Use plogis() to unconvert.
logodds <- function(x) {log(x/(1-x))}
ptrans <- function(p,a) {p^a/(p^a+(1-p)^a)}
  # transformation function
Lsig <- logodds(Sig)
```

```
# create matrix, 100 forecasts at each value
Lsigmat0 <- matrix(rep(Lsig,100),100,100)

# computes values used for optimization, as a
  function of Noise and Reg
Setup <- function(Noise,Reg) {
  Lsigmat <- t(t(Lsigmat0)+Noise
    *qnorm((0.5:99.5)/100)) # add noise
  Lsigreg <- Reg*Lsigmat # regress, multipying by Reg
  Respsreg <<- plogis(Lsigreg) # de-transformed
  Medreg <<- apply(Respsreg,1,median)
    # de-transformed aggregated median
}

# functions to minimize for optimization
f2 <- function(x)
  {sum((Sig-ptrans(rowMeans(Respsreg),x))^2)}
f4 <- function(x) {sum((Sig-ptrans(Medreg,x))^2)}

# set up matrix for results of optimization for the
  50 values
Tests <- matrix(NA,50,4)
colnames(tests) <- c(``Noise,'' ``Regression,'' ``a,''
  ``error'')
Tests[,1] <- rep(1:10,5)
Tests[,2] <- rep(c(1,0.8,0.6,0.4,0.2),
                  c(10,10,10,10,10))

# fill the right two columns of the matrix
for (i in 1:50) {
  Setup(Tests[i,1],Tests[i,2])
  Tests[i,3:4] <- unlist(optimize(f2,
                          interval=c(0,20)))
}
```

References

Ariely D, Wing-Tung A, Bender RH, Budescu DV, Dietz CB, Gu H, Wallsten TS, Zauberman G (2000) The effects of averaging subjective probability estimates between and within judges. *J. Experiment. Psych.: Appl.* 6(2):130–147.

Drăgulescu A, Yakovenko VM (2001) Evidence for the exponential distribution of income in the USA. *Eur. Phys. J. B* 299(1–2): 585–589.

Erev I, Wallsten TS, Budescu DV (1994) Simultaneous over- and underconfidence: The role of error in judgment processes. *Psych. Rev.* 101(3):519–527.

Fox CR, Ülkümen G (2011) Distinguishing two dimensions of uncertainty. Brun W, Keren G, Kirkebøen G, Montgomery H, eds. *Perspectives on Thinking, Judging, and Decision Making: A Tribute to Karl Halvor Teigen* (Universitetsvorlaget, Oslo, Norway), 21–35.

Hernández-Orallo J, Flach P, Ferri C (2012) A unified view of performance metrics: Translating threshold choice into expected classification loss. *J. Machine Learn. Res.* 13(October):2813–2869.

Johnstone DJ, Jose VRR, Winkler RL (2011) Tailored scoring rules for probabilities. *Decision Anal.* 8(4):256–268.

Juslin P, Winman A, Olsson H (2000) Naive empiricism and dogmatism in confidence research: A critical examination of the hardeasy effect. *Psych. Rev.* 107(2):384–396.

Karmarkar US (1978) Subjectively weighted utility: A descriptive extension of the expected utility model. *Organ. Behav. Human Performance* 21(1):61–72.

Mellers BA, Ungar L, Baron J, Ramos J, Gürçay B, Fincher K, Scott S, et al. (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psych. Sci.* Forthcoming.

Murphy AH (1966) A note on the utility of probabilistic predictions and the probability score in the cost-lost ratio decision situation. *J. Appl. Meteorology* 5(4):534–537.

Pauker SG, Kassirer JP (1980) The threshold approach to clinical decision making. *New England J. Medicine* 302(20):1109–1117.

Satopää VA, Baron J, Foster DP, Mellers BA, Tetlock PE, Ungar LH (2014) Combining multiple probability predictions using a simple logit model. *Internat. J. Forecasting* 30(2):344–356.

Shlomi Y, Wallsten TS (2010) Subjective recalibration of advisors' probability estimates. *Psych. Bull. Rev.* 17(4):492–498.

Turner BM, Steyvers M, Merkle EC, Budescu DV, Wallsten TS (2013) Forecast aggregation via recalibration. *Machine Learn.*, ePub ahead of print October, http://link.springer.com/article/10.1007%2Fs10994-013-5401-4.

Wallsten TS, Diederich A (2001) Understanding pooled subjective probability estimates. *Math. Soc. Sci.* 41(1):1–18.

Yates JF (1990) *Judgment and Decision Making* (Prentice Hall, Englewood Cliffs, NJ).

Jonathan Baron is a professor of psychology at the University of Pennsylvania. Baron's research examines judgments and decisions about public policies, specifically, what sorts of judgments, when put into effect through politics, prevent governments from doing the most good. Baron is the sole author of *Rationality and Intelligence* (1985); *Thinking and Deciding*, 4th ed. (2008), a widely used textbook for advanced undergraduates and beyond; *Morality and Rational Choice* (1993); *Judgment Misguided: Intuition and Error in Public Decision Making* (1998); and *Against Bioethics* (2006). He has also co-edited three books and published more than 200 articles and chapters. He is editor of the journal *Judgment and Decision Making* and past president of the Society for Judgment and Decision Making. He holds a B.A. from Harvard (psychology, 1966) and a Ph.D. from University of Michigan (experimental psychology, 1970), and he is a fellow of the American Association for the Advancement of Science, the Association for Psychological Science, the Eastern Psychological Association, and the Society of Experimental Psychologists.

Barbara A. Mellers is the Heyman University Professor at the University of Pennsylvania with appointments in School of Arts and Sciences and the Wharton School. She received a Ph.D. in psychology from the University of Illinois at Urbana–Champaign. Her research examines human judgment and decision making. She has studied the effects of emotions on choice and the effects of context and question format on judgments. Her primary interest has been to understand why people deviate from basic laws of probability, principles of logic, and axioms of choice, and she has developed models to capture the way people actually make judgments and decisions. In recent years, she has worked to develop better methods of eliciting judgments of uncertainty and aggregating

over multiple judgments to arrive at the most accurate possible predictions of geopolitical events.

Philip E. Tetlock is the Annenberg University Professor at the University of Pennsylvania (with cross-appointments in School of Arts and Sciences and the Wharton School). He has long-standing interests in the empirical and conceptual challenges of pinning down the elusive concept of "good judgment" in real world as well as laboratory settings. He is one of the principal investigators in the Good Judgment Project, a forecasting tournament sponsored by the Intelligence Advanced Research Projects Activity, which was the source of data for this article.

Eric Stone is a statistical programmer for the joint University of Pennsylvania, University of California, Berkeley team in the IARPA Forecasting World Events tournament. He received his M.S. in statistics from Temple University in 2012, where he was a university fellow. He has presented on and developed data sonification software, and he has

conducted research on identifying individual forecasting skills. Previously, he served as a graduate intern at the USDA's National Agricultural Statistics Service and worked as a project director for market research and consulting firm ORC International. He earned his B.A. in psychology from Oberlin College in 2007.

Lyle H. Ungar is a professor of computer and information science at the University of Pennsylvania, where he also holds appointments in multiple departments in the schools of Engineering, Arts and Sciences, Medicine, and Business. Dr. Ungar received a B.S. from Stanford University and a Ph.D. from MIT. He has supervised 20 Ph.D. dissertations, published more than 200 articles, and is co-inventor on 11 patents. His research areas include machine learning, data and text mining, and psychology, with a current focus on statistical natural language processing, spectral methods, and the use of social media to understand the psychology of individuals and communities.