Judgment and Decision Making, Vol. 8, No. 1, January 2013, pp. 25–27

A cautionary note on global recalibration

Joseph B. Kadane* Baruch Fischhoff†

Abstract

We report a mathematical result that casts doubt on the possibility of recalibration of probabilities using calibration curves. We then discuss how to interpret this result in the light of behavioral research.

Keywords: calibration, coherence, probability, judgment.

1 Recalibration

There has long been interest in bridging the gap between the subjective and objective interpretations of probability. The subjective view, championed by DeFinetti (1937/1964) and Savage (1954), takes statements of probability to indicate the beliefs of the person providing them, as would be reflected in their willingness to bet on fair gambles using those values. The objective view holds that probability is a function of the external world and, hence, that a subjective (or personal) probability can be wrong in some objective sense.

Both views agree, however, on the following laws of probability:

(i) for all events A, $P\{A\} \geq 0$

(ii) if S is the sure event, $P\{S\} = 1$

(iii) if A and B are disjoint events (i.e., they can't both happen), then

$$P\{A \cup B\} = P\{A\} + P\{B\},$$

where $A \cup B$ is the event that either A or B occurs. If P assigns numbers to events in such a way that these assumptions are satisfied, P is said to be coherent.

One point of intersection of subjective and objective views is in studies of calibration. As expressed by Lichtenstein, Fischhoff and Phillips (1982, p. 307), "A judge is calibrated if, over the long run, for all propositions assigned a given probability, the proportion that is true equals the probability assigned. Judges' calibration can be empirically evaluated by observing their probability assessments, verifying the associated propositions, and then observing the proportion true in each response category."

They go on to propose that "calibration may be reported by a *calibration curve*. Such a curve is derived as follows:

1. Collect many probability assessments for items whose correct answer is known or will shortly be known to the experimenter.

2. Group similar assessments, usually within ranges (e.g. all assessments between .60 and .69 are placed in the same category).

3. Within each category, compute the proportion correct (i.e. the proportion of items for which the proposition is true or the alternative is correct).

4. For each category, plot the mean response (on the abscissa) against the proportion correct (on the ordinate)."

Empirical studies typically find that the calibration curve deviates, often substantially, from the identity line. In such studies, the subjective probability judgments are evaluated by an objective standard, the observed proportions of correct answers. (For research into the behavioral and measurement properties of these curves, see O'Hagan, Buck, Daneshkhah, Eiser, et al. 2006; Brenner, Griffin & Koehler, 2005; Budescu & Johnson, 2011; and the articles in the special issue of the *Journal of Behavioral Decision Making* [Budescu, Erev & Wallsten, 1997].) Observing such miscalibration, a natural reaction is to propose a *global* recalibration of such probability judgments to more accurate ones. According to this logic, if I know my calibration curve, f, and I was about to give p as my probability of an event, I should instead give my recalibrated probability $f(p)$, adjusting for the expected miscalibration. This suggestion runs into a fundamental mathematical difficulty:

*Department of Statistics, Carnegie Mellon University, Baker Hall, Pittsburgh, PA 15213. Email: kadane@stat.cmu.edu.

†Department of Social and Decision Sciences, Department of Engineering and Public Policy, Carnegie Mellon University. Email: baruch@cmu.edu.

Proposition 1

a) Suppose recalibration depends only on the subjective probability of the event. That is, every event judged before recalibration to have probability p is recalibrated to $f(p)$. Suppose, in addition, the following are true:

b) the subjective probabilities obey the rules of probability given above

c) the recalibrated probabilities obey the rules of probability

d) there are n exhaustive and mutually exclusive events I regard as equally likely.

Then

$$f(k/n) = k/n \text{ for } 1 \leq k \leq n. \quad (1)$$

Proof of Proposition 1:

By *a)* I assign equal probability to the events in d), namely $1/n$. Recalibrated, I assign $f(1/n)$ to each. Because of *c)*, I must have

$$nf(1/n) = 1, \quad (2)$$

i.e. $f(1/n) = 1/n$.

Now consider an event that is the union of $k \leq n$ of the original events. By *b)*, I assign this event probability k/n. Recalibrated, it now has probability $f(k/n)$. But by *c)*, using (2), I have that

$$f(k/n) = kf(1/n) = k/n. \quad (3)$$

□

Proposition 2:

Suppose e) that I regard some random variable X as having a continuous distribution. Then for every positive integer n, there are events satisfying d) above.

Proof of Proposition 2:

Let the cumulative distribution function of X be F. Choose $n \geq 2$. Let

$$\begin{aligned} A_1 &= \{x|F(x) < 1/n\} \\ A_n &= \{x|F(x) \geq (n-1)/n\} \quad (4) \\ \text{and } A_i &= \{x|i/n \leq F(x) < (i+1)/n\}, \\ &\quad \text{for } i = 2, \ldots, n-1. \end{aligned}$$

By construction, the sets A_i are disjoint (meaning that an x can be in only one), exhaustive (meaning that every x is in one of them), and have (uncalibrated) probability

$$P\{X \in A_i\} = 1/n \quad i = 1, \ldots, n. \quad (5)$$

□

Corollary: If e) holds and if assumptions a), b), and c) of Proposition 1 hold, then

$$f(k/n) = k/n \text{ for all } n \geq 1 \text{ and all } k, 1 \leq k \leq n. \quad (6)$$

Proof:

Under e), Proposition 1 shows that assumption d) of Proposition 1 holds for all $n \geq 1$. The conclusion then follows from Proposition 1. □

Equation (6) shows that under these assumptions, the only possible recalibration curve f is the identity function from the rational numbers to the rational numbers in $[0, 1]$.

To explain the import of the conclusion of the corollary, note that:

(i) The identity function is a function such that for each value of the input, the output equals the input.

(ii) The consequence of the corollary is that the only calibration curve that satisfies assumptions *a), b), c)* and *e)* is a straight line from $(0, 0)$ to $(1, 1)$.

Thus, assuming e) holds, if recalibration depends only on the number originally assigned to the event and if the original judgment and the recalibrated one both satisfy the laws of probability, then they are also the same, meaning that no recalibration can occur. Preliminary versions of this result are given in Kadane and Lichtenstein (1982), Seidenfeld (1985) and Garthwaite, Kadane and O'Hagen (2005).

There are four assumptions for this result. The first is that recalibration depends only on the probability assigned. It would not hold say for a weather forecaster who over-predicts rain and, hence, under-predicts not-rain. In such cases, knowing the event (rain or not-rain) may allow recalibration. For example, a forecaster might give rain on a particular day a probability of 80% and, hence, not-rain a probability of 20%. Recalibrating these as 70% and 30% might yield more useful forecasts. What cannot be done, according to the Proposition, is to recalibrate *every* event given probability 80% to be probability 70%. Similarly, an overconfident person might propose an interquartile range that is characteristically too narrow. This means that the overconfident person assigns 25^{th} and 75^{th} percentiles to some unknown quantity that are too close to each other. Then the set outside that range would be characteristically too broad, but both sets would have subjective probability 1/2. Both the over-prediction of rain and the over-confident, too-narrow interquartile range are barred by assumption *a)*.

A second assumption, *c)* above, is that the recalibration results in assessments that obey the laws of probability. Because the recalibrated probabilities are frequencies, this must be the case. The role of assumption d) is to

https://doi.org/10.1017/S1930297500004460 Published online by Cambridge University Press

ensure that there are enough events to work with. Proposition 2 is one way to ensure this. A second is to imagine decomposing large lumps of probability into smaller units. For example, an event of probability .2 could be thought of as 1 out of five equally likely events, but it can also be thought of as 200 of 1000 equally likely events. Proposition 2 applies if I am willing to agree to some random variable having a normal distribution, or a uniform distribution, or a chi-square distribution, or any other continuous distribution imaginable. Hence, attention must be focused on the fourth assumption, *b)* above, examined in the next section.

2 Discussion of assumption *b)*

Probability judgments can be miscalibrated when individuals have coherent, but misinformed beliefs. In such cases (as just shown), their judgments cannot be recalibrated and still be probabilities. Recalibration might, however, be possible if the flawed judgments were not, in fact, probabilities, but just numeric responses elicited on probability-like scales. Knowing whether judgments are probabilities requires evaluating their internal consistency (or coherence). Such tests are, however, rare in calibration studies. When consistency checks are performed, they are typically modest, conducted with the hope of getting a hearing from skeptics poised to dismiss any numeric expression of uncertainty (e.g. Bruine de Bruin, Fischhoff, Brilliant & Caruso, 2006; Fischhoff, Parker, Bruine de Bruin, Downs, Palmgren, Dawes & Manski, 2000).

Those numeric judgments might be made more useful by recalibration, so that "when someone says X%, treat it as Y%." Unfortunately for such simple solutions, experimental psychologists have long known that translating internal states into observable behavior involves complex psychological processes, subject to sometimes subtle details of how judgments are elicited (e.g., Woodworth & Schlosberg (1954)). For example, numeric judgments can be affected by whether response scales use integral or decimal values, where the first judgment falls in the range of possibilities, and what range respondents expect to use (Poulton, 1989, 1994). That can be true even with familiar response modes, such that beliefs may differ when elicited in terms of probabilities and odds (vonWinterfeldt & Edwards, 1986). As a result, there is no unique Y% for each X%.

Such sensitivity to procedural detail also limits the generalizability of the studies used to support recalibration proposals. For example, one common task in those studies asks respondents to choose the more likely of two alternatives (e.g., absinthe is *(a)* a liqueur or *(b)* a precious stone), and then assign a probability (on the 0.5-1.0 range). Those judgments tend to be too high (indicating overconfidence) with relatively hard questions (e.g., 60% correct) and too low (indicating underconfidence) with relatively easy ones (e.g., 80% correct). A plausible explanation is that respondents enter the tasks expecting some difficulty level (e.g., getting 70% correct) and then anchor on that expectation, leaving their judgments too high for hard tasks and too low for easy ones (Lichtenstein et al., 1982). If asked how many they got right after answering the questions, people often provide an answer that differs from their mean response (Sniezek & Buckley, 1991).

Because numeric responses are sensitive to procedural detail, recalibration could make matters worse, rather than better, for example, increasing judgments that should be decreased, depending on whether the task was easy or hard, relative to respondents' expectations. As a result, recalibration requires matching the conditions of an elicitation session to those in studies in which miscalibration has been observed.

A better strategy, though, is to get better-calibrated responses in the first place. The best way to do that is to provide the conditions needed for any learning: prompt, unambiguous feedback, with clearly understood incentives for candid expressions of uncertainty, augmented by whatever insight the research can provide regarding the psychological processes involved in evaluating evidence (Lichtenstein et al., 1982). The consistency checks that are part of formal expert elicitations (Morgan & Henrion, 1990; O'Hagan et al., 2006) should help people to check their work. The success of those efforts, however, is an empirical question, obligating those who elicit judgments to evaluate their coherence and calibration. In the best case, those responses will prove to be probability judgments needing no recalibration. Shlomi and Wallsten (2010) provide a nice example, with observers able to see how and how well judges use probability numbers.

3 Conclusion

Many studies have found probability judgments to be miscalibrated, in the sense that they deviate from observed probabilities of being correct. Seeing that, it might be tempting to recalibrate the probabilities that people give to more realistic ones. This note identifies a limit to the kinds of recalibration that make sense. If one accepts our argument that the requirement that recalibrated probabilities be coherent (assumption c) and assumption e)) are benign, then global recalibration (other than the identity function) entails that the original elicited numbers do not obey the laws of probability. While this is not a surprise, given the results of the huge literature following the path of Kahneman and Tversky, it does mean that

https://doi.org/10.1017/S1930297500004460 Published online by Cambridge University Press

global recalibration is not a remedy for miscalibration of subjective probabilities.

References

Brenner, L., Griffin, D., & Koehler, D. (2005). Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organization Behavior and Human Decision Processes, 97*, 64–81.

Bruine de Bruin, W., Fischhoff, B., Brilliant, L., & Caruso, D. (2006). Expert judgments of pandemic influenza. *Global Public Health, 1*, 178–193.

Budescu, D., Erev, I., & Wallsten, T. (1997). Introduction to this Special Issue on Stochastic and Cognitive Models of Confidence. *Journal of Behavioral Decision Making, 10*, 153–155 (whole issue, 153–285).

Budescu, D. and Johnson, T. (2011). A model-based approach for the analysis of the calibration of probability judgments. *Judgment and Decision Making, 6*, 857–869.

DeFinetti, B. (1937/1964). La prevision: Ses lois logiques, se sources subjectives. *Annals de l'Institut Henri Poincaré*. English translation in H. E. Kyburg, Jr, & H. E. Smokler, (Eds.) *Studies in Subjective Probability*, New York: Wiley, 1964, 7, 1–6.

Fischhoff, B., Parker, A., Bruine de Bruin, W., Downs, J., Palmgren, C., Dawes, R., & Manski, C. (2000). Teen expectations for significant life events. *Public Opinion Quarterly, 64*, 189–205.

Garthwaite, P., Kadane, J., & O'Hagan, A. (2005). Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association, 100*, 680–701, at page 696.

Kadane, J. and Lichtenstein, S. (1982). A Subjectivist View of Calibration. Technical Report 233, Carnegie Mellon University.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*, pp. 306–334. New York: Cambridge University Press.

Morgan, M. and Henrion, M. (1990). *Uncertainty*. New York: Cambridge University Press.

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., & et al. (2006). *Uncertain judgements: Eliciting expert probabilities*. Chichester: Wiley.

Poulton, E. (1989). *Bias in quantifying judgment*. Hillsdale, NJ: Erlbaum.

Poulton, E. (1994). *Behavioral decision research*. Hillsdale, NJ: Erlbaum.

Savage, L. (1954). *Foundations of Statistics*. New York: Wiley.

Seidenfeld, T. (1985). Calibration, coherence and scoring rules. *Philosophy of Science, 52*, 274–294 at page 280.

Shlomi, Y. and Wallsten, T. (2010). Subjective recalibration of advisor' probability estimates. *Psychonomic Bulletin and Review, 17*, 492–498.

Sniezek, J. and Buckley, T. (1991). Confidence depends on level of aggregation. *Journal of Behavioral Decision Making, 4*, 263–272.

vonWinterfeldt, D. and Edwards, W. (1986). *Decision analysis and behavioral research*. New York: Cambridge University Press.

Woodworth, R. and Schlosberg, H. (1954). *Experimental psychology*. New York: Holt.

https://doi.org/10.1017/S1930297500004460 Published online by Cambridge University Press