22. Calibration of probabilities: The state of the art to 1980

Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D. Phillips

From the subjectivist point of view (de Finetti, 1937/1964), a probability is a degree of belief in a proposition. It expresses a purely internal state; there is no "right," "correct," or "objective" probability residing somewhere "in reality" against which one's degree of belief can be compared. In many circumstances, however, it may become possible to verify the truth or falsity of the proposition to which a probability was attached. Today, one assesses the probability of the proposition "it will rain tomorrow." Tomorrow, one looks at the rain gauge to see whether or not it has rained. When possible, such verification can be used to determine the adequacy of probability assessments.

Winkler and Murphy (1968b) have identified two kinds of "goodness" in probability assessments: normative goodness, which reflects the degree to which assessments express the assessor's true beliefs and conform to the axioms of probability theory, and substantive goodness, which reflects the amount of knowledge of the topic area contained in the assessments. This chapter reviews the literature concerning yet another aspect of goodness, called calibration.

If a person assesses the probability of a proposition being true as .7 and later finds that the proposition is false, that in itself does not invalidate the assessment. However, if a judge assigns .7 to 10,000 independent propositions, only 25 of which subsequently are found to be true, there is something wrong with these assessments. The attribute that they lack is called calibration; it has also been called realism (Brown & Shuford, 1973), external validity (Brown & Shuford, 1973), realism of confidence (Adams & Adams, 1961), appropriateness of confidence (Oskamp, 1962), secondary validity (Murphy & Winkler, 1971), and reliability (Murphy,

This is a revised version of the paper that originally appeared in H. Jungermann and G. deZeeuw (Eds.), *Decision Making and Change in Human Affairs.* Dordrecht-Holland: D. Reidel Publishing Co., 1977. Reprinted by permission.

1973). Formally, 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.

The experimental literature on the calibration of assessors making probability judgments about discrete propositions is reviewed in the first section of this chapter. The second section looks at the calibration of probability density functions assessed for uncertain numerical quantities. Although calibration is essentially a property of individuals, most of the studies reviewed here have reported data grouped across assessors in order to secure the large quantities of data needed for stable estimates of calibration.

Discrete propositions

Discrete propositions can be characterized according to the number of alternatives they offer:

No alternatives: "What is absinthe?" The assessor provides an answer, and then gives the probability that the answer given is correct. The entire range of probability responses, from 0 to 1, is appropriate.

One alternative: "Absinthe is a precious stone. What is the probability that this statement is true?" Again, the relevant range of the probability scale is 0 to 1.

Two alternatives: "Absinthe is (a) a precious stone; (b) a liqueur." With the *half-range* method, the assessor first selects the more likely alternative and then states the probability (≥.5) that this choice is correct. With the *full-range* method, the subject gives the probability (from 0 to 1) that the prespecified alternative is correct.

Three or more alternatives: "Absinthe is (a) a precious stone; (b) a liqueur; (c) a Caribbean island; (d) . . . " Two variations of this task may be used: (1) the assessor selects the single most likely alternative and states the probability that it is correct, using a response $\geq 1/k$ for k alternatives or (2) the assessor assigns probabilities to all alternatives, using the range 0 to 1.

For all these variations, calibration may be reported via 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).

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

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).

Perfect calibration would be shown by all points falling on the identity line.

For half-range tasks, badly calibrated assessments can be either *overconfident,* whereby the proportions correct are less than the assessed probabilities, so that the calibration curve falls below the identity line, or *underconfident,* whereby the proportions correct are greater than the assessed probabilities and the calibration curve lies above the identity line.

For full-range tasks with zero or one alternative, overconfidence has two possible meanings. Assessors could be overconfident in the truth of the answer; such overconfidence would be indicated by a calibration curve falling always below the identity line. Alternatively, assessors could be overconfident in their ability to discriminate true from false propositions. Such overconfidence would be shown by a calibration curve below the identity line in the region above .5 and above the identity line in the region below .5.

Several numerical measures of calibration have been proposed. Murphy (1973) has explored the general case of k-alternative items, starting with the Brier score (1950), a general measure of overall goodness or probability assessments such that the smaller the score, the better. The Brier score for N items is:

$$B = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{r}_i - \mathbf{c}_i)(\mathbf{r}_i - \mathbf{c}_i)'$$

where $\mathbf{r}_i$ is a vector of the assessed probabilities for the k alternatives of item i, $\mathbf{r}_i = (r_{1i}, \ldots r_{ki})$, $\mathbf{c}_i$ is the associated outcome vector, $\mathbf{c}_i = (c_{1i}, \ldots, c_{ji}, \ldots, c_{ki})$, where c_{ji} equals one for the true alternative and zero otherwise, and the prime (′) denotes a column vector. Murphy showed that the Brier score can be partitioned into three additive parts. To do so, sort the N response vectors into T subcollections such that all the response vectors, $\mathbf{r}_t$, in subcollection t are identical. Let n_t be the number of responses in subcollection t, and let $\bar{\mathbf{c}}_t$ be the proportion-correct vector for subcollection t:

$$\bar{\mathbf{c}}_t = (\bar{c}_{1t}, \ldots, \bar{c}_{jt}, \ldots, \bar{c}_{kt}), \text{ where } \bar{c}_{jt} = \sum_{t=1}^{n_t} c_{jt}/n_t$$

Let $\bar{\mathbf{c}}$ be the proportion-correct vector across all responses,

$$\bar{\mathbf{c}} = (\bar{c}_1, \ldots, \bar{c}_j, \ldots, \bar{c}_k), \text{ where } \bar{c}_j = \frac{1}{N}\sum_{i=1}^{N} c_{ji}$$

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

Finally, let **u** be the unity vector, a row vector whose k elements are all one.

Then Murphy's partition of the Brier score is:

$$B = \bar{\mathbf{c}}(\mathbf{u} - \bar{\mathbf{c}})' + \frac{1}{N}\sum_{t=1}^{T} n_t(\mathbf{r}_t - \bar{\mathbf{c}}_t)\,(\mathbf{r}_t - \bar{\mathbf{c}}_t)' - \frac{1}{N}\sum_{t=1}^{T} n_t(\bar{\mathbf{c}}_t - \bar{\mathbf{c}})\,(\bar{\mathbf{c}}_t - \bar{\mathbf{c}})'$$

The first term is not a function of the probability assessments; rather, it reflects the relative frequency of true events across the k alternatives. For example, suppose all the items being assessed had the same two alternatives, {rain, no rain}. Then the first term of the partition is a function of the base rate of rain across the N items (or days). If it always (or never) rained, this term would be zero. Its maximum value, $(k - 1)/k$, would indicate maximum uncertainty about the occurrence of rain. The second term is a measure of calibration, the weighted average of the squared difference between the responses in a category and the proportion correct for that category. The third term, called resolution, reflects the assessor's ability to sort the events into subcategories for which the proportion correct is different from the overall proportion correct.

Murphy's partition was designed for repeated predictions of the same set of events (e.g., rain vs. no rain). When the alternatives have no common meaning across items (e.g., in a multiple-choice examination), then all that the first term indicates is the extent to which the correct answer appears equally often as the first, second, etc., alternative.

When only one response per item is scored, Murphy's partition (Murphy, 1972) reduces to:

$$B' = \bar{c}(1 - \bar{c}) + \frac{1}{N}\sum_{t=1}^{T} n_t(r_t - \bar{c}_t)^2 - \frac{1}{N}\sum_{t=1}^{T} n_t(\bar{c}_t - \bar{c})^2,$$

where $\bar{c}$ is the overall proportion correct and $\bar{c}_t$ is the proportion correct in subcategory t. When the scored responses are the responses that are greater than or equal to .5 (as with the two-alternative, half-range task), the first term reflects the subject's ability to pick the correct alternative and thus might be called knowledge. As before, the second term measures calibration, and the third resolution.

Similar measures of calibration have been proposed by Adams and Adams (1961) and by Oskamp (1962). None of these measures of calibration discriminates between overconfidence and underconfidence. The sampling properties of these measures are not known.

Meteorological research

In 1906, W. Ernest Cooke, government astronomer for Western Australia, advocated that each meteorological prediction be accompanied by a single number that would "indicate, approximately, the weight or degree of probability which the forecaster himself attaches to that particular prediction." (Cooke, 1906b, p. 274). He reported (Cooke, 1906a, 1906b) results from 1,951 predictions. Of those to which he had assigned the highest degree of probability ("almost certain to be verified"), .985 were correct. For his middle degree of probability ("normal probability"), .938 were correct, while for his lowest degree of probability ("doubtful"), .787 were correct.

In 1951, Williams asked eight professional weather bureau forecasters in Salt Lake City to assess the probability of precipitation for each of 1095 12-hour forecasts, using one of the numbers 0, .2, .4., .6, .8, and 1.0. Throughout most of the range, the proportion of precipitation days was lower than the probability assigned. This might reflect a fairly natural form of hedging in public pronouncements. People are much more likely to criticize a weather forecast that leaves them without an umbrella when it rains than one that leads them to carry an umbrella on dry days.

Similar results emerged from a study by Murphy and Winkler (1974). Their forecasters assessed the probability of precipitation for the next day twice, before and after seeing output from a computerized weather prediction system (PEATMOS). The 7,188 assessments (before and after PEATMOS) showed the same overestimation of the probability of rain found by Williams.

Sanders (1958) collected 12,635 predictions, using the 11 responses 0, .1., ...9, 1.0, for a variety of dichotomized events: wind direction, wind speed, gusts, temperatures, cloud amount, ceiling, visibility, precipitation occurrence, precipitation type, and thunderstorm. These data revealed only a slight tendency for the forecasters' probability assessments to exceed the proportion of weather events that occurred.[1] Root (1962) reported a symmetric pattern of calibration of 4,138 precipitation forecasts: Assessed probabilities were too low in the low range and too high in the high range, relative to the observed frequencies.

Winkler and Murphy (1968a) reported calibration curves for an entire year of precipitation forecasts from Hartford, Connecticut. Each forecast was for either a 6-hour or a 12-hour time period, with a lead time varying from 5 to 44 hours. Unfortunately, it was unclear whether the forecasters had included "a trace of precipitation" (less than .01 inch) in their predictions. The data were analyzed twice, once assuming that "precipitation" included the occurrence of traces and again without traces. The inclusion or exclusion of traces had a substantial effect on calibration, as did the time period. Six-hour forecasts with traces included and 12-hour forecasts excluding traces exhibited excellent calibration. The calibration curve for 12-hour forecasts with traces lay above the identity line; the curve for 6-hour forecasts excluding traces lay well below it. Variations in lead time did not affect calibration.

[1] The references by Cooke (1906), Williams (1951), and Sanders (1958) were brought to our attention by Raiffa (1969).

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

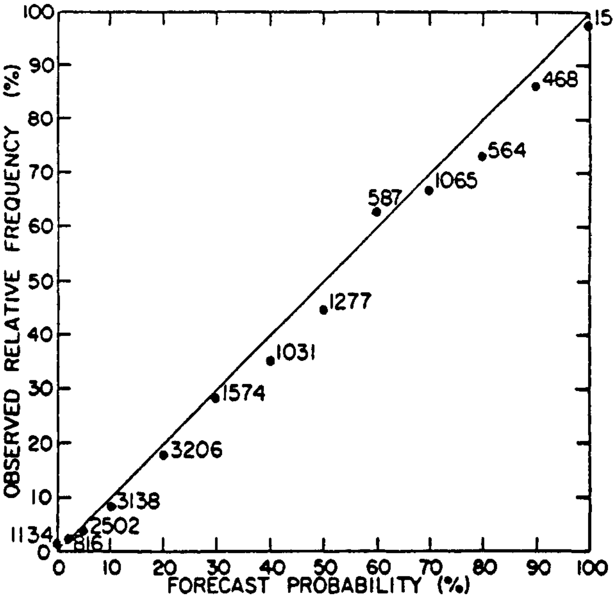

Figure 1. Calibration data for precipitation forecasts. The number of forecasts is shown for each point. (*Source:* Murphy & Winkler, 1977a.)

National Weather Service forecasters have been expressing their forecasts of precipitation occurrence in probabilistic terms since 1965. The calibration for some parts of this massive data base has been published (Murphy & Winkler, 1977a; U.S. Weather Bureau, 1969). Over the years the calibration has improved. Figure 1 shows the calibration for 24,859 precipitation forecasts made in Chicago during the four years ending June 1976. This shows remarkably good calibration; Murphy (1980) says the data for recent years are even better! He attributes this superior performance to the experience with probability assessment that the forecasters have gained over the years and to the fact that these data were gathered from real on-the-job performance.

Early laboratory research

In 1957, Adams reported the calibration of subjects who used an 11-point confidence scale: The subject was "instructed to express his confidence in terms of the percentage of responses, made at that particular level of confidence, that he expects to be correct. . . . Of those responses made with confidence *p*, about *p*% should be correct" (pp. 432–433).

In Adams's task, each of 40 words were presented tachistoscopically 10 times successively, with increasing illumination each time, to 10 subjects. After each exposure subjects wrote down the work they thought they saw and gave a confidence judgment. The resulting calibration curve showed

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

that the proportions that were correct greatly exceeded the confidence ratings along the entire response scale (except for the responses of 100). Great caution must be taken in interpreting these data: Because each word was shown 10 times, the responses are highly interdependent. It is unknown what effect such interdependence has on calibration. Subjects may have chosen to "hold back" on early presentations, unwilling to give a high response when they knew that the same word would be presented several more times.

The following year, Adams and Adams (1958) reported a training experiment, using the same response scale but a new, three-alternative, single-response task: For each of 156 pairs of words per session, subjects were asked whether the words were antonyms, synonyms, or unrelated. The mean calibration scores (based on the absolute difference, $|r_t - \bar{c}_t|$) of 14 experimental subjects, who were shown calibration tallies and calibration curves after each of five sessions, decreased by 48% from the first session to the last. Six control subjects, whose only feedback was a tally of their unscored responses, showed a 36% mean increase in discrepancy scores.

Adams and Adams (1961) discussed many aspects of calibration (using the term *realism of confidence*), anticipating much of the work done by others in recent years, and presented more bits of data, including the grossly overconfident calibration curve of a schizophrenic who believed he was Jesus Christ. In a nonsense-syllable learning task, they found large overconfidence on the first trial and improvement after 16 trials. They also briefly described a transfer of training experiment: On day 1, subjects made 108 decisions about the percentage of blue dots in an array of blue and red dots. On days 2 and 4, the subjects decided on the truth or falsity of 250 general knowledge statements. On day 3, they lifted weights, blindfolded. On day 5, they made 256 decisions (synonym, antonym, or unrelated) about pairs of words. Eight experimental subjects, given calibration feedback after each of the first four days, showed on the fifth day a mean absolute discrepancy score significantly lower than that of 8 control (no feedback) subjects, suggesting some transfer of training. Finally, Adams and Adams reported that across 56 subjects taking a multiple-choice final examination in elementary psychology, poorer calibration was associated with greater fear of failure ($r = .36$). Neither knowledge nor overconfidence was related to fear of failure.

Oskamp (1962) presented subjects with 200 MMPI profiles[2] as stimuli. Half the profiles were from men admitted to a Veterans Administration (VA) hospital for psychiatric reasons; the others were from men admitted for purely medical reasons. The subjects' task was to decide, for each profile, whether the patient's status was psychiatric or medical and to state

[2] The MMPI (Minnesota Multiphasic Personality Inventory) is a personality inventory widely used for psychiatric diagnosis. A profile is a graph of 13 subscores from the inventory.

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

the probability that their decision was correct. Each profile had been independently categorized as hard (61 profiles), medium (88), or easy (51) on the basis of an actuarially derived classification system, which correctly identified 57%, 69%, and 92% on the hard, medium, and easy profiles, respectively.

All 200 profiles were judged by three groups of subjects: 28 undergraduate psychology majors, 23 clinical psychology trainees working at a VA hospital, and 21 experienced clinical psychologists. The 28 inexperienced judges were later split into two matched groups and given the same 200 profiles again. Half were trained during this second round to improve accuracy; the rest were trained to improve calibration.

Oskamp used three measures of subjects' performance: accuracy (percentage correct), confidence (mean probability response), and appropriateness of confidence (a calibration score):

$$\frac{1}{N}\sum_t n_t\,|r_t - \bar{c}_t|$$

All three groups tended to be overconfident, especially the undergraduates in their first session (accuracy 70%, confidence .78). However, all three groups were underconfident on the easy profiles (accuracy 87%, confidence .83).

The subjects trained for accuracy increased their accuracy from 67% to 73%, approaching their confidence level, .78, which did not change as a result of training.[3] The subjects trained for calibration lowered their confidence from .78 to .74, bringing it closer to their accuracy, 68%, which remained unchanged. As would be expected from these changes, the calibration score of both groups improved.

Signal detection research

In the early days of signal detection research, investigators looked into the possibility of using confidence ratings rather than yes-no responses in order to reduce the amounts of data required to determine stable receiver operating characteristic (ROC) curves. Swets, Tanner, and Birdsall (1961) asked four observers to indicate their confidence that they had heard signal plus noise rather than noise alone for each of 1,200 trials. Although three of the four subjects were terribly calibrated, the four calibration curves were widely different. One subject exhibited a severe tendency to assign too small probabilities (e.g., the signal was present over 70% of the times when that subject used the response category ".05-.19").

Clarke (1960) presented one of five different words, mixed with noise, to listeners through headphones. The listeners selected the word they

[3] MMPI buffs might note that with this minimal training the undergraduates showed as high an accuracy as either the best experts or the best actuarial prediction systems.

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

thought they heard and then rated their confidence by indicating one of five categories defined by slicing the probability scale into five ranges. After each of 12 practice tests of 75 items, listeners scored their own results and noted the percentage of correct identifications in each rating category, thus allowing them to change strategies on the next test. Clarke found that although all five listeners appeared well calibrated when data were averaged over the five stimulus words, analyses for individual words showed that the listeners tended to be overconfident for low-intelligibility words and underconfident for words of relatively high intelligibility.

Pollack and Decker (1958) used a verbally defined 6-point confidence rating scale that ranged from "Positive I received the message correctly" to "Positive I received the message incorrectly." With this rating scale it is impossible to determine whether an individual is well calibrated, but it is possible to see shifts in calibration across conditions. Calibration curves for easy words generally lay above those for difficult words, whatever the signal-to-noise ratio, and the curves for high signal-to-noise ratios lay above those for low signal-to-noise ratios, whatever the word difficulty.

In most of these studies, calibration was of secondary interest; the important question was whether confidence ratings would yield the same ROC curves as Yes–No procedures. By 1966, Green and Swets concluded that, in general, rating scales and Yes–No procedures yield almost identical ROC curves. Since then, studies of calibration have disappeared from the signal detection literature.

Recent laboratory research

Overconfidence. The most pervasive finding in recent research is that people are overconfident with general-knowledge items of moderate or extreme difficulty. Some typical results showing overconfidence are presented in Figure 2. Hazard and Peterson (1973) asked 40 armed forces personnel studying at the Defense Intelligence School to respond with probabilities or with odds to 50 two-alternative general-knowledge items (e.g., "Which magazine had the largest circulation in 1970, *Playboy* or *Time*?"). Lichtenstein (unpublished) found similar results, using the same items but only the probability response, with 19 Oregon Research Institute employees, as did Phillips and Wright (1977) with different items, using British undergraduate students as subjects.

Numerous other studies using general-knowledge questions have shown the same overconfidence (Fischhoff, Slovic, & Lichtenstein, 1977; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977, 1980a, 1980b; Nickerson & McGoldrick, 1965). Cambridge and Shreckengost (1978) found overconfidence with Central Intelligence Agency analysts. Fischhoff and Slovic (1980) found severe overconfidence using a variety of impossible or nearly impossible tasks (e.g., predicting the

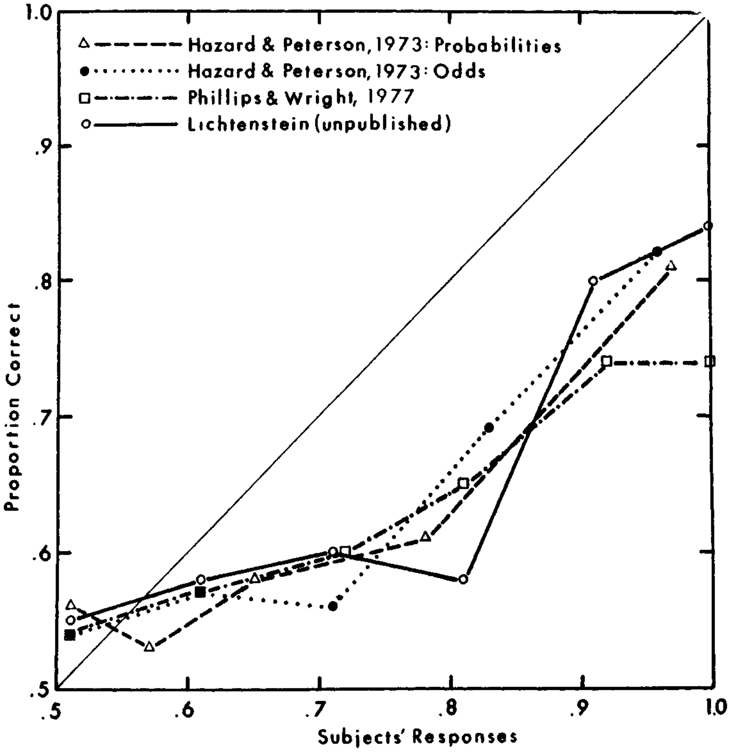

Figure 2. Calibration for half-range, general-knowledge items.

winners in 6-furlong horse races, diagnosing the malignancy of ulcers). Pitz (1974) reported overconfidence using a full-range method.

Fischhoff, Slovic, and Lichtenstein (1977) focused on the appropriateness of expressions of certainty. Using a variety of methods (no alternatives, one alternative, and two alternatives with half range and full range), they found that only 72% to 83% of the items to which responses of 1.0 were given were correct. In the full-range tasks, items assigned the other extreme response, zero, were correct 20% to 30% of the time. Using an odds response did not correct the overconfidence. Answers assigned odds of 1,000:1 of being correct were only 81% to 88% correct; for odds of 1,000,000:1 the correct alternative was chosen only 90% to 96% of the time. Subjects showed no reluctance to use extreme odds; in one of the experiments almost one-fourth of the responses were 1,000:1 or greater. Further analyses showed that extreme overconfidence was not confined to just a few subjects or a few items.

The effect of difficulty. Overconfidence is most extreme with tasks of great difficulty (Clarke, 1960; Nickerson & McGoldrick, 1965; Pitz, 1974). With

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

essentially impossible tasks (discriminating between European and American handwriting, Asian and European children's drawings, and rising and falling stock prices) calibration curves did not rise at all; for all assessed probabilities, the proportion of correct alternatives chosen was close to .5 (Lichtenstein & Fischhoff, 1977). Subjects were not reluctant to use high probabilities in these tasks; 70% to 80% of all responses were greater than .5.

As tasks get easier, overconfidence is reduced. Lichtenstein and Fischhoff (1977) allowed one group of subjects in the handwriting discrimination task to study a correctly labeled set of sample stimuli before making its probability assessments. This experience made the task much easier (71% correct versus 51% for the no-study group) and the study group was only slightly overconfident. Lichtenstein and Fischhoff (1977) performed post hoc analyses of the effect of difficulty on calibration using two large collections of data from general-knowledge, two-alternative half-range tasks. They separated easy items (those for which most subjects chose the correct alternative) from hard items and knowledgeable subjects (those who selected the most correct alternatives) from less knowledgeable subjects. They found a systematic decrease in overconfidence as the percentage correct increased. Indeed, the most knowledgeable subjects responding to the easiest items were *under*confident (e.g., 90% correct when responding with a probability of .80). This finding was replicated with two new groups of subjects given sets of items chosen to be hard or easy on the basis of previous subjects' performance. The resulting calibration curves are shown in Figure 3, along with the corresponding calibration curves from the post hoc analyses.

In the research just cited, difficulty was defined on the basis of subjects' performance (Clarke, 1960; Lichtenstein & Fischhoff, 1977). More recently, Lichtenstein and Fischhoff (1980a), following the lead of Oskamp (1962), developed a set of 500 two-alternative general-knowledge items for which difficulty could be defined independently. The items were of three types: Which of two cities, states, countries, or continents is more populous (e.g., Las Vegas vs. Miami), which of two cities is farther in distance from a third city (e.g., "Is Melbourne farther from Rome or from Tokyo?"), and which historical event happened first (e.g., Magna Carta signed vs. Mohammed born). Thus, each item had associated with it two numbers (populations, distances, or elapsed time to the present). The ratio of the larger to the smaller of those numbers was taken as a measure of difficulty: The 250 items with the largest ratios were designated as *easy*; the remaining, as *hard*. This a priori classification was quite successful; over 35 subjects, the percentage correct was 81 for easy items and 58 for hard items. These results, too, showed overconfidence for hard items and underconfidence for easy items.

The hard-easy effect seems to arise from assessors' inability to appre-

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

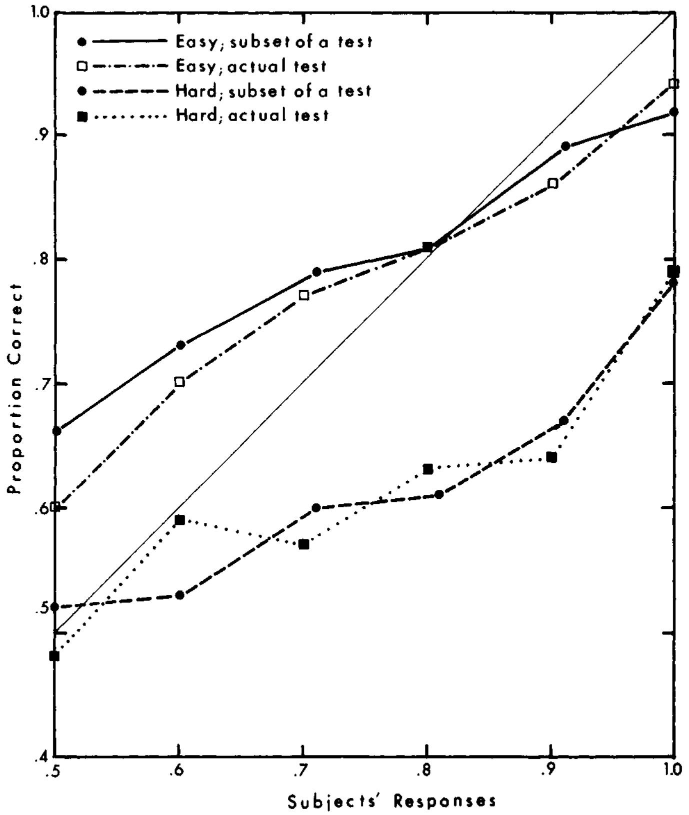

Figure 3. Calibration for hard and easy tests and for hard and easy subsets of a test.

ciate how difficult or easy a task is. Phillips and Chew (unpublished) found no correlation across subjects between percentage correct and the subjects' ratings on an 11-point scale of the difficulty of a set of just-completed items. However, subjects do give different distributions of responses for different tasks; Lichtenstein and Fischhoff (1977) reported a correlation of .91 between percentage correct and mean response across 16 different sets of data. But the differences in response distributions are less than they should be: Over those same 16 sets of data, the proportion correct varied from .43 to .92, while the mean response varied only from .65 to .86.

Ferrell and McGoey (1980) have recently developed a model for the calibration of discrete probability assessments that addresses the hard-easy effect. The model, based on signal detection theory, assumes that assessors transform their feelings of subjective uncertainty into a decision variable, X, which is partitioned into sections with cutoff values $\{x_i\}$. The

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

assessor reports probability r_i whenever X lies between x_{i-1} and x_i. Ferrell and McGoey assume that, in the absence of feedback about calibration performance, the assessor will not change the set of cutoff values, $\{x_i\}$, as task difficulty changes. This assumption leads to a prediction of overconfidence with hard items and underconfidence with easy items. Application of the model to much of the data from Lichtenstein and Fischhoff (1977) showed a moderately good fit to both the calibration curves and the distribution of responses under the assumption that the cutoff values remained constant as difficulty changed. Thus the hard–easy effect is seen as an inability to change the cutoffs involved in the transformation from feelings of certainty to probabilistic responses.

Effect of base rates. One-alternative (true–false) tasks may be characterized by the proportion of true statements in the set of items. To be well calibrated on a particular set of items one must take this base-rate information into account. The signal detection model of Ferrell and McGoey (1980) assumes that calibration is affected independently by (a) the proportion of true statements and (b) the assessor's ability to discriminate true from false statements. Assuming that the cutoff values, $\{x_i\}$, are held constant, the model predicts quite different effects on calibration from changing the proportion of true statements (while holding discriminability constant) as opposed to changing discriminability (while holding the proportion of true statements constant). Ferrell and McGoey presented data supporting their model. Students in three engineering courses assessed the probability that the answers they wrote for their examinations would be judged correct by the grader. Post hoc analyses separating the subjects into four groups (high vs. low percentage of correct answers and high vs. low discriminability) revealed the calibration differences predicted by the model. Unpublished data collected by Fischhoff and Lichtenstein, shown in Figure 4, also suggest support for the model. Four groups of subjects received 25 one-alternative general-knowledge items (e.g., "The Aeneid was written by Homer") differing in the proportion of true statements: .08, .20, .50, and .71. The groups showed dramatically different calibration curves, of roughly the same shape as predicted by Ferrell and McGoey for their base-rate changing, discriminability constant case.

Individual differences. Unqualified statements that one person is better calibrated than another person are difficult to make, for two reasons. First, at least several hundred responses are needed in order to get a stable measure of calibration. Second, it appears that calibration strongly depends on the task, particularly on the difficulty of the task. Indeed, Lichtenstein and Fischhoff (1980a) have suggested that each person may have an "ideal" test (i.e., a test whose difficulty level leads to neither

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

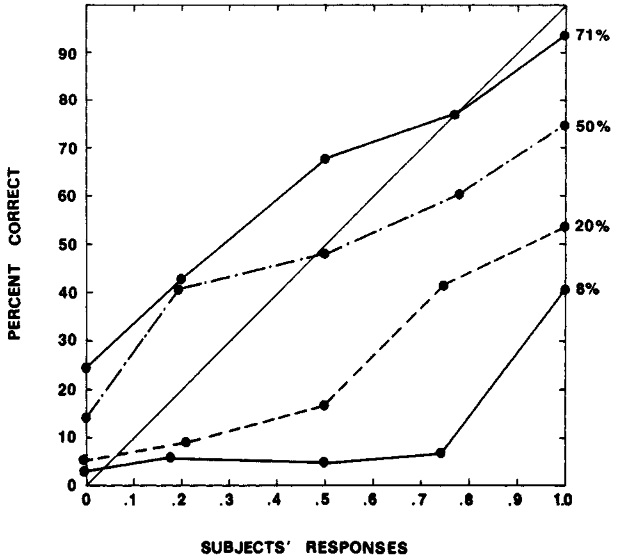

Figure 4. The effect on calibration due to changes in the percentage of true statements. (*Source:* Fischhoff & Lichtenstein, unpublished.)

overconfidence nor underconfidence, and thus the test on which the person will be best calibrated). However, the difficulty level of the "ideal" test may vary across people. Thus, even when one person is better than another on a particular set of items, the reverse may be true for a harder or easier set.

Comparisons between different groups of subjects have generally shown few differences when difficulty was controlled. Graduate students in psychology, who presumably are more intelligent than the usual subjects (undergraduates who answered an ad in the college newspaper), were no different in calibration (Lichtenstein & Fischhoff, 1977). Nor have we found differences in calibration or overconfidence between males and females (Lichtenstein & Fischhoff, 1981).

Wright and Phillips (1976) studied the relationships among several personality measures (authoritarianism, conservatism, dogmatism, and intolerance of ambiguity), verbal expressions of uncertainty (e.g., the number of words such as *unlikely* used in short written answers to 45 questions), and several measures of calibration. The only relationships they found between six personality scales and seven calibration measures were two modest correlations (.41 and .34) with the authoritarianism (*F*) scale. The calibration of certainty responses (i.e., responses of 1.0) was uncorrelated with the calibration of uncertainty (<1.0) responses. The measures of verbal uncertainty were uncorrelated with any of the numer-

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

ical calibration measures. The authors concluded that probabilistic thinking is neither a single factor nor strongly related to individual differences on personality measures.

Wright et al. (1978) have studied cross-cultural differences in calibration. The calibration of their British sample was shown in Figure 2 (identified there as Phillips & Wright, 1977). Their other samples were Hong Kong, Indonesian, and Malay students. The Asian groups showed essentially flat calibration curves. The authors speculated that fate-oriented Asian philosophies might account for these differences.

Corrective efforts. Fischhoff and Slovic (1980) tried to ward off overconfidence on the task of discriminating Asian from European children's drawings by using explicitly discouraging instructions:

> All drawings were taken from the Child Art Collection of Dr. Rhoda Kellogg, a leading proponent of the theory that children from different countries and cultures make very similar drawings. . . . Remember, it may well be impossible to make this sort of discrimination. Try to do the best you can. But if, in the extreme, you feel totally uncertain about the origin of all of these drawings, do not hesitate to respond with .5 for every one of them. (p. 792)

These instructions lowered the mean response by about .05, but substantial overconfidence was still found.

Will increased motivation improve calibration? Sieber (1974) compared the calibration of two groups of students on a course-related set of four-alternative items. One group was told that they were taking their mid-term examination. The other group was told that the test was not the mid-term but would be used to coach them for the mid-term. The two groups did not differ in the number of correct alternatives chosen, but the presumably more motivated group, whose performance would determine their grade, showed significantly *worse* calibration (greater overconfidence).

Training assessors by giving them feedback about their calibration has shown mixed results. As mentioned, Adams and Adams (1958) found modest improvement in calibration after five training sessions and, in a later study (1961), some generalization of training. Choo (1976), using only one training session with 75 two-alternative general-knowledge items, found little improvement and no generalization.

Lichtenstein and Fischhoff (1980b) trained two groups of subjects by giving extensive, personalized calibration feedback after each of either 2 or 10 sessions composed of 200 two-alternative general-knowledge items. They found appreciable improvement in calibration, all of which occurred between the first and the second session. Modest generalization occurred for tasks with different difficulty levels, content, and response mode (four rather than two alternatives), but no improvement was found with a

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

fractile assessment task (described in the next section) or on the discrimination of European from American handwriting samples.

Another approach to improving calibration is to restructure the task in a way that discourages overconfidence. In a study by Koriat, Lichtenstein, and Fischhoff (1980), subjects first responded to 30 two-alternative general-knowledge items in the usual way. They then received 10 additional items. For each item they wrote down all the reasons they could think of that supported or contradicted either of the two possible answers, and then made the usual choice and probability assessments. This procedure significantly improved their calibration. An additional study helped to pinpoint the effective ingredient of this technique. After responding as usual to an initial set of 30 items, subjects were given 30 more items. For each, they first chose a preferred answer, then wrote (a) one reason supporting their chosen answer, (b) one reason contradicting their chosen answer, or (c) two reasons, one supporting and one contradicting. Then they assessed the probability that their chosen answer was correct. Only the group asked to write contradicting reasons showed improved calibration. This result, as well as correlational analyses on the data from the first study, suggests that an effective partial remedy for overconfidence is to search for reasons why one might be *wrong.*

Expertise. Students taking a college course are, presumably, experts, at least temporarily, in the topic material of the course. Sieber (1974) reported excellent calibration for students taking a practice mid-term examination (i.e., the group of students who were told that the test was *not* their mid-term). Over 98% of their 1.0 responses and only .5% of their 0 responses were correct. Pitz (1974) asked his students to predict their grade for his course; they also were well calibrated.

Would these subjects have been as well calibrated on items of equivalent difficulty that were not in their area of expertise? Lichtenstein and Fischhoff (1977) asked graduate students in psychology to respond to 50 two-alternative general-knowledge items and 50 items covering knowledge of psychology (e.g., "the Ishihara test is (a) a perceptual test, (b) a social anxiety test"). The two subtests were of equal difficulty, and the calibration was similar for the two tasks.

Christensen-Szalanski and Bushyhead (1981) reported nine physicians' assessments of the probability of pneumonia for 1,531 patients who were examined because of a cough. Their calibration was abysmal; the curve rose so slowly that for the highest confidence level (approximately .88), the proportion of patients actually having pneumonia was less than .20. Similar results have been reported for diagnoses of skull fracture and pneumonia by Lusted (1977) and for diagnoses of skull fracture by DeSmet, Fryback, and Thornbury (1979). The results of these field studies with physicians are in marked contrast with the superb calibration of

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

weather forecasters' precipitation predictions. We suspect that several factors favor the weather forecasters. First, they have been making probabilistic forecasts for years. Second, the task is repetitive; the question to be answered (Will it rain?) is always the same. In contrast, a practicing physician is hour by hour considering a wide array of possibilities (Is it a skull fracture? Does she have strep? Does he need further hospitalization?). Finally, the outcome feedback for weather forecasters is well defined and promptly received. This is not always true for physicians; patients fail to return or are referred elsewhere, or diagnoses remain uncertain.

People who bet on or establish the odds for horse races might also be considered experts. Under the pari-mutuel (or totalizator) method, the final odds are determined by the amount of money bet on each horse, allowing a kind of group calibration curve to be computed. Such curves (Fabricand, 1965; Hoerl & Fallin, 1974) show excellent calibration, with only a slight tendency for people to bet too heavily on the long shots. However, such data are only inferentially related to probability assessment. More relevant are the calibration results reported by Dowie (1976), who studied the forecast prices printed daily by a sporting newspaper in Britain. These predictions, in the form of odds, are made by one person for all the horses in a given race; about eight people made the forecasts during the year studied. The calibration of the forecasts for 29,307 horses showed a modest underconfidence for probabilities greater than .4 and superb calibration for probabilities less than .4 (which comprised 98% of the data).

The burgeoning research on calibration has led to the development of a new kind of expertise: calibration experts, who know about the common errors people make in assessing probabilities. Lichtenstein and Fischhoff (1980a) compared the calibration of 8 such experts with 12 naive subjects and 15 subjects who had previously been trained to be well calibrated. The normative experts not only overcame the overconfidence typically shown by naive subjects but apparently overcompensated, for they were underconfident. The experts were also slightly more sensitive to item difficulty than the other two groups.

Future events. Wright and Wishudha (1979) have speculated that calibration for future events may be different from that for general-knowledge questions. If true, this would limit extrapolation from research with general-knowledge questions to the prediction of future events. Unfortunately, Wright and Wishudha's general-knowledge items were more difficult than their future events, which could account for the superior calibration of the latter.

Fischhoff and Beyth (1975) asked 150 Israeli students to assess the probability of 15 then-future events, possible outcomes of President

Nixon's much-publicized trips to China and Russia (e.g., "President Nixon will meet Mao at least once"). The resulting calibration curve was quite close to the identity line. However, Fischhoff and Lichtenstein (unpublished) have recently found that the calibration of future events showed the same severe overconfidence as was shown for general-knowledge items of comparable difficulty. Phillips and Chew (unpublished) obtained calibration curves for three sets of items: general knowledge, future events, and past events (e.g., "a jumbo jet crashed killing more than 100 people sometime in the past 30 days"). All three curves showed overconfidence. Calibration for future and past events was identical, and somewhat better than for the general-knowledge items. The difficulty levels of the three sets of items could not account for these results.

Jack Dowie and colleagues are now collecting calibration data at the Open University in Milton Keynes, England, from several hundred students in the course on risk, using course-related questions, general-knowledge questions, and future-event questions. The students received a general introduction to the concept of calibration and were given feedback about their performance and calibration. Preliminary results (Dowie, 1980) suggest that they were moderately overconfident. Calibration was best on general-knowledge items and worst on course-related items, but the significance and origins of these differences remain to be investigated.

Continuous propositions: Uncertain quantities

The fractile method

Uncertainty about the value of an uncertain continuous quantity (e.g., What proportion of students prefer Scotch to bourbon? What is the shortest distance from England to Australia?) may be expressed as a probability density function across the possible values of that quantity. However, assessors are not usually asked to draw the entire function. Instead, the elicitation procedure most commonly used is some variation of the fractile method. In this method, the assessor states values of the uncertain quantity that are associated with a small number of predetermined fractiles of the distribution. For the median or .50 fractile, for example, the assessor states a value of the quantity such that the true value is equally likely to fall above or below the stated value; the .01 fractile is a value such that there is only 1 chance in 100 that the true value is smaller than the stated value. Usually three or five fractiles, including the median, are assessed. In a variant called the *tertile* method, the assessor states two values (the .33 and .67 fractiles) such that the entire range is divided into three equally likely sections.

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

Two calibration measures are commonly reported. The *interquartile* index is the percentage of items for which the true value falls inside the interquartile range (i.e., between the .25 and the .75 fractiles). The perfectly calibrated person will, in the long run, have an interquartile index of 50. The *surprise index* is the percentage of true values that fall outside the most extreme fractiles assessed. When the most extreme fractiles assessed are .01 and .99, the perfectly calibrated person will have a surprise index of 2. A large surprise index shows that the assessor's confidence bounds have been too narrow to encompass enough of the true values and thus indicates overconfidence (or hyperprecision; Pitz, 1974). Underconfidence would be indicated by an interquartile index greater than 50 and a low surprise index; no such data have been reported in the literature.

The impetus for investigating the calibration of probability density functions came from a 1969 paper by Alpert and Raiffa (1969, **21**). Alpert and Raiffa worked with Harvard Business School students, all familiar with decision analysis. In group 1, all subjects assessed five fractiles, three of which were .25, .50, and .75. The extreme fractiles were, however, different for four subgroups; .01 and .99 (group A); .001 and .999 (group B); "the minimum possible value" and "the maximum possible value" (group C); and "astonishingly low" and "astonishingly high" (group D). The interquartile and surprise indices for these four subgroups are shown in Table 1. Discouraged by the enormous number of surprises, Alpert and Raiffa then ran three additional groups (2, 3, and 4) who, after assessing 10 uncertain quantities, received feedback in the form of an extended report and explanation of the results, along with perorations to "Spread Those Extreme Fractiles!" The subjects then responded to 10 new uncertain quantities. Results before and after feedback are shown in Table 1. The subjects improved, but still showed considerable overconfidence.

Hession and McCarthy (1974) collected data comparable to Alpert and Raiffa's first experiment, using 55 uncertain quantities and 36 graduate students as subjects. Their instructions urged subjects to make certain that the interval between the .25 fractile and the .75 fractile did indeed capture half of the probability. "Later discussion with individual subjects made it clear that this consistency check resulted in most cases in a readjustment, decreasing the interquartile range originally assessed" (p. 7) – thus making matters worse! This instructional emphasis, not used by Alpert and Raiffa, may explain why Hession and McCarthy's subjects were so badly calibrated, as shown in Table 1.

Hession and McCarthy also gave their subjects a number of individual difference measures: authoritarianism, dogmatism, rigidity, Pettigrew's Category-width Scale, and intelligence. The correlations of the subjects' test scores with their interquartile and surprise indices were mostly quite low, although the authoritarian scale correlated –.31 with the interquar-

Table 1. *Calibration summary for continuous items: Percentage of true values falling within interquartile range and outside the extreme fractiles*

		Observed interquartile index[a]	Surprise index	
	N		Observed	Ideal
Alpert & Raiffa (1969)				
Group 1-A (.01, .99)	880	33	46	2
Group 1-B (.001, .999)	500		40	.2
Group 1-C ("min" & "max")	700		47	?
Group 1-D ("astonishingly high/low")	700		38	?
Groups 2, 3, & 4				
before training	2,270	34	34	2
after training	2,270	44	19	2
Hession & McCarthy (1974)	2,035	25	47	2
Selvidge (1975)				
Five fractiles	400	56	10	2
Seven fractiles (incl. .1 & .9)	520	50	7	2
Moskowitz & Bullers (1978)				
Proportions				
Three fractiles	120	—	27	2
Five fractiles	145	32	42	2
Dow-Jones				
Three fractiles	210	—	38	2
Five fractiles	210	20	64	2
Pickhardt & Wallace (1974)				
Group 1,				
first round	?	39	32	2
fifth round	?	49	20	2
Group 2,				
first round	?	30	46	2
sixth round	?	45	24	2
T. A. Brown (1973)	414	29	42	2
Lichtenstein & Fischhoff (1980b)				
Pretest	924	32	41	2
Post-test	924	37	40	2
Seaver, von Winterfeldt, & Edwards (1978)				
Fractiles	160	42	34	2
Odds-fractiles	160	53	24	2
Probabilities	180	57	5	2
Odds	180	47	5	2
Log odds	140	31	20	2
Schaefer & Borcherding (1973)				
First day, fractiles	396	23	39	2
Fourth day, fractiles	396	38	12	2
First day, hypothetical sample	396	16	50	2
Fourth day, hypothetical sample	396	48	6	2

Table 1 *(cont.)*

	N	Observed interquartile index[a]	Surprise index	
			Observed	Ideal
Larson & Reenan (1979) "Reasonably Certain"	450	—	42	?
Pratt (1975) "Astonishingly high/low"	175	37	5	?
Murphy & Winkler (1974) Extremes were .125 & .875	132	45	27	25
Murphy & Winkler (1977b) Extremes were .125 & .875	432	54	21	25
Staël von Holstein (1971a)	1,269	27	30	2

Note: N = total number of assessed distributions.
[a] The ideal percentage of events falling within the interquartile range is 50, for all experiments except Brown (1973). He elicited the .30 and .70 fractiles, so the ideal is 40%.

tile score and +.47 with the surprise score ($N = 28$). This is consistent with Wright and Phillips's (1976) finding that authoritarianism was modestly related to calibration.

Selvidge (1975) extended Alpert and Raiffa's work by first asking subjects four questions about themselves (e.g., "Do you prefer Scotch or bourbon?"). Their responses determined the true answer for these *group-generated* proportions (e.g., what proportion of the subjects answering the questionnaire preferred Scotch to bourbon?). One group gave five fractiles, .01, .25, .5, .75, and .99. Another group gave those five plus two others: .1 and .9. As shown in Table 1, the seven-fractile group did a bit better. The five-fractile results are not as different from Alpert and Raiffa's results as they appear. Three of Alpert and Raiffa's uncertain quantities were group-generated proportions similar to Selvidge's items. On these three items, Alpert and Raiffa found 57% in the interquartile range and 20% surprises. Finally, for one of the items, half the subjects in the five-fractile group were asked to give .25, .5, and .75 first, and then to give .01 and .99, while the other half were instructed to assess the extremes first. Selvidge found fewer surprises for the former order (8%) than for the latter (16%).

Moskowitz and Bullers (1978) also used group-generated proportions, but found many more surprises than did Selvidge. One group gave the same five fractiles that Selvidge used (in the order .5, .25, .75, .01, .99). Another group was asked for only three assessments (the mode of the

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

distribution and the .01 and .99 fractiles). Before making their assessments, the three-fractile group received a presentation and discussion of some typical reference events (e.g., "Consider a lottery in which 100 people are participating. Your chance of holding the winning ticket is 1 in 100") designed to give assessors a better understanding of the meaning of a .01 probability. As shown in Table 1, the three-fractile group had fewer surprises than the five-fractile group. In another experiment using the same two methods, Moskowitz and Bullers asked 44 undergraduate commerce students to assess the average value of the Dow-Jones industrial index for 1977, 1974, 1965, 1960, and 1950. Each subject gave assessments before and after engaging in three-person discussions. Since no systematic differences due to the discussions were found, the data have been combined in Table 1. Again, the three-fractile group (who had received the presentation on the meaning of .01) had fewer surprises than the five-fractile group. The performance of the five-fractile group was extremely bad.

Pickhardt and Wallace (1974) replicated Alpert and Raiffa's work with variations. Across several groups they reported 38% to 48% surprises before feedback and not less than 30% surprises after feedback. Two variations, using or not using course grade credit as a reward for good calibration and using or not using scoring rule feedback, made no difference in the number of surprises. Pickhardt and Wallace also studied the effects of extended training: Two groups of 18 and 30 subjects (number of uncertain quantities not reported) responded for five and six sessions with calibration feedback after every session. Modest improvement was found, as shown in Table 1.

Finally, Pickhardt and Wallace (1974) studied the effects of increasing knowledge on calibration in the context of a production simulation game called PROSIM. Thirty-two graduate students each made 51 assessments during a simulated 17 "days" of production scheduling. Each assessment concerned an event that would occur 1, 2, or 3 "days" hence. The closer the time of assessment to the time of the event, the more the subject knew about the event. Overconfidence decreased with this increased information: There were 32% surprises with 3-day lags, 24% with 2-day lags, and 7% with 1-day lags. No improvement was observed over the 17 "days" of the stimulation.

T.A. Brown (1973) asked 31 subjects to assess seven fractiles (.01, .10, .30, .50, .70, .90, .99) for 14 uncertain quantities. The results, shown in Table 1, are particularly discourging, because each question was accompanied by extensive historical data (e.g., for "Where will the Consumer Price Index stand in December, 1970?" subjects were given the consumer price index for every quarter between March 1962 and June 1970). For 11 of the questions, had the subjects given the historical minimum as their .01 fractile and the historical maximum as their .99 fractile, they would have

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

had no surprises at all. The other 3 questions showed strictly increasing or strictly decreasing histories, and the true value was close to any simple approximation of the historical trend. The subjects must have been relying heavily on their own erroneous knowledge to have given distributions so tight as to produce 42% surprises.

Lichtenstein and Fischhoff (1980b) elicited five fractiles (.01, .25, .5, .75, .99) from 12 subjects on 77 uncertain quantities both before and after the subjects received extensive calibration training on two-alternative discrete items. As shown in Table 1, the subjects did not significantly improve their calibration of uncertain quantities.

Other methods

Seaver, von Winterfeldt, and Edwards (1978) studied the effects of five different response modes on calibration. Two groups used the fractile method, either five fractiles (.01, .25, .50, .75, .99) or the odds equivalents of those fractiles (1:99, 1:3, 1:1, 3:1, 99:1). Three other groups responded with probabilities, odds, or odds on a log-odds scale to one-alternative questions that specified a particular value of the uncertain quantity (e.g., "What is the probability that the population of Canada in 1973 exceeded 25 million?"). Five such fixed values were given for each uncertain quantity, and from the responses the experimenters estimated the interquartile and surprise indices. For each method, seven to nine students responded to 20 uncertain quantities. As shown in Table 1, the groups giving probabilistic and odds responses had distinctly better surprise indices than those using the fractile method. It is unclear whether this superiority is due to the information communicated by the values chosen by the experimenter. The log-odds response mode did not work out well.

Schaefer and Borcherding (1973) asked 22 students to assess 18 group-generated proportions in each of four sessions. Each subject used two assessment techniques: (a) the fractile method (.01, .125, .25, .5, .75, .875, .99), and (b) the hypothetical sample method. In the latter method, the assessor states the size, *n*, and the number of successes, *r*, of a hypothetical sample that best reflects the assessor's knowledge about the uncertain quantity (i.e., I feel as certain about the true value of the proportion as I would feel were I to observe a sample of *n* cases with *r* successes). Larger values of *n* reflect greater certainty about the true value of the proportion. The ratio r/n reflects the mean of the probability density function. Subjects had great difficulty with this method, despite instructions that included examples of the beta distributions underlying this method. After every session, subjects were given extensive feedback, with emphasis on their own and the group's calibration. The results from the first and last sessions are shown in Table 1. Improvement was found for both methods. Results from the hypothetical sample method started out worse (50%

surprises and only 16% in the interquartile range) but ended up better (6% surprises and 48% in the interquartile range) than the fractile method.

Barclay and Peterson (1973) compared the tertile method (i.e., the fractiles .33 and .67) with a "point" method in which the assessor is asked to give the modal value of the uncertain quantity, and then two values, one above and one below the mode, each of which are half as likely to occur as is the modal value (i.e., points for which the probability density function is half as high as at the mode). Using 10 almanac questions as uncertain quantities and 70 students at the Defense Intelligence School in a within-subjects design, they found for the tertile method that 29% (rather than 33%) of the true answers fell in the central interval. For the point method, only 39% fell between the two half-probable points, whereas, for most distributions, approximately 75% of the density falls between these points.

Pitz (1974) reported several results using the tertile method. For 19 subjects estimating the populations of 23 countries, he found only 16% of the true values falling inside the central third of the distributions. In another experiment he varied the items according to the depth and richness of knowledge he presumed his subjects to have. With populations of countries (low knowledge) he found 23% of the true values in the central third; with heights of well-known buildings (middling knowledge), 27%; and with ages of famous people (high knowledge), 47%, the last being well above the expected 33%. In another study, he asked 6 subjects to assess tertiles and a few days later to choose among bets based on their own tertile values. He found a strong preference for bets involving the central region, just the reverse of what their too-tight intervals should lead them to.

Larson and Reenan's (1979) subjects first gave their best guess at the true answer (i.e., the mode) and then two more values that defined an interval within which they were "reasonably certain" the correct answer lay. Forty-two percent of the true values lay outside this region. Note how similar this surprise index is to the indices of Alpert and Raiffa's subjects given the verbal phrases "minimum/maximum" (47%) and "astonishingly high/low" (38%).

Real tasks with experts

Pratt (1975) asked a single expert to predict movie attendance for 175 movies or double features shown in two local theaters over a period of more than one year. The expert assessed the median, quartiles, and "astonishingly high" and "astonishingly low" values. As shown in Table 1, the interquartile range tended to be too small. Even though the expert received outcome feedback throughout the experiment, the only evidence of improvement in calibration over time came in the first few days.

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

Three experiments used weather forecasters for subjects. In two experiments, Murphy and Winkler (1974, 1977b) asked weather forecasters to give five fractiles (.125, .25, .5, .75, .875) for tomorrow's high temperature. The results, shown in Table 1, indicate excellent calibration. These subjects had fewer surprises in the extreme 25% of the distribution than did most of Alpert and Raiffa's subjects in the extreme 2%! Murphy and Winkler found that the five subjects in the two experiments who used the fractile method were better calibrated than four other subjects who used a fixed-width method. For the fixed-width method, the forecasters first assessed the median temperature (i.e., the high temperature for which they believed there was a .5 probability that it would be exceeded). Then they stated the probability that the temperature would fall with intervals of 5°F and of 9°F centered at the median. These forecasters were overconfident; the probability associated with the temperature falling inside the interval tended to be too large. The superiority of the fractile method over the fixed-width method stands in contrast to Seaver, von Winterfeldt, and Edwards's finding that fixed-value methods were superior, perhaps because the fixed intervals used by Murphy and Winkler (5°F and 9°F) were noninformative.

Staël von Holstein (1971a) used three fixed-value tasks: (a) average temperature tomorrow and the next day (dividing the entire response range into eight categories), (b) average temperature 4 and 5 days from now (eight categories), and (c) total amount of rain in the next 5 days (four categories). From each set of responses (four or eight probabilities summing to 1.0) he estimated the underlying cumulative density function. He then combined the 1,269 functions given by 28 participants. From the group cumulative density function shown in his article, we have estimated the surprise and interquartile indices (see Table 1). In contrast to other weather forecasters, these subjects were quite poorly calibrated, perhaps because the tasks were less familiar.

Summary of calibration with uncertain quantities

The overwhelming evidence from research using fractiles to assess uncertain quantites is that people's probability distributions tend to be too tight. The assessment of extreme fractiles is particularly prone to bias. Training improves calibration somewhat. Experts sometimes perform well (Murphy & Winkler, 1974, 1977b), sometimes not (Pratt, 1975; Staël von Holstein, 1971a). There is some evidence that difficulty is related to calibration for continuous propositions. Pitz (1974) and Larson and Reenan (1979) found such an effect, and Pickhardt and Wallace's (1974) finding that 1-day lags led to fewer surprises than 3-day lags in their simulation game is relevant here. Several studies (e.g., Barclay & Peterson, 1973; Murphy & Winkler, 1974) have reported a correlation between the spread of the assessed distribution and the absolute difference between the assessed median and the true answer, indicating that subjects do have a partial sensitivity to how much they do or don't know. This finding parallels the correlation between the percentage correct and the mean response with discrete propositions.

Discussion

Why be well calibrated?

Why should a probability assessor worry about being well calibrated? Von Winterfeldt and Edwards (1973) have shown that in most real-world decision problems with continuous decision options (e.g., invest X dollars) fairly large assessment errors make relatively little difference in the expected gain. However, several considerations argue against this reassuring view. First, in a two-alternative situation, the payoff function can be quite steep in the crucial region. Suppose your doctor must decide the probability that you have condition A, and should receive treatment *A*, versus having condition B and receiving treatment *B*. Suppose that the utilities are such that treatment *A* is better if the probability that you have condition A is greater than or equal to .4; otherwise treatment *B* is better. If the doctor assesses the probability that you have A as $p(\text{A}) = .45$ but is poorly calibrated, so that the appropriate probability is .25, then the doctor would use treatment *A* rather than treatment *B* and you would lose quite a chunk of expected utility. Real-life utility functions of just this type are shown by Fryback (1974).

Furthermore, when the payoffs are very large, when the errors are very large, or when such errors compound, the expected loss looms large. For instance, in the Reactor Safety Study (U.S. Nuclear Regulatory Commission, 1975) "at each level of the analysis a log-normal distribution of failure rate data was assumed with 5 and 95 percentile limits defined" (Weatherwax, 1975, p. 31). The research reviewed here suggests that distributions built from assessments of the .05 and .95 fractiles may be grossly biased. If such assessments are made at several levels of an analysis, with each assessed distribution being too narrow, the errors will not cancel each other but will compound. And because the costs of nuclear-power-plant failure are large, the expected loss from such errors could be enormous.

If good calibration is important, how can it be achieved? Cox (1958) recommended that one externally recalibrate people's assessments by fitting a model to a set of assessments for items with known answers. From then on, the model is used to correct or adjust responses given by the assessor. The technical difficulties confronting external recalibration are substantial. When eliciting the assessments to be modeled, one would have to be careful not to give the assessors any more feedback than they

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

normally receive, for fear of their changing their calibration as it is being measured. As Savage (1971) pointed out, "You might discover with experience that your expert is optimistic or pessimistic in some respect and therefore temper his judgments. Should he suspect you of this, however, you and he may well be on the escalator to perdition" (p. 796). Furthermore, since research has shown that the type of miscalibration observed depends on a task's difficulty level, one would also have to believe that the future will match the difficulty of the events used for the recalibration.

The theoretical objections to external recalibration may be even more serious than the practical objections. The numbers produced by a recalibration process will not, in general, follow the axioms of probability theory (e.g., the numbers associated with mutually exclusive and exhaustive events will not always sum to one, nor will it be generally true that $P(A) \cdot P(B) = P(A,B)$ for independent events); hence, these new numbers cannot be called probabilities.

A more fruitful approach would be to train assessors to become well calibrated. Under what conditions might one expect that assessors could achieve this goal? One should not expect assessors to be well calibrated when the explicit or implicit rewards for their assessments do not motivate them to be honest in their assessments. As an extreme example, an assessor who is threatened with beheading should any event occur whose probability was assessed at $<.25$ will have good reason not to be well calibrated with assessments of .20. Although this example seems absurd, more subtle pressures such as "avoid being made to look the fool" or "impress your boss" might also provide strong incentives for bad calibration. Any rewards for either wishful thinking or denial could also bias the assessments.

Receiving outcome feedback after every assessment is the best condition for successful training. Dawid (in press) has shown that under such conditions assessors who are honest and coherent subjectivists will expect to be well calibrated regardless of the interdependence among the items being assessed. In contrast, Kadane (1980) has shown that in the absence of trial-by-trial outcome feedback, honest, coherent subjectivists will expect to be well calibrated if and only if all the items being assessed are independent. This theorem puts strong restrictions on the situations under which it would be reasonable to expect assessors to learn to be well calibrated. Even if the training process could be conducted using only events that assessors believed were independent, there may be good reason to doubt the independence of the real-life tasks to which the assessors would apply their training. Important future events may be interdependent either because they are influenced by a common underlying cause or because the assessor evaluates all of them by drawing on a common store of knowledge. In such circumstances, one would not want or expect to be well calibrated.

The possibility that people's biases vary as a function of the difficulty of

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

the tasks poses a further obstacle to calibration training in the absence of immediate outcome feedback. The difficulty level of future tasks may be impossible to predict, thus rendering the training ineffective.

Calibration as cognitive psychology

Experiments on calibration can be used to learn how people think. Even if the immediate practical significance of each study is limited, it may still provide greater understanding of how people develop and express feelings of uncertainty and certainty. However, a striking aspect of much of the literature reviewed here is its "dust-bowl empiricism." Psychological theory is often absent, either as motivation for the research or as explanation of the results.

Not all authors have avoided theorizing. Slovic (1972a) and Tversky and Kahneman (1974, 1) argued that, as a result of limited information-processing abilities, people adopt simplifying rules or heuristics. Although generally quite useful, these heuristics can lead to severe and systematic errors. For example, the tendency of people to give unduly tight distributions when assessing uncertain quantities could reflect the heuristic called "anchoring and adjustment." When asked about an uncertain quantity, one naturally thinks first of a point estimate such as the median. This value then serves as an anchor. To give the 25th or 75th percentile, one adjusts downward or upward from the anchor. But the anchor has such a dominating influence that the adjustment is insufficient; hence the fractiles are too close together, yielding overconfidence.

Pitz (1974), too, accepted that people's information-processing capacity and working memory capacity are limited. He suggested that people tackle complex problems serially, working through a portion at a time. To reduce cognitive strain, people ignore the uncertainty in their solutions to the early portions of the problem in order to reduce the complexity of the calculations in later portions. This could lead to too-tight distributions and overconfidence. Pitz also suggested that one way people estimate their own uncertainty is by seeing how many different ways they can arrive at an answer, that is, how many different serial solutions they can construct. If many are found, people will recognize their own uncertainty; if few are found, they will not. The richer the knowledge base from which to build alternative structures, the less the tendency toward overconfidence.

Phillips and Wright (1977) presented a three-stage serial model. Their model distinguishes people who tend naturally to think about uncertainty in a probabilistic way from those who respond in a more black-and-white fashion. Their work on cultural and individual differences (Wright & Phillips, 1976, Wright et al., 1978) has attempted, with partial success, to identify distinct cognitive styles in processing this type of information.

Koriat et al. (1980) also took an information-processing approach. They discussed three stages for assessing probabilities. First one searches one's

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press

memory for relevant evidence. Next one assesses that evidence to arrive at a feeling of certainty or doubt. Finally, one translates the certainty feeling into a number. The manipulations used by Koriat et al. were designed to alter the first two stages, by forcing people to search for and attend to contradictory evidence, thereby lowering their confidence.

Ferrell and McGoey's (1980) model, on the other hand, deals entirely with the third stage, translation of feelings of certainty into numerical responses. By assuming that, without feedback, people are unable to alter their translation strategies as either the difficulty of the items or the base rate of the events changes, the model provides strong predictions that have received support from calibration data.

Structure and process theories of probability assessment are beginning to emerge; we hope that the further development of such theories will serve to integrate this rather specialized field into the broader field of cognitive psychology.

https://doi.org/10.1017/CBO9780511809477.023 Published online by Cambridge University Press