# A Subjectivist View of Calibration

Joseph B. Kadane

Carnegie-Mellon University

Sarah Lichtenstein

Decision Research, A Branch of Perceptronics, Eugene, Oregon

Decision Research Report 82-6, 1982

Correspondence may be addressed to:

Sarah Lichtenstein
Decision Research
A Branch of Perceptronics
1201 Oak Street
Eugene, Oregon 97401

Running Head: A Subjectivist View of Calibration

Abstract

Calibration concerns the relationship between subjective probabilities and the long-run frequencies of events. Theorems from the statistical and probability literature are reviewed to discover the conditions under which a coherent Bayesian expects to be calibrated. If the probability assessor knows the outcomes of all previous events when making each assessment, calibration is always expected. However, when such outcome feedback is lacking, the assessor expects to be well calibrated on an exchangeable set of events if and only if all the events in question are viewed as independent. Although this strong condition has not been tested in previous research, we speculate that the past findings of pervasive overconfidence are not invalid. Although experimental studies of calibration hold promise for the development of cognitive theories of confidence, their value for the practice of probability assessment seems more limited. Efforts to train probability assessors to be calibrated may be misplaced.

# A Subjectivist View of Calibration

The need for subjectively assessed probabilities has become widely recognized in the last ten years. Such probabilities are routinely used by weather forecasters (Murphy, in press) and are gaining adherents in medicine (Lusted, 1968), business (Brown, Kahr & Peterson, 1974), intelligence analysis (Cambridge & Shreckengost, Note 1), and technological risk assessment (even the "Rasmussen Report", USNRC, 1975, used subjectively assessed probabilities to quantify failure rates).

Accompanying this interest has been a burgeoning experimental literature exploring the validity of such assessments. This validity has usually been sought in a characteristic called calibration (also sometimes called reliability, Murphy, 1973). For probabilities assigned to the outcomes of discrete events (e.g., will it rain tomorrow?), probability assessments are calibrated, or well calibrated, if, in the long run, the proportion of true events is equal to the probability assigned to the events. Thus just 70% of all events to which one assigned a probability of .7 should be true. For probability density functions assessed over the range of an uncertain continuous quantity (e.g., how many inches of rain will fall tomorrow?), the assessments are well calibrated if, in the long run, the proportion of true values that fall at or below the n'th fractile of the assessed probability density functions is equal to n. Thus, for example, just 50% of the true values should fall at or below the .50 fractile, or median, of the assessed distributions.

For a review of the research literature on calibration see Lichtenstein, Fischhoff, and Phillips (1982).

The research on calibration seems to have been motivated both by a concern for its practical implications and by an interest in the cognitive processes involved in thinking under uncertainty. Practically speaking, people want other people to be well calibrated. If you are a surgeon recommending that your patient undergo an operation, the patient can more easily evaluate your advice when your statement, "There is only X% chance of serious complications" can be interpreted as a statement consistent with your previous surgical outcomes. A drug company can more effectively plan for the future when the predictions its staff makes about possible new drugs are well-calibrated probabilities (Balthasar, Boschi, & Menke, 1978).

For the cognitive psychologist, measures of calibration provide a tool for investigating the conditions that affect our feelings of uncertainty and the processes underlying these feelings. The research has been directed towards such questions as: Are people overconfident? When? Why?

The subjectivist, or Bayesian, view of probability (de Finetti, 1974; Savage, 1954) rejects the idea that relative frequencies provide the definitional foundation for probabilities. But subjectivists have seemed to regard calibration, which is frequency-based, as somehow a good thing. Raiffa, for example, has written:

> As consumers, we should like probabilistic reports to be <u>externally validated</u> by empirical frequencies. We should want our experts to calibrate themselves in such a way that if we were to group together a large number of forecasts in the

> .8 probability category (say), then roughly .8 of the forecasts should turn out to be correct (Note 2, p. 5).

Roberts made a similarly intuitive appeal:

> Even among people who are enthusiastic about expressing numerical probability assessments, there is a feeling that these assessments will have been in vain unless they are borne out by subsequent frequencies (Note 3, p. 5).

More recently, subjectivists have been no more specific about the theoretical foundations of calibration:

> To most people, it seems reasonable to say that events which are assigned a subjective probability of 30% should, on the average, occur 30% of the time (Harrison, 1977, p. 322).

The present paper attempts to fill this gap in subjectivist theory by exploring the conditions under which a person (called you) who subscribes to the subjectivist theory of probability would expect to be well calibrated and under what conditions you would expect not to be well calibrated.

We begin by setting the stage with some general considerations. We then discuss separately two situations that turn out to be critical for calibration. In the first situation we assume that each time you make a probability assessment, you know the outcomes of all the events for which you had previously assessed probabilities. Two theorems are presented for this outcome-feedback situation, one for probability density functions assessed for uncertain continuous quantities and the other for probabilities of discrete events. The results of these theorems are very general: you will always expect to be well calibrated.

In the second situation, outcome feedback is lacking. That is, you must make a large number of probability assessments before learning the outcome of any one of the events being assessed. Here the results are very different: you will expect to be well calibrated only under strong (and thus perhaps rarely met) assumptions.

Along the way, we will try to explicate the implications of these theorems for psychological research on calibration and for the practice of probability assessment.

## General Considerations

Subjectivism and coherence. We assume, throughout this paper, the subjectivist view of probabilities, under which probabilities are coherent degrees of belief, beliefs you would be willing to bet on. Coherence is the key concept in the subjectivist theory; indeed, the usual axioms of probability (i.e., probabilities are numbers from 0 to 1 such that the probability of a union of two mutually exclusive events is the sum of the probabilities of the events and the probabilities of mutually exclusive and exhaustive events sum to one) can be derived from the single idea of coherence (de Finetti, 1980; Kemeny, 1955; Kyberg, & Smokler, 1980; Lehman, 1955; Ramsey, 1980; Savage, 1954; Shimony, 1955). Coherence is defined in terms of betting, specifically, in terms of a Dutch book. A set of probabilities are coherent if no Dutch book can be made from them. A Dutch book is a set of two or more bets placed on the outcomes of one or more uncertain events such that the person holding the bets will surely lose money, regardless of the outcome(s) of the event(s).

As an example of incoherent probabilities, suppose you believe that P(A) = .75 and that P(not-A) = .80. Then I can form two bets that individually seem fair but together constitute a Dutch book, as follows:

Bet 1: If A occurs, you win $1.

If not-A occurs, you lose $3.

Bet 2: If A occurs, you lose $4.

If not-A occurs, you win $1.

Bet 1 is based on your first belief, that A is three times as likely as not-A to occur. Bet 2 is based on your second belief, that not-A is four times as likely to occur as A. But the two bets taken together guarantee that you will lose money. If A occurs, you will win $1 and lose $4, for a net loss of $3. If not-A occurs, you will lose $3 and win $1, for a net loss of $2.

Unless otherwise specified, we assume throughout this paper that probabilities are coherent. In particular, all the probabilities in the theorems below are coherent.

Telling the truth. We note here, in order to exclude them from further consideration, situations in which the payoffs (either monetary or otherwise) ensuing from your assessments motivate you to lie, that is, to report as probabilities something other than your true beliefs. We would not, in general, expect such probabilities to be well calibrated. As an extreme example, suppose you are told that the occurrence of any event to which you have previously assessed a probability of .25 or less will lead to your immediate execution by firing squad, yet you are required to use, at least occasionally, small but non-zero probabilities. Under these conditions, you would expect to be quite badly calibrated (you would certainly not want 20% of your .20 assessments to occur). We thus begin by assuming that our search for conditions under which you expect to be well calibrated will be limited to those situations for which the payoffs encourage you

to tell the truth. Such payoff functions are called proper scoring rules (Staël von Holstein, 1970) or, equivalently, reproducing scoring systems (Shuford, Albert & Massengill, 1966). This limitation to proper scoring rules is not a trivial one in practice. While we know of no assessors who will be shot at dawn if "wrong", we suspect that many real uses of probabilities entail payoffs such as "being made to look the fool" or "increased chance of job promotion" that may not be proper.

Communicating with others. We exclude from this paper any discussion of how you might view other people's probability assessments or how you should use information concerning both their probability assessments and their possible lack of calibration to alter your own beliefs. These topics are discussed by DeGroot and Fienberg (Note 4), Lindley (Note 5), Lindley, Tversky, and Brown (1979), and Morris (1974; 1977).

Almost sure convergence. Finally, a technical note about the theorems that follow: The theorems use a particular kind of convergence of a sequence of random variables to a probability distribution, called almost sure convergence (Loeve, 1960). This is a strong sense of convergence, and can be thought of as meaning, "it is a virtual certainty to you that...."

With these general matters behind us, we turn now to the theorems about calibration.

## Outcome Feedback

A fundamental distinction in the theory of calibration is whether one knows the outcome of all previous events before one is required to state one's probability for the next event. If one has that information, we say that one has outcome feedback.

The following result, due to Pratt (Note 6) and reported here with his permission, concerns the assessment of continuous uncertain quantities, $\tilde{\theta}_i$, in the case of outcome feedback.

Pratt's theorem. Suppose that you are going to assess judgmentally the distributions for a set of parameters $\tilde{\theta}_1$, $\tilde{\theta}_2$, ... Let $S_n$ be the number of $\tilde{\theta}_i$ among $\tilde{\theta}_1$, ... , $\tilde{\theta}_n$ such that $\tilde{\theta}_i$ falls below the p'th fractile of its assessed distribution, and assume for convenience the assessed distributions are all continuous.

Theorem 1. Under these conditions, if you know $\tilde{\theta}_1$, ... , $\tilde{\theta}_k$ when you make your assessment of $\tilde{\theta}_{k+1}$,

(a) You must regard $S_n$ as binomially distributed with parameters n and p;

(b) You are almost sure that $\frac{1}{n} S_n \to p$.

Pratt's proof is given in the Appendix.

An example might help make the theorem clear. Suppose it is your job to predict how many people will attend a particular movie house each night. For each night, you assess a distribution over the then-uncertain parameter, total number of tickets sold ($\tilde{\theta}_i$). Typically, you express each distribution as a set of fractiles. The .25 fractile (p) for night i might be, say, 100, meaning "my probability that attendance will be equal to or less than 100 is .25." The next morning you find out how many tickets were in fact sold (outcome feedback) and then make an assessment for the coming night. Over n such nights, you count the number of times ($S_n$) that the actual attendance was equal to or smaller than the attendance you associated with fractile p (did 100 or fewer people attend on night i?). Part (b) of Theorem 1 shows the convergence of the proportion $S_n/n$ to the fractile p; thus it says that you are virtually certain you are (in the long run) well calibrated.

The importance of Pratt's result is its generality. It places no conditions on what kind of events you are assessing or on what beliefs you hold concerning possible interrelationships among the events. It simply says that under conditions of outcome feedback for the assessment of continuous probability density functions, coherence alone implies calibration.

Dawid's theorem. The case of probabilities assessed for discrete events has been explored by Dawid (in press), who has proven a theorem as general as Pratt's.

Suppose that $A_1, A_2, \ldots, A_j, \ldots$ is a set of events. Suppose that $p_1, p_2, \ldots$ are someone's subjective probabilities for $A_1, A_2 \ldots$. The assessment of $p_i$ is made only after the outcomes of $A_1, \ldots, A_{i-1}$ are known.

Now we wish to form a sum, like $S_n$ in Pratt's Theorem, whose convergence is to be studied. To have a hope of convergence, this sum must have infinitely many events in it. Thus, although it might be natural to form the sum by including all the events, and only those events, within some prespecified interval of subjective probability, we could not be certain, in advance, of having infinitely many events in the sum. For this reason, Dawid introduces the technical device of a probability $\xi_i$ that the event $A_i$ is selected for inclusion in the sum. Here $\xi_i$ can depend on the outcomes of $A_{i-1}, A_{i-2}, \ldots$ in an arbitrary way, and so in particular, $\xi_i$ can depend on $p_i$. Now we can let $n_j = \sum_{i=1}^{j} \xi_i$, which is roughly the number of events from the first j to be included in the sum. Finally, let $S_j = \sum_{i=1}^{j} \xi_i Y_i$, where $Y_i$ takes the

value 1 if $A_i$ occurs and zero otherwise. Then $S_j$ can be interpreted roughly as the number of the first j events that are selected and occur. Finally let $T_j = \sum_{i=1}^{j} \xi_i p_i$, which is roughly the sum of the subjective probabilities for the events that are included in the sum. Then Dawid proves the following:

Theorem 2. If $j \to \infty$ and the $\xi_i$ are chosen so that $n_j \to \infty$, then $\frac{S_j}{n_j} - \frac{T_j}{n_j} \to 0$ almost surely.

This theorem says (roughly) that the difference between the proportion of events that occur, $S_j/n_j$, and the average assessed probabilities, $T_j/n_j$, goes to zero; thus, you expect with near certainty to be well calibrated.

Like Pratt's Theorem, Dawid's Theorem requires no assumptions on the probability structure of the set of events, $A_1$, $A_2$ ..., except feedback. We summarize the import of Pratt and Dawid's work by saying that coherence implies calibration in the presence of outcome feedback.

How is calibration achieved? How might you accomplish the calibration you almost surely expect from Theorems 1 and 2? One obvious possibility is to keep a running tally of assessments and outcomes ("So far I've used .7 six times, of which three were right", etc.). Such a tally could be summarized in a calibration curve, which shows, for all your assessments, the relationship between your assessments (on the abcissa) and the proportion of events that occurred (on the ordinate). If the running tally or the calibration curve indicate poor calibration, you might wish to use the data to find a transformation function $\gamma(E;p)$, where $\gamma$ depends only on p and is a continuous function of p. You intend to say $\gamma$ whenever, in the future, you believe p, and in this way be well calibrated.

If your initial beliefs, p, are coherent, however, you are confronted with the following paradox :

Let $A_1$ and $A_2$ be any two disjoint events having probabilities $p_1$ and $p_2$, respectively. Then since $\gamma$ is to be a probability, it must satisfy $\gamma(A_1 \cup A_2;\ p_1 + p_2) = \gamma(A_1;p_1) + \gamma(A_2;p_2)$

Hence we must have (using continuity of $\gamma$)

$$\gamma(A;\ p) = \alpha p \qquad \text{for all } p,\ 0 \leq p \leq 1, \text{ for some number } \alpha .$$

Noting that $\gamma$ must assign probability one to the universal set,

$$\gamma(\Omega;1) = 1 \text{ implies } \alpha = 1.$$

Hence we have

$$\gamma(A;\ p) = p \qquad \text{for all } p,\ 0 \leq p \leq 1.$$

We conclude that if p and $\gamma$ are both coherent, they are identical (see also Edwards, 1962, Theorem 3). This is a bit of an embarrassment for you, because you had hoped to recalibrate your probabilities on the basis of the function $\gamma$.

Since a calibration curve or calibration tally presents a kind of $\gamma$ function (e.g., for all the times you said ".8," only 65% occurred), it cannot be used to "correct" your assessments. Indeed, we know of no valid model for how people should change their opinions when assessments they believe are coherent systematically deviate from the observed frequencies of events. We speculate that the fault in the transformation-function approach lies in the idea, expressed mathematically by the requirement that the function $\gamma$ depends on p alone, that all events of probability p should be transformed in the same way to some new probability $\gamma$. Instead, we believe that what people should learn from outcome feedback is something about the properties of the world,

not about the properties of p. For example, a person who assesses movie attendance at a particular movie house may learn that musicals are less popular than previously thought, Robert Redford is a big draw, and space adventures are growing in popularity. Such information will lead to calibration through selective changes in belief, with some probabilities increasing and some decreasing.

Under this view, the innocent-sounding assumption of Theorems 1 and 2, that you know the outcome of all previous events before you state your probability for the next event, is seen as a psychologically strong requirement: Not only must you have a perfect memory for past outcomes but also you must be able to use the outcome information to develop a better understanding of the world in which these events occur. Given the oft-documented limitations on human cognitive abilities (Kahneman, Slovic & Tversky, 1982), can people achieve good calibration if given outcome feedback? We know of no laboratory research speaking to this issue, but an abundance of field research on weather forecasters (e.g., Murphy, in press; Murphy & Winkler, 1977) indicates that the answer is yes in their case. The U.S. Weather Bureau instituted probabilistic assessments in precipitation forecasts nationwide in 1965. These forecasts are made under conditions that come close[1] to satisfying Theorem 2. As a group, the forecasters are superbly well calibrated.

In the laboratory, however, subjects will not typically have, or receive, the thorough knowledge of the content area that weather forecasters have. Moreover, in many experiments (see Lichtenstein, Fischhoff & Phillips, 1982) there isn't any single content area. For example, subjects are first asked how many eggs were produced in the U.S. in 1979; next they are asked how many dimples there are on a golf ball, and so forth across a wide variety of topics. Our

speculations on the importance of content learning to the achievement of calibration with outcome feedback leads us to predict that when the content is the same for each item (e.g., predicting the winning horse from a large number of horse-racing past performance charts), laboratory subjects will quickly learn to be well calibrated, whereas they will find greater difficulty in achieving calibration with items of diverse content.

We further speculate on the more applied problem of how to design training programs for assessors who will be receiving outcome feedback in the course of their work. Initial training in the meaning of probabilities will be necessary, to ensure an understanding of coherence. The assessor should understand, for example, that the assessment of a probability of 1.0 indicates not only a willingness, but even an eagerness, to accept a bet which pays, say, $1 if the event in question occurs, and which has a loss of all one's present and future worth if the event does not occur.

Thereafter, however, we would not recommend a training emphasis on calibration, per se. A reliance on calibration tallies or calibration curves may encourage the trainee to search, futilely, for a transformation function, $\gamma$. Furthermore, such summaries discard the very information that, we believe, is essential for satisfying the conditions of Theorems 1 and 2: which events occurred, and which didn't? Thus we would recommend that training in probability assessment be fully integrated with training in the content material. Medical schools, for example, should teach about uncertainty in diagnosis while teaching about diagnosis.

No Outcome Feedback

We turn now to a consideration of calibration without feedback, and begin by exploring whether exchangeability is sufficient to ensure calibration.

Exchangeability. A set of events is exchangeable to you if each event has the same probability, each pair of events has the same probability of occurrence as each other pair, each triple the same, etc. Thus exchangeability is similar to independence of events with the same probability, but more general, since independence requires, for example, that the probability of two events both occurring be the product of their single event probabilites, while exchangeability does not specify what the joint probability is, as long as it is the same for all possible pairs.

Exchangeability is an entirely subjective notion. You must examine your beliefs about joint occurrence to determine whether the events in a set are exchangeable. It is thus presumptuous for us to tell you when events are or are not exchangeable for you.

As an example of exchangeable but not independent events, suppose that $E_i$ is the event that it snows on block i of Minneapolis tonight. Suppose that you judge each block to have the same probability of being snowed on tonight, because it either will snow everywhere, which you consider it will with probability s, or it will snow nowhere, which for you has probability 1 - s. Under these assumptions, the events $E_i$ are exchangeable but not independent to you; you believe that the probability of snow in each pair of blocks is s, each triple of blocks is s, and so forth. But you are sure that you will not be well calibrated on these events, because you are sure that either all of them will occur or none of them will occur.

The following two theorems formalize our conclusion from the Minneapolis snow example: exchangeability does not ensure calibration.

deFinetti's Theorem. A fundamental theorem of deFinetti (see Feller, 1971, p. 228) gives the following structure for a set of exchangeable events: Let $A_1, \ldots$ be an infinite exchangeable set. Then the probability of exactly k out of n events occurring is given by

$$\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} f(p)dp \qquad (1)$$

for all k and n and some density f. Thus a Bayesian with an exchangeable opinion on a set views the events as if a p were drawn from a distribution with density f(p), and then, conditional on p, the events are Bernoulli with probabilty p. Theorem 3 follows from equation (1).

Theorem 3. If $A_1, A_2, \ldots$ are exchangeable, the proportion k/n of events that occur converges to a random variable with density f(p).

Every event in an exchangeable set has, ex ante, the same probability to you, namely, the mean of the distribution with density f(p), as can be seen by substituting k = n = 1 in (1). So before receiving outcome feedback, you will report that mean, $\int_0^1 pf(p)dp$, as your probability for each event. You expect the proportion of events that occur to converge to p, but you are not sure what the value of p is. Your beliefs about the value of p are expressed as f(p). If f(p) is heaped up at its mean, you will have some modest hope that k/n will converge to a value close to its mean, so that you will be close to being well calibrated, but you would not be astonished if k/n were to converge on any other value of p for which f(p) has non-zero probability. In the Minneapolis snow example, f(p) has spikes at p = 0 and p = 1 ("...it will either snow everywhere...or it will snow nowhere") and is zero elsewhere, so here you are sure you will not be well calibrated. In general, Theorem 3 says that with exchangeable events, you will not automatically expect to be well calibrated.

Two further examples may help to illustrate Theorem 3. First, we ask you to assess the probability that a ball drawn from an urn will be red, for each of many such draws (with replacement). The urn contains either three red balls and one white ball or three white balls and one red ball, but you don't know which; you put probability 1/2 on each of these two possibilities. Thus you assess, for each draw, the probability of a red ball as .5. All draws are exchangeable for you, but you do not expect that 50% of the draws will be red balls. Instead, you expect that either 25% or 75% of them will be red. Your f(p) is spiked at the values .25 and .75, and you would be extraordinarily surprised, indeed, quite suspicious of our honesty, should we report to you that you were well calibrated on a long series of draws.

Consider now a national election. Suppose that you are assessing the probability of winning for a set of liberal candidates across the nation, each of whom, you believe, has probability .6 of winning. Further assume that these races are exchangeable to you. But suppose that at the same time you are aware of a nation-wide effort by ultra-conservatives to encourage all other conservatives to vote in unprecedented number. If such an effort succeeds, you expect fewer than 60% of the liberals to win; if the effort fails or backfires, you expect more than 60% to win. Your beliefs about the probability of success by the ultra-conservatives are expressed in f(p), which has a mean of .6. For most reasonably smooth, widely spread f(p) functions, no particular set of outcomes will surprise you much more than any other. Still, these considerations, and the realization that you likely will not be well calibrated, do not lead you to change your assessment of .6 for each race in the set.

The following theorem adds a further constraint on your beliefs, in order to characterize sets of exchangeable events for which calibration is a consequence of coherence, under conditions of no outcome feedback.

Theorem 4. Let $A_1, A_2, \ldots, A_i, \ldots$ be a set of events exchangeable to you, let $S_n$ be the number of them that occur among the first n, and let p be the probability that any given $A_i$ occurs. Then you are almost sure that

$$\frac{S_n}{n} \to p$$

if and only if the set is a set of independent events to you. Theorem 4 follows immediately from Theorem 3.

Thus, with no outcome feedback, if events are both exchangeable to you and independent to you, you will expect to be well calibrated.

Other possible conditions. Theorems 3 and 4 do not speak to the question of calibration in the absence of both outcome feedback and exchangeability. A more extensive mathematical treatment of calibration in the absence of outcome feedback may be found, though expressed in a different vocabulary, in ergodic theory (Breiman, 1968), interpreted subjectively.

Ergotic theory arose in the physics of systems composed of large numbers of identical particles (Reif, 1965, p. 583ff). Interest centers on conditions under which the average over time of some system parameter under a fixed system state is equal to the average at a fixed time across all possible system states. The parallel with calibration is as follows: Suppose the system parameter under study is a 0-1 variable. Then its average over time is a relative frequency. The

average at a fixed time across all system states may be interpreted as the subjective probability of the system parameter's occurrence at that time. The question can thus be rephrased: Under what conditions does the relative frequency approach the subjective probability? For large classes of stochastic processes, these relative frequencies converge to a particular random variable, as in Theorem 3, and under very special circumstances they converge to a constant, here interpreted as subjective probability, as in Theorem 4. However, none of these classes has the subjective intuitive appeal of exchangeability, so we will not report those further results here. Instead, we simply note that there do exist other conditions under which one would expect to be well calibrated, but we doubt that these conditions would ever be met either in laboratory or in real-life settings of subjective probability assessment.

Violations of independence. Independence, like exchangeability, is an entirely subjective notion, speaking to the beliefs you hold about events. Two ways of testing independence are to ask yourself either "Will knowing the outcome of event B change my belief about A?", that is, $P(A|B) = P(A)$ is required for independence, or "Is the probability of the joint occurrence of A and B the same as the product of the single probabilities?", that is, $P(A \cap B) = P(A) \cdot P(B)$ is required for independence. Since independence is an essential condition for you to expect to be well calibrated with exchangeable events in the absence of outcome feedback, we here briefly discuss ways in which independence might be violated.

The examples of exchangeability given above (snow in Minneapolis, red and white balls drawn from an urn, and the elections predictions) are all examples of non-independence; in each case the non-independence is driven by a causal

link between the events. Similarly, a physician's beliefs about the likelihood of a highly contagious flu bug present in the community will generate non-independence in the physician's diagnostic assessments of otherwise unrelated patients.

Non-independence might also result from events whose definitions have some logical relation. For example, suppose I ask you to assess the probability that the population of one city exceeds the population of another, for a large set of pairs of cities (Lichtenstein & Fischhoff, Note 7). Suppose that several of these city-pairs have one city in common. For example, you assess $p_1$ that Seattle is more populous than Phoenix and $p_2$ that Seattle is more populous than Wichita, etc. If you realize that you might be quite wrong on the population of Seattle, then you will recognize that you will tend to systematic error in your assessments of all the Seattle pairs; the product of $p_1$ and $p_2$ will thus not be equal to your probability that Seattle is more populous than both Phoenix and Wichita.

In general, beliefs about the possible interconnectedness of your beliefs as well as beliefs about the interconnectedness of events will lead to nonindependence.

Implications for experimental research. Virtually all the published research on calibration (except the literature on weather forecasters) has been conducted without outcome feedback. When feedback has been given, it usually has involved information about calibration (i.e., calibration tallies or curves) without providing the subjects with answers to the items assessed (Adams, & Adams, 1958; Lichtenstein, & Fischhoff, 1980; Oskamp, 1962). Thus the experimenters should have expected to observe good calibration and discussed the psychological

import of contrary findings only if the items used in the research were viewed by the subjects as independent. No one has ever asked research subjects about independence, but we suspect that strict independence among items has been rare.

The use of general-knowledge or "trivia" items has been popular in calibration research. Here, independence may be violated because of overlapping content of items. When collecting the large number of items needed for such research, it is difficult to avoid such overlap entirely. One subject of Lichtenstein and Fischhoff's (1980) took great glee in pointing out to the experimenters such problems (for example, the same alternative was used as one of two possible answers to two different questions; it could not have been the correct alternative for both).

Some research had used items all of which refer to the same content (e.g., 100 handwriting samples, for each of which the task was to assess the probability that it was written by a European rather than an American; Lichtenstein & Fischhoff, 1977). Research on assessments of uncertain quantities has sometimes used group-generated proportions for their items (e.g., Alpert & Raiffa, 1982 ; Moskowitz & Bullers, Note 8; Selvidge, Note 9). In these tasks, the subjects were first asked facts about themselves, like "Do you prefer bourbon to scotch?"; then they assessed the proportion of subjects answering yes to those questions. Suppose in these situations that the assessor uses some strategy or theory to aid in making all the assessments. For example, subjects assessing group-generated proportions may believe that "Most people have preferences like mine." For the handwriting task, an assessor may believe that cramped

writing is more likely to be European than American. If the assessor has uncertainty about the validity of the theory, that uncertainty would lead to lack of subjective independence among the events.

A relatively new area of research in calibration deals with the assessment of future events (most of this research has not yet been published, but see Fischhoff and Beyth, 1975). In developing a large number of items all of which will be decided within a relatively short period of time (so that the experimenters can score the items), it is difficult to avoid items with relationships among them (e.g., will Democrat X win the upstate election? Will Democrat Y win downstate?).

For these reasons, we suspect that the items used in past research would not have all been judged strictly independent by all subjects. Thus, one could say that the finding that subjects are often badly calibrated (usually, overconfident) has little meaning, since Theorem 4 says that one wouldn't expect good calibration. However, we reject this reasoning. We believe that the non-independence in most past studies was so small as to have virtually no impact on the results. With large numbers of items, most subjects do not remember previous items when responding to the current one. Moreover, subjects who use simplifying theories to aid them in the task ("cramped writing is more likely to be European") probably do not question or doubt their own theories, and it is the doubt about such a theory, not the theory itself, which induces non-independence. Finally, with items of diverse content, one could reasonably expect that the effects of the few interdependencies would tend to cancel each other out (some leading to too many items being correct, others to too few) in the overall calibration. Therefore,

we find ourselves still believing the results of past research purporting to show that people are generally overconfident in the extent and accuracy of their own knowledge. Still, research that specifically addresses problems of subjective non-independence is needed.

Let us suppose for a moment that previous laboratory findings will replicate when subjective independence of items is carefully satisfied; specifically, suppose that people tend to be badly calibrated with trivia items, but that this overconfidence can be eliminated to some degree by showing the assessors their own calibration curves (without outcome feedback; Lichtenstein & Fischhoff, 1980) or by making the assessors focus on reasons why they might be wrong (Koriat, Lichtenstein & Fischhoff, 1980). How should we view such results in the light of Theorem 4? The subjects in psychological experiments are typically not Bayesians, indeed, are not informed at all about probability theory. The instructions they receive do not educate them about the fine points of coherence. Thus we would not be surprised if subjects' responses aren't coherent probabilities and are quite easy to change.

Despite some promising beginnings (Pitz, 1974; Ferrell & McGoey, 1980), we still do not have an adequate theory about how people form and express feelings of uncertainty. Even when the exchangeability and independence assumptions are not met, research on calibration can serve to advance our understanding of these processes. For example, experimental manipulations that change calibration (e.g., Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977) can suggest what processes are involved in the formation of uncertainty. In these endeavors to build cognitive psychology, however, we should take care not to claim that good calibration is a goal to which reasonable people should always strive.

On becoming rich. Calibration tasks do not provide evidence about the assessor's coherence (such evidence would be sought in the assessments of not only P(A) and P(B), but also P(A|B), P(B|A), P(A∩B), and so forth). Thus we cannot use calibration tasks to become rich by developing Dutch books on other assessors. However, if people are generally overconfident in assessing probabilities, seemingly attractive bets having negative expected value to the assessor can easily be written. In a wide variety of tasks involving the assessments of fractiles for uncertain quantities, for example, naive assessors are, overall, so overconfident that some 40% of all true answers lie outside their central 99% confidence intervals (Lichtenstein et al., 1982 ). Such assessors should be willing to accept bets that seem quite favorable to them yet provide the offerer excellent odds of winning. Subjects' willingness in such situations has been reported by Pitz (1974), whose subjects showed a systematic preference for betting on the central regions rather than on the tails of their own assessed distributions, and by Fischhoff, Slovic, and Lichtenstein (1977), some of whose subjects expressed eagerness to enter what was in fact a losing "trivia-hustling" game.

Implications for practice. Some authors (e.g., Brown, Kahr & Peterson, 1974, pp. 437-438) have recommended that probability assessors study their own calibration on trivia items and apply the lessons so learned to their real-world assessments. We recommend against such a training procedure. First, a justification of this procedure would require strong and quite dubious assumptions about exchangeability and independence between the set of trivia questions and the set of real-world events to be considered later. Secondly, there is no evidence that the lessons learned from such exercises ("I'm overconfident with trivia

items") would be appropriately generalized when people make assessments in their area of expertise. Finally, with sufficient effort by the designers, trivia items can be written that strike everyone as independent (What's the relationship between golf ball dimples and egg production?). This may lull assessors into carelessly assuming, without critical examination, that the real-world events they are assessing are also independent.

A reasonably stable measurement of calibration requires a large amount of data. It may be rare to find situations outside the laboratory in which an assessor makes a large number (e.g., 500) of assessments before receiving any outcome feedback. More typically, an assessor may make only a few assessments, five or twenty, before receiving outcome feedback. If so, the situation is more like that of Theorems 1 and 2; concern for independence is lessened, and the focus of the assessor's attention in evaluating the feedback is to learn which events occurred and why.

When a large number of assessments are made in practical settings before outcome feedback is received, the question of independence is crucial in evaluating calibration. Sometimes non-independence will be obvious, as when a set of assessments about future economic events all depend on, say, the growth of the national economy. Then the message from Theorems 3 and 4 is that assessors should just not care about their calibration. Indeed, it may be misleading even to look at a calibration tally.

Still, there may be occasions when it is appropriate to study calibration in practical settings. What can be learned from such

exercises? Suppose that you assess probabilities (without outcome feedback at the time) for a set of events that you believe are exchangeable and independent. Suppose the data (when later you get outcome feedback) suggest that you are quite badly calibrated. How should you regard these data? You might attribute the source of your trouble to one or more of the following causes:

1. It might be that you are indeed well calibrated in the long run, but the finite sample of data is a rare one. This possibility becomes less likely as the sample size increases, but can never be entirely ruled out.

2. Despite your best efforts, your probability assessments may be incoherent. Calibration tallies or curves do not, in themselves, provide evidence of incoherence. It may be, however, that if you assessed probabilities of joint events (unions, intersections, etc.), you would discover a way to write a Dutch book against yourself. The existence of such a Dutch book would show that your assessments were incoherent.

3. Your expressed probabilities and assumptions may not have truly represented your beliefs. For example, you may have been too quick to assume independence among the events, without critical examination. You should especially be alert to the existence of previously unnoticed conditioning events (e.g., assuming that the U.S. would not enter a severe depression when forecasting future energy needs) that linked all your assessments.

4. It may be that your beliefs about the world are wrong. You may have, at the time, unquestioningly believed that the incidence of cramped handwriting is more likely among Europeans than Americans,

but when outcome feedback becomes available, you learn that your belief was false. The theorems presented above are subjectivist; they describe the conditions under which you believe you are well calibrated. But they do not guarantee that you will, in fact, be well calibrated.

The theorems presented above will not help you to choose among these four possibilities. Only further exploration and consideration of the events in question may inform you of the source of your difficulty. Calibration, we conclude, is dependent more on the characteristics of and interrelationships among the events than on your feelings of uncertainty concerning each event considered in isolation.

Reference Notes

1. Cambridge, R. M., & Shreckengost, R. C. Are you sure? The subjective probability assessment test. Unpublished manuscript. Langley, Va.: Office of Training, Central Intelligence Agency, 1978.

2. Raiffa, H. Assessments of probabilities. Rough draft, Harvard University, January 1969.

3. Roberts, H. V. On the meaning of the probability of rain. A paper prepared for presentation to the First National Conference on Statistical Meteorology, American Meteorological Society, May 1968 in Hartford, Connecticut.

4. DeGroot, M. H., & Fienberg, S. E. Assessing probability assessors: Calibration and refinement. Technical Report 205, Department of Statistics, Carnegie-Mellon University, Pittsburgh, Pa., 1981.

5. Lindley, D. V. The improvement of probability judgments. Unpublished manuscript, 1981.

6. Pratt, John W. Must subjective probabilities be realized as relative frequencies? Memo dated September 27, 1962.

7. Lichtenstein, S., & Fischhoff, B. How well do probability experts assess probabilities? Technical Report PTR-1092-80-8. Woodland Hills, Ca.: Perceptronics, Inc., 1980.

8. Moskowitz, H., & Bullers, W. I. Modified PERT versus fractile assessment of subjective probability distributions. Paper No. 675, Purdue University, 1978.

9. Selvidge, J. Experimental comparison of different methods for assessing the extremes of probability distributions by the fractile method. Management Science Report Series, Report 75-13. Boulder, Co.: Graduate School of Business Administration, University of Colorado, 1975.

References

Adams, P. A., & Adams, J. K. Training in confidence judgments. American Journal of Psychology, 1958, 71, 747-751.

Alpert, M., & Raiffa, H. A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, in press.

Balthasar, H. A., Boschi, R. A. A. & Menke, M. M. Calling the shots in R & D. Harvard Business Review, 1978, 56, 151-160.

Breiman, L. Probability. Reading: Addison-Wesley, 1968.

Brown, R. V., Kahr, A. S., & Peterson, C. Decision analysis for the manager. New York: Holt, Rinehart, & Winston, 1974.

Dawid, A. P. The well-calibrated Bayesian. Journal of the American Statistical Association, in press.

de Finetti, B. The theory of probability, 2 volumes. New York: Wiley, 1974.

de Finetti, B. Foresight: Its logical laws, its subjective sources. In H. E. Kyborg, Jr., and H. E. Smokler (Eds.), Studies in subjective probability, 2nd ed. New York: Kreiger, 1980.

Edwards, W. Subjective probabilities inferred from decisions. Psychological Review, 1962, 69, 109-135.

Feller, W. An introduction to probability theory and its applications, Vol. II. New York: Wiley, 1971.

Ferrell, W. R., & McGoey, P. J. A model of calibration for subjective probabilities. Organizational Behavior and Human Performance, 1980, 26, 32-35.

Fischhoff, B., & Beyth, R. "I knew it would happen"--remembered probabilities of once-future things. Organizational Behavior and Human Performance, 1975, 13, 1-16.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance, 1977, 3, 552-564.

Harrison, J. M. Independence and calibration in decision analysis. Management Science, 1977, 24, 320-328.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Kemeny, J. G. Fair bets and inductive probabilities. Journal of Symbolic Logic, 1955, 20, 263-273.

Koriat, A., Lichtenstein, S., & Fischhoff, B. Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 1980, 6, 107-118.

Kyburg, H. E., Jr., & Smokler, H. E. Introduction. In H. E. Kyburg, Jr. and H. E. Smokler (Eds.), Studies in subjective probability, 2nd ed. New York: Krieger, 1980.

Lehman, R. S. On confirmation and rational betting. Journal of Symbolic Logic, 1955, 20, 251-262.

Lichtenstein, S., & Fischhoff, B. Do those who know more also know more about how much they know? The calibration of probability judgments. Organizational Behavior and Human Performance, 1977, 20, 159-183.

Lichtenstein, S., & Fischhoff, B. Training for calibration. Organizational Behavior and Human Performance, 1980, 26, 149-171.

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

Lindley, D. V., Tversky, A., & Brown, R. V. On the reconciliation of probability assessments (with discussion). Journal of the Royal Statistical Society, A, 1979, 142, Part 2, 146-180.

Loeve, M. Probability Theory, Princeton: Van Nostrand, 1960.

Lusted, L. B. Introduction to medical decision making. Springfield, IL: Thomas, 1968.

Morris, P. A. Decision analysis expert use. Management Science, 1974, 20, 1233-1241.

Morris, P. A. Combining expert judgments: A Bayesian approach. Management Science, 1977, 23, 679-693.

Murphy, A. H. A new vector partition of the probability score. Journal of Applied Meteorology, 1973, 12, 595-600.

Murphy, A. H. Subjective quantification of uncertainty in weather forecasts in the United States. Meteorologische Rundschau, in press.

Murphy, A. H. & Winkler, R. L. Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? National Weather Digest, 1977, 2, 2-9.

Oskamp, S. The relationship of clinical experience and training methods to several criteria of clinical prediction. Psychological Monographs, 1962, 76 (28, Whole No. 547).

Pitz, G. F. Subjective probability distributions for imperfectly known quantities. In L. W. Gregg (Ed.), Knowledge and cognition. Potomac, Maryland: Lawrence Erlbaum, 1974.

Ramsey, F. P. Truth and probability. In H. E. Kyburg, Jr., and H. E. Smokler (Eds.), Studies in subjective probability, 2nd ed. New York: Kreiger, 1980.

Reif, F. Fundamentals of statistical and thermal physics. New York: McGraw-Hill, 1965.

Savage, L. J. The foundations of statistics. New York: Wiley, 1954.

Shimony, A. Coherence and the axioms of confirmation. Journal of Symbolic Logic, 1955, 20, 1-28.

Shuford, E. H., Jr., Albert, A., & Massengill, H. E. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.

Staël von Holstein, C.-A. S. Assessment and evaluation of subjective probability distributions. Stockholm School of Economics, Economic Research Institute, Stockholm, 1970.

U. S. Nuclear Regulatory Commission. Reactor safety study: An assessment of accident risks in U.S. commercial nuclear power plants. WASH 1400 (NUREG-75/014), Washington, D. C.: The Commission, 1975.

## Footnotes

The preparation of this article was supported in part by the Office of Naval Research under Contract N00014-80-C-0150 to Perceptronics, Inc.

Request for reprints may be sent to Sarah Lichtenstein, Decision Research, A Branch of Perceptronics, 1201 Oak Street, Eugene, Oregon 97401.

The discussions which led to the development of this paper were initiated at an invitational workshop on "Expert Judgments for Policy Analysis" sponsored by the Department of Engineering and Public Policy of Carnegie-Mellon University and the Biomedical and Environmental Assessment Division at Brookhaven National Laboratory with funds from the Alfred P. Sloan Foundation, the National Science Foundation, the MPC Corporation and the United States Energy, Research and Development Administration. We are grateful, also, to the many friends with whom we have discussed these ideas, including Ruth Beyth-Marom, Baruch Fischhoff, Seymour Geisser, Max Henrion, Richard Jeffrey, Jill Larkin, Dennis Lindley, Don MacGregor, L. D. Phillips, Mark Schervish, Teddy Seidenfeld, Amos Tversky, and Bob Winkler.

1. These forecasters usually make three forecasts at a time, for example, one for the forthcoming six hours and two more for the two 12-hour periods thereafter. They thus do not know the outcomes of the first two of these when assessing the third.

## Appendix

Proof of Theorem 1. (Pratt)

Let $\tilde{X}_i = 1$ if $\tilde{\theta}_i$ falls below the p-point of its assessed distribution, $\tilde{X}_i = 0$ otherwise. Then

$$P(\tilde{X}_1 = 1) = p \tag{1}$$

$$P(\tilde{X}_2 = 1 | \theta_1) = p \tag{2}$$

and in general

$$P(\tilde{X}_{k+1} = 1 | \theta_1, \ldots, \theta_k) = p \quad \text{for all } k \text{ and all } \theta_1, \ldots, \theta_k. \tag{3}$$

Since $\tilde{X}_1$ is a function of $\tilde{\theta}_1$, it follows from (2) and the properties of conditional expectation that

$$P(\tilde{X}_2 = 1 | X_1) = p. \tag{4}$$

Since $\tilde{X}_1, \ldots, \tilde{X}_{k-1}$ are functions of $\tilde{\theta}_1, \ldots, \tilde{\theta}_{k-1}$, it follows similarly that

$$P(\tilde{X}_{k+1} = 1 | X_1, \ldots, X_k) = p \quad \text{for all } k \text{ and for all } X_1, \ldots, X_k. \tag{5}$$

From (1), (4), and (5) it follows that $\tilde{X}_1, \tilde{X}_2, \ldots$ is (marginally, i.e., initially) a Bernoulli process with parameter p and hence $\tilde{S}_n = \sum_{k=1}^{n} \tilde{X}_k$ is Binomial. This proves part (a). Part (b) follows immediately.