CALIBRATION, COHERENCE, AND SCORING RULES*

TEDDY SEIDENFELD†

Department of Philosophy
Washington University in St. Louis

Can there be good reasons for judging one set of probabilistic assertions more *reliable* than a second? There are many candidates for measuring "goodness" of probabilistic forecasts. Here, I focus on one such aspirant: calibration. Calibration requires an alignment of announced probabilities and observed relative frequency, e.g., 50 percent of forecasts made with the announced probability of .5 occur, 70 percent of forecasts made with probability .7 occur, etc.

To summarize the conclusions: (i) Surveys designed to display calibration curves, from which a recalibration is to be calculated, are useless without due consideration for the interconnections between questions (forecasts) in the survey. (ii) Subject to feedback, calibration in the long run is otiose. It gives no ground for validating one coherent opinion over another as each coherent forecaster is (almost) sure of his own long-run calibration. (iii) Calibration in the short run is an inducement to hedge forecasts. A calibration score, in the short run, is improper. It gives the forecaster reason to feign violation of total evidence by enticing him to use the more predictable frequencies in a larger finite reference class than that directly relevant.

1. Introduction—Calibration and Calibration Curves. The radio announcer reports a "30 percent chance of precipitation" for tomorrow. A phone call for the local weather forecast yields the same message. But the Channel 4 TV weatherman says there is only a 20 percent chance of rain for tomorrow and he is billed as having the most reliable weather predictions in town. Not surprisingly, it is Channel 4 itself that advertises the superiority of the Channel 4 weather predictions.

The immodest claim made by Channel 4 on behalf of its own skills at meteorological prognostication prompts an important question. Can there be good reasons for judging one set of probabilistic assertions more *reliable* than another? The problem is hardly new (see Finetti 1972, chap.

*Received December 1983; revised June 1984.

†I thank Jay Kadane and Mark Schervish for helpful discussions about their important work on calibration, and Isaac Levi for his constructive criticism of this and earlier drafts. Also, I have benefited from conversations with M. De Groot and J. K. Ghosh.

Preliminary versions of this paper were delivered at the Meeting of the Society for Philosophy and Psychology, May 13–16, 1982, London, Ontario; and at Session TA10, "Modeling Uncertainty," of the TIMS/ORSA conference, April 27, 1983, Chicago, Illinois.

Research for this work was sponsored by a Washington University Faculty Research Grant.

Philosophy of Science, 52 (1985) pp. 274–294.

3) and specific proposals for measuring the "goodness" of probabilistic weather forecasts date back some thirty years, at least (see Brier 1950). Needless to say, what we learn about rating weather predictions can be applied to appraising expert handicappers of all sorts.

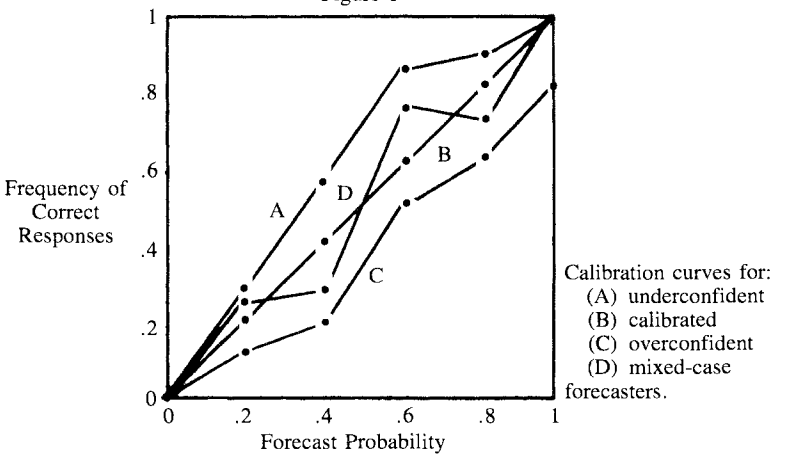
There are many approaches to assessing "goodness" of probabilistic forecasts (see Murphy and Epstein 1967). Here I am concerned with one such aspirant: *calibration*, which carries philosophically interesting consequences for the debate between personal and frequency interpretations of probability.

DEFINITION: A set of probabilistic predictions are (*well*) *calibrated* if p percent of all predictions reported at probability p are true.

In other words, a forecaster's predictions are calibrated if half of those made at "probability .5" are true, if 70 percent of those made at "probability .7" are true, etc. Simply put, calibration is an alignment of relative frequency of occurrence and assertions grouped by constant "probability."[1]

The contrast afforded by calibration is not intended merely for descriptive purposes. Better, it is thought, that the forecasts be calibrated than not. Efforts have been made to tabulate responses to questionnaires, to construct calibration curves, so that individual profiles of those well calibrated, or overconfident, or underconfident in their probability assessments might be made observable. Figure 1 illustrates the technique for a graphic display of departures from calibration (see Lichtenstein and

Figure 1



[1]There is the obvious generalization to calibration $\pm \epsilon$, but nothing said here depends upon whether ϵ is small instead of $\epsilon = 0$.

Fischhoff 1977). Well-experienced decision theorists have speculated that it is a common feature of ordinary forecasts that they are overconfident about expected values (see Pratt and Schlaifer forthcoming; Alpert and Raiffa 1982; and Lichtenstein, Fischhoff, and Philipps 1982).

What merit is there in having calibrated weather forecasts? Alternatively, what risk do I run in relying upon a weatherman who is poorly calibrated?[2] Of course, that depends upon what I do with the forecasts I hear. Suppose, for simplicity, I listen to one forecaster and adopt his announced "chance of precipitation" as my personal probability for rain/snow tomorrow.[3] Then is there an advantage to me to use a calibrated weatherman? M. J. Schervish (1983) adduces an interesting reason to think there may be, but rightly judges it insufficient for an affirmative answer.

If we consider simple (binary) decisions I am to make, where the relevant states of uncertainty are "rain" versus "no rain," we may contrast the value to me of different forecasters by the efficacy of the decisions I would make on condition that I adopt their forecast probabilities as my own. Without loss of generality (and assuming states are probabilistically independent of acts), a simple decision is summarized in the 2×2 matrix:

Figure 2

	S_1	S_2
A_1	0	$-x$
A_2	$-(1 - x)$	0

S_1 and S_2 are the states "rain" and "no rain," and x $(0 \leq x \leq 1)$ is the "loss" upon choosing option A_1 when state S_2 occurs and $(1 - x)$ the "loss" upon choosing option A_2 when state S_1 occurs. To maximize expected utility in this choice, adopt A_1 if the probability of "rain" is as great as the "loss" x, i.e.,

if $P(S_1) \geq x$, then choose A_1.

Consider a sequence of distinct but similar decisions, where the respective "losses" are constant. A forecaster's net performance can be gauged by the average "loss" incurred when decisions are made based on his announced "probability of rain." Following Schervish (1983, p. 3) say

[2]Murphy and Winkler (1977) report that weathermen tend to be well calibrated.

[3]This is how De Groot and Eriksson (forthcoming) define calibration from a subjectivist point of view. That is, forecast B is calibrated from the perspective of subjectivist A if A is prepared to adopt B's announced forecasts as his own, i.e., if

$$P_A(E|P_B(E) = r) = r;$$

where $P_B(E)$ is B's announced forecast for event E, and $P_A(\cdot|\cdot)$ is A's subjective conditional probability. On this account, each coherent subjectivist is calibrated by his own lights.

that forecaster 1 *performs at least as well as* forecaster 2 on a sequence of distinct but similar decisions, if for each "loss" (for all x) the average "loss" incurred when decisions are based on forecaster 1 is no greater than that incurred when decisions are based on forecaster 2. In other words, forecaster 1 outperforms forecaster 2 if similar decisions made with his forecast 1 probabilities dominate those of his rival (as a function of average "loss").

One merit in a calibrated forecaster is made clear by comparing the *performance* of a poorly calibrated weatherman with that of his calibrated counterpart. Over a (finite) sequence of daily forecasts, let forecaster 2 be miscalibrated in his announced "chance of precipitation." On the basis of forecaster 2's calibration curve (see Figure 1), construct a well-calibrated forecaster 1 simply by using the observed relative frequency of "rain" corresponding to each of forecaster 2's forecasts. For example, if the relative frequency of "rain" is 70 percent on days when forecaster 2 announces a "chance" of .6, then forecaster 1's prediction is a "chance" of .7 whenever forecaster 2 announces a "60 percent chance of rain." In this fashion we construct a well-calibrated counterpart (forecaster 1) to forecaster 2.

There is an interesting result about forecasters and their calibrated counterparts.

THEOREM 1: (Schervish 1983, T.13): If a forecaster is not well calibrated over a given (finite) sequence of events, then his well-calibrated counterpart *outperforms* him in similar decisions taken over this sequence.

In this sense, it pays to rely on calibrated forecasts. However, the result has little operational value, as construction of the "counterpart" depends upon the calibration curve for the miscalibrated forecaster. That is information typically unavailable until after the decisions must be made. Moreover, if choices can await construction of a calibration curve for a forecaster, then, as one learns not only whether or not the forecaster is calibrated but also whether or not it "rained" on a given day, the decision theoretic claim of Theorem 1 is moot. Under these conditions the choices are made under "certainty" and forecaster reliability is irrelevant.

Though the data summarized in a calibration curve may come too late to affect my decisions involving those past events for which the calibration curve is constructed, am I not now in a better position to use the weatherman's *next* forecast after seeing his calibration curve for, say, his last 1,000 daily forecasts? If the weatherman is (practically) well calibrated for this long stretch, am I obligated by minimal principles of induction to use his announced forecasts as my personal probability for rain? The answer depends upon whose account of "direct inference" you

adopt. As I understand Kyburg's position (Kyburg 1974), the frequency information contained in the weatherman's calibration curve (in the absence of other frequency data about tomorrow's weather) fixes the *epistemological* probability of "rain" close to the announced "chance." But on accounts of direct inference that adhere to a (Bayesian) requirement of *conditionalization,* e.g., Levi's program (see Levi 1981 for pertinent discussion), no such inference is mandated.[4]

Suppose, either because inductive logic is Kyburgian or because I am prepared to make the requisite irrelevance assumptions, I think the weatherman's observed calibration determines my reaction to his forecast for "rain" tomorrow. Thus, if the forecast is for an "x percent chance of precipitation" but the calibration curve reveals a y percent of occurrence of "rainy" days whenever the forecast is "x percent," then I assign a probability of $.y$ to "rain" tomorrow. Under these conditions, the fact that forecasts are calibrated, as opposed to being overconfident, is *not* what makes them informative to me.

For example, a poorly calibrated weatherman, one who is quite overconfident in his high-probability predictions, may nonetheless be more valuable a source of weather information than some well-calibrated rival. In an extreme case, a weatherman is calibrated if he announces the same "chance" of precipitation day after day, where that is the correct overall percentage of "rainy" days, e.g., 20 percent for the Chicago area. Hence, the poorly calibrated weatherman might be a better discriminator for judging rain tomorrow, though his discriminations are not *positively* correlated with his announced "chance."

Knowing all this, of course, I can convert the poorly calibrated forecaster into his well-calibrated counterpart by a simple correction, i.e., transform the high-probability forecasts into (suitably) low-probability forecasts as dictated by the calibration curve. In that case, there are two sets of calibrated predictions: the well-calibrated one-note weatherman ("There's a 20 percent chance, again!") and the recalibrated forecasts of the weatherman with better resolving power.[5]

These informal considerations raise a paradox. If I can improve the calibration of the overconfident weatherman by a transformation that depends solely upon his calibration curve, why can't he do the same for

[4]In a program such as Levi's, the "direct inference" depends upon additional assumptions of confirmational irrelevance not fixed by the information contained in a calibration curve. I have examined several consequences of this debate for statistical inference in an earlier essay (Seidenfeld 1978).

[5]For a useful discussion of how well-calibrated forecasters might be compared see the papers by De Groot and Fienberg 1981, 1982; and by De Groot and Eriksson forthcoming. These use a partial ordering of calibrated distributions coincident with the partial ordering induced by the statistical notion of sufficiency of experiments (see Blackwell and Girshick 1954, chap. 12).

himself? It would appear that each forecaster is in a position to become calibrated if only he pays attention to his own track record as reported by his calibration curve. Hence, principles of rationality should ensure whatever anyone needs by way of calibration on the condition that past performance is available.[6] How can calibration be an index of reliability if it is so easy to attain?

The "paradox" is resolved, I maintain, by distinguishing two senses of calibration. To repeat, *qua norm,* calibration requires that a sequence of probability-p level forecasts have a relative frequency of p percent correct assertions. We can impose calibration as a requirement for some (hypothetical) long-range, denumerable sequence of predictions. Or we can impose it as a requirement for some finite set of forecasts. The former has obvious ties to the limiting-frequency interpretation of probability. The latter suggests a finite-frequency interpretation. On either reading, calibration is linked to a frequency interpretation.

Understood this way, calibration (in either sense) stands outside the minimal conditions imposed by a subjective (Bayesian) interpretation. It is an all too familiar objection to subjectivism that its standards of rationality are overly liberal; that it tolerates coherent but wildly unreasonable views which bear little semblance to the facts. But our informal analysis suggests that calibration is easily achieved, even for a subjectivist, merely by attending to what you learn from your own (running) calibration curve: construct your own calibrated counterpart. And decision theoretic concerns as Schervish's argument provides) suggest that, even for a subjectivist, it pays to be calibrated. All told, it seems that subjectivism includes a commitment to calibrated personal probabilities.

In the remainder of this paper, I contrast the minimal requirements of subjectivism with calibration (taken as a norm) in each of its two versions: long- and short-range calibration. The lesson we learn from this contrast is rather surprising. Long-range calibration is, at best, otiose. Short-range calibration is a norm that promotes lies and deceit!

2. Calibration and Subjectivism—basic considerations. The subjective interpretation of probability is usually identified with the positions advocated by Finetti (1972, 1974) and Savage (1954). Though there is lively dispute over exactly what constitutes a "subjectivist," I take the following three as core postulates.

1. *Coherence:* A rational agent has (at time t) a belief-state modeled

[6]This conclusion stands in sharp contrast to the claim, "Earlier studies . . . have reported some evidence that people who know more are better calibrated," (Lichtenstein and Fischhoff 1977, p. 163), unless these authors cleverly intend to refer to those with better *self*-knowledge, e.g., those with better memories!

by a (finitely additive) conditional probability $P_K(\cdot|\cdot)$ defined over an algebra of propositions.

2. The agent obeys a *total evidence* requirement: the space of possibilities is relativized to background knowledge K (deductively closed and consistent). K is what the agent accepts as evidence (at t).

3. *Conditionalization:* Conditional probability $P_K(\cdot|\cdot \,\&\, e)$ determines the coherent probability $P_{K'}(\cdot|\cdot)$ for the expanded body of knowledge K' obtained by adding evidence e to K. Loosely put, conditionalization equates the "updated" unconditional probability $P_{K'}(\cdot)$ with the "initial" conditional probability $P_K(\cdot|e)$.[7]

Imagine you (a subjectivist) take a "test" designed to examine your skill at forecasting, as measured by your observed calibration. As a subjectivist, what do you learn from the calibration curve for the forecasts you gave on the "test"? Two warnings are in order here.

First, you *cannot* hope to construct your calibrated but coherent counterpart by a transformation based on the observed calibration curve unless the "test" questions are logically independent (given your background information K). In general, if $P(\cdot)$ is a coherent (fine) probability over an algebra of propositions, then $\tau[P(\cdot)]$ is coherent if and only if τ is the identity function. (This theorem is reported in Kadane and Lichtenstein 1982.[8]) Thus recalibration is a content-dependent problem.

In other words, if a calibration curve is to reveal overconfidence or underconfidence on the part of the forecaster, then (on pain of incoherence) the "test" items must be logically independent from the point of view of the forecaster. Testing meteorological forecasts of "rain" for successive days seems a reasonable try at logical independence. But a battery

[7]Typically, Bayesians understand conditionalization in a dynamic, temporal sense—where K' is a body of knowledge held at time t' subsequent to time t. This construal makes conditionalization into a norm, regulating changes in degrees-of-belief as learning occurs. Also, however, there is an atemporal sense to conditionalization (see Levi 1980, chap. 10) Atemporally, conditionalization is the commitment at t for relating the current belief state K to hypothetical belief states K' that can arise by accepting new evidence.

The difference between these two senses of conditionalization is important to a proper understanding of what is settled by the Dutch Book argument—which purports to establish subjective, expected utility theory merely from (weak) assumptions about rational self-interest. For instance, Shimony's 1955 account, where conditional wagers are modeled by "called-off bets," involves just the atemporal sense, not the dynamic sense, of conditionalization. Thus, Dutch Book does not preclude changing your mind although you leave unchanged your corpus of knowledge, K.

[8]The proof is straightforward. Let $\tau[P(\cdot)]$ be some recalibration of the coherent (fine) probability $P(\cdot)$. Finite additivity assures the linearity of τ for, $\tau[P(g \text{ or } h)] = \tau[P(g)] + \tau[P(h)]$ when g and h are contraries. But coherence requires probability 1 is a fixed point under τ for, $\tau[P(t)] = P(t) = 1$ for each tautology t. Hence, τ is the identity function.

As an aside, this nice result constrains "counter-inductive" policies merely through coherence.

of relational questions with common terms and, say, a transitive relation obviously will not do, e.g., "Is the Niger longer than the Nile?" "Is the Niger longer than the Danube?" etc., fails. Nor will independence be achieved when the questioned items are "atomic," but a simple background generalization links them together biconditionally. Consider the pair of questions: "Is grey tin a good electrical conductor?" "Is white tin?" If the forecaster holds the (false) generalization that electrical properties of elements do not depend upon allotropic form, these two questions are equivalent.

Second, there is no reason for you to accept calibration of forecasts as a norm if the sequence of test questions involves events you deem probabilistically dependent. (See Kadane and Lichtenstein 1982.) For example, asked for your current predictions about the outcomes of 100 repeated flips of a coin that you judge to be either two-headed or two-tailed, each with probability .5, your opinion is the constant $P(\text{heads on flip}_i) = .5$ $(i = 1, \ldots, 100)$. However, you are certain to be maximally miscalibrated on these 100 forecasts as the outcome is either a string of all heads or all tails; there is an observed relative frequency of 100 percent or 0 percent when you have predicted outcomes with a (constant) probability of .5.

Both these problems arise whether calibration is taken in the short- or long-range sense. Fortunately, there is a common solution. Provide the forecaster with feedback between successive predictions, informing him if the previous forecast is accurate prior to eliciting the next forecast. Then, on condition the new evidence (the feedback) is consistent with background information, the sequence of predictions constitutes an *absolutely fair* process (as explained in the Appendix). This is also sufficient to avoid the problem of logical dependence.

Daily weather predictions of "rain" constitute a simple illustration. The meteorologist observes Monday's weather before Tuesday's forecast is given. Allowing that the observed weather pattern does not refute accepted theory, the forecaster faces a sequence of predictions (with feedback) where it is possible to recalibrate by transforming predictions in light of a calibration curve. However, the forecaster now is in a position to calculate his own calibration curve. Thus we find ourselves in the pradoxical circumstances discussed at the close of the previous section. The forecaster needs no assistance to recalibrate; self-help will do all that an outsider can.

3. Long-range Calibration and Subjectivism. Suppose our weatherman is a subjectivist. His daily forecasts, an "X percent chance of rain for tomorrow," reflect the current meteorological conditions—including the feedback of today's precipitation rate. What is the relation between

his forecast probabilities and calibration, in the long run? As Pratt noted twenty years ago (see note 12) and as was recently rediscovered by A. P. Dawid (1982), variations on the strong law of large numbers (for martingales) apply. In effect, the informal argument of section 1 (that a forecaster can recalibrate himself merely by attending to his past performance) can be made rigorous. Moreover, the three core postulates of subjectivism are sufficient to assure long-range calibration, almost surely. (Readers may skip these details and proceed to the discussion following Theorem 3.1, p. 283).

In detail, we have the following. Our weatherman contemplates making a sequence of forecasts regarding the events e_i ($i = 1,\ldots$), "rain" on day$_i$. He knows he will receive feedback information f_i subsequent to the ith forecast and prior to the $i+1$st. The information f_i (together with background knowledge) is sufficient to determine the truth or falsity of the ith prediction, e_i. In other words, f_i must (in context) be at least informative as the binary indicator:

$$I_i = \begin{cases} 1, \text{ if } e_i \\ 0, \text{ otherwise.} \end{cases}$$

The upshot is that: $P(e_1) = p_1$ is the first probabilistic forecast (for "rain" on day$_1$); $P(e_2|f_1) = p_2$ is the second such prediction; and generally $P(e_i|f_1,\ldots,f_{i-1}) = p_i$ is the subjectivist model for the ith day's forecast. Of course, the magnitude of p_i depends upon all the evidence $(f_1,\ldots,f_{i-1})$, in general. But the predictions are independent, when relativized to their (respective) background conditions.[9]

To test these predictions for miscalibration, consider an arbitrary, infinite subsequence identified by the sequence ξ_i ($i = 1,\ldots$), where each ξ_i is either 0 or 1. That is, we are to include the ith forecast in the test subsequence just in case $\xi_i = 1$.[10] Following Dawid (1982, p. 606), define:

$$\nu_k = \sum_{i=1}^{k} \xi_i \qquad \text{(i)}$$

$$rf_k = \nu_k^{-1} \cdot \sum_{i=1}^{k} \xi_i I_i \qquad \text{(ii)}$$

[9]Note that p_i is asserted under background conditions logically weaker than those for the jth forecast ($j > i$). Technically, the claim of independence involves the sequence of (*absolutely fair*) random variables $\{X_i\}$, $X_i = (p_i - I_i)$.

[10]The value ξ_i need be determined (measurable) only after $(f_1,\ldots,f_{i-1})$ are specified, i.e., after the $i-1$st feedback. The sequence of ξ's is analogous to a place selection function used to determine random subsequences in a von Mises collective. (See Spielman 1976, for related results.) Last, we assume infinitely many ξ_i are positive, in the spirit of long-range calibration.

and

$$\bar{p}_k = \nu_k^{-1} \cdot \sum_{i=1}^{k} \xi_i p_i. \qquad \text{(iii)}$$

In words, (i) counts the number of forecasts chosen from among the first k to form the test subsequence; (ii) is the relative frequency with which events e_i occur, relativized to those tested (by the subsequence) among the initial k predictions. Last, (iii) gives the average forecast probability within the tested subsequence, over the first k predictions.

THEOREM 3.1: Assuming $\xi_i = 1$ infinitely often, with probability 1, $\lim_{k\to\infty} (rf_k - \bar{p}_k) = 0$.

Proof: See the Appendix.

The theorem asserts calibration, almost surely. For example, if the meteorologist uses probabilities from the interval $[r - \epsilon, r + \epsilon]$ infinitely often in his forecasts, then he can be almost sure that on this (sub)sequence of his predictions the observed relative frequency of "rainy" days is between $r - \epsilon$ and $r + \epsilon$.[11]

In his classic work *The Foundations of Statistics*, L. J. Savage gives the oft-cited result (Savage 1954, sec. 3.6) that nonextreme coherent opinions converge (almost surely) to the truth with enough shared evidence (of the appropriate sort). Like that well-known consequence of subjectivism, this theorem of Pratt and Dawid displays yet another asymptotic property of coherent beliefs (subject to accumulating evidence).[12] If long-range calibration is thought to signal realism, as an objective validation of subjective beliefs, then mere coherence is sufficient to establish (almost certain) agreement with the facts, in the long run. The agreement is "internal" (in Putnam's 1981 phrase), as the result holds—the asymptotic calibration is judged—from the point of view of the forecaster. The increasing body of *consistent* evidence must be accepted by the forecaster. What he is prepared to admit as evidence (as feedback) on day$_i$ depends upon his background assumptions at that stage; similarly for Sav-

[11]To construct this subsequence, set $\xi_i = 1$ just in case $r - \epsilon \leq p_i \leq r + \epsilon$. This inequality is determined to hold or fail after the i-1st feedback, in accord with the measurability conditions.

[12]We obtain Dawid's (1982) result from Theorem 3.1 by restricting feedback (f_i) to the indicator I_i, i.e., the weatherman learns only if day$_i$ was "rainy" or not. To obtain Pratt's result, let the feedback on day i (prior to forecast p_{i+1}) be the (real-valued) amount of rainfall on day$_i$, and require each forecast to fix, e.g., an upper 90 percent estimate of rainfall for the next day. Then $p_i = .9$ ($i = 1,\ldots$) and the weatherman's prediction amounts to a lower (probability = .9) interval-estimate of rainfall. (J. W. Pratt [1962], "Must subjective probabilities be realized as relative frequencies?" Manuscript.)

age's "convergence of opinion." What is lost by this constraint? There is no privileged epistemological vantage point free of this modicum of "relativism."

4. Calibration in the Short Run and Scoring Rules. If calibration in the long range is of no concern to a subjectivist (theorem 3.1) what then of calibration in the short run? Is there some useful criterion of short-range calibration that demarcates the expert from the duffer? Can some index of calibration be used to identify the best weather forecasts over the next year's set of daily predictions? In general, a short-run scoring rule for gauging calibration must compare observed calibration after the fixed period (the calibration curve) with the well-calibrated forecaster and compute a "distance" between the two. (See Rao 1980 for a useful discussion of "distance.")

The idea of a scoring rule for measuring adequacy of one's predictions is hardly news to a subjectivist. The standard defense of the subjective interpretation of probability as personal betting rates, the "Dutch Book" argument, trades on this theme. Following Finetti (1974, p. 87) consider the procedure which requires you to choose a value $\bar{x}$ (called a *prevision*) for each real-valued random quantity X, with the understanding that you are obliged to accept any gamble of the form (with payoff):

$$c(X - \bar{x});$$

where c is some real number selected at the discretion of an "opponent."

In the case where X is the indicator for some event e ($X = 1$ if e, $X = 0$ otherwise), you are obliged to provide a (familiar) betting rate r ($=\bar{x}$) on the condition that you are prepared to:

win the amount $S - rS$ if e occurs

and

lose the amount rS if e fails to occur,

with $c = S > 0$. That is, you are asked for a betting rate on e at which you are indifferent between betting on/against e.[13]

DEFINITION: Your previsions are *coherent* if there is no finite selection of gambles ($c \neq 0$) that ensures you a (uniformly negative) loss. Otherwise, your previsions are *incoherent*.

Finetti's "Dutch Book" theorem establishes that the rates for bets on events satisfy the axioms of (finitely additive) probability just in case they are

[13]For a good discussion of why the supposition of a fair betting rate on each event is excessive in the name of rationality, see Kyburg (1978) and Levi (1980, chap. 4).

coherent.[14] Finetti argues that, in general, coherence of previsions is equivalent to their being finitely additive linear functionals meeting the added constraint that *infemum* $X \leqq \bar{x} \leqq$ *supremum* X. However, this claim requires additional limitations on the structure of payoffs (see Seidenfeld and Schervish 1983).

The distinction between coherent and incoherent previsions yields a binary classification into those which are decision-theoretically admissible and those uniformly inadmissible against the simple alternative: no-bet ($c = 0$, for each c), based on a finite selection of gambles. However, from the standpoint of (short-range) calibration, admissibility alone is too liberal an account of rationality, as coherence fails to distinguish well-calibrated from poorly calibrated forecasters. Nor does success in gambling serve as a useful measure of (short-range) calibration.[15]

Gambles are not alone among scoring rules that serve to demarcate coherent previsions from incoherent ones.

DEFINITION: Call a scoring rule for previsions a *proper* scoring rule if it is (always) in the forecaster's interest to announce his true prevision. Under a proper scoring rule, a forecaster maximizes his expected score (or minimizes it in case of penalty scores) by reporting his honest opinions.

Thus, a scoring rule for (previsions on) events is proper just in case it evokes the gambler's fair betting rates.

Gambles, as defined above ($c[X - x]$) are proper, assuming the units for wagers are linear in utility. An agent who mimics a set of (coherent) previsions different from his own faces a selection of "c's" (by an opponent) that results in an expected loss. (Of course, if he mimics an incoherent set of previsions, he exposes himself to a sure loss, by "Dutch Book.") This possibility is ruled out exactly when he reports his honest previsions, when all gambles strike him as fair, and there is no expected loss or gain from a (finite) selection of wagers.

[14]The extension to conditional bets is achieved (Shimony 1955; Finetti 1974, p. 135) with the use of *called-off* bets. To fix a conditional prevision for X, given f, use an unconditional wager with payoff: $I_f \cdot c(X - \bar{x})$, where I_f is the indicator variable for event f. Then coherence across previsions, including conditional previsions, yields the multiplication theorem for conditional probabilities.

For a helpful discussion of the limitations in this approach regarding coherence of conditional previsions given events of probability 0, see Levi 1980, sec. 5.6.

[15]For example, consider two gamblers who share the same coherent belief states: they are epistemic twins. One will wager against the other on a finite sequence of mutually fair bets, at their common betting rates. But the one shows a net profit just in case the other shows a net loss. Unless they choose sides ($c \gtrless 0$) as a function of their odds (player$_1$ always takes the long odds), their individual gains and losses bear no correlation to their common calibration.

As Finetti notes (1974, p. 93), gambles have an operational awkwardness when used as scoring rules since they require (among other considerations) an "opponent" whose behavior in picking out what to gamble on might, e.g., influence the bettor's opinions that are to be elicited. An alternative proper scoring rule is to penalize the forecaster by an amount

$$L = \left(\frac{X - \bar{x}}{k}\right)^2, \qquad (*)$$

for some known constant $k \neq 0$.[16] Any linear transformation of a proper scoring rule is, again, a proper scoring rule. Thus, we can offer a small prize, sufficient to compensate for penalty and induce the agent's cooperation.

The scoring rule (*) allows a demarcation between coherent and incoherent previsions by the criterion of admissibility; that is, as was the case with gambles, this scoring rule allows us to declare a set of previsions incoherent if for some finite set of random variables there is an alternative collection of previsions with (uniformly) smaller losses under (*) (regardless of the values of these random variables).[17]

The scoring rule (*) has several interesting characteristic features. In his seminal paper, "Elicitation of Personal Probabilities and Expectations," Savage (1971) shows that the quadratic loss L is unique among proper scores for either condition: (*i*) L is symmetric in honest and reported previsions. That is, the forecaster's expected loss when r is announced and s is his prevision equals his expected loss when s is reported and r is his prevision. (ii) L depends upon $(r - s)$ only. Also, the quadratic loss (*) is the algebraically simplest proper score. For example, the alternative loss function:

$$L' = |X - \bar{x}| \qquad (**)$$

leads to announced forecasts of 0–1 "previsions" for events. When scores are calculated according to (**) for forecasts on events X_e, the forecaster minimizes his expected L'-loss by the simple rule:

announce $\bar{x}_e = 1$ if $P(e) > .5$

announce $\bar{x}_e = 0$ if $P(e) < .5$

and announce an arbitrary value for $\bar{x}_e$ if $P(e) = .5$.

[16]Using the proper score (*) in place of gambles does not avoid the operational worry that selection of random variables for scoring might influence the bettor's previsions. Since all scoring rules must be restricted to a proper subset of the algebra when applied, I see no way round this problem.

[17]To verify that L is proper with respect to previsions for events, note that a forecaster minimizes his expected loss by reporting $\bar{x} = r = P(e)$, for $X = I_e$.

In general, L' is minimized by reporting the median of one's personal distribution for X, instead of the mean of that distribution (which is what a proper score elicits).[18]

For purposes of differentiating among rival coherent forecasters, proper scores such as L have an advantage over gambles. Gambles elicit previsions that make each one fair—and so too is the sum of finitely many fair when payoffs are linear in utility. That is not the case with quadratic loss, as G. W. Brier noticed some thirty-three years ago (1950).

Brier set himself the task of constructing a proper scoring rule for verifying weather forecasts. On each of N occasions one of R possible (exclusive and exhaustive) meteorological events transpires. Thus, the weatherman gives daily predictions over an R-fold partition of meteorological states, for N days. Let I_{ij} be the indicator for the jth (of R) event(s) on the ith (of N) occasion(s). Let p_{ij} be the weatherman's probabilistic forecast for I_{ij}, i.e., the announced probability for event$_j$ on day$_i$. Brier proposed the following score:

$$\text{(Brier score) } B = (1/N) \sum_{j=1}^{R} \sum_{i=1}^{N} (p_{ij} - I_{ij})^2.$$

Brier-score is merely Finetti's quadratic loss L, with $k = \sqrt{N}$, summed over the RN simple predictions, and scaled with a range of (0,2) (assuming the forecaster is coherent).

A perfect score ($B = 0$) is achieved by using 0–1 forecasts exclusively and committing no errors. A worst score ($B = 2$) is achieved again by using 0–1 forecasts only and getting none of the probability-1 predictions correct. Each coherent forecaster expects a Brier-score less than 1 ($= \lim_{n \to \infty} [1 - r^{-1}]$). A forecaster with a uniform probability over the RN simple events ($p_{ij} = 1/R$) earns the constant score $(R - 1)/R$. Thus, both the expected score and the range of possible scores depend upon the forecaster's distribution: the full [0,2] range is uniquely associated with the distribution of lowest, i.e., best, expected score—the two-point 0–1 distribution.

Let us contrast Brier-score with a short-range calibration scoring rule closely related to it. As identified by Bross (1953), first tied to Brier-score by Sanders (1958), and later investigated by Murphy (see his 1973a,

[18]In his discussion of the value of evidence, P. Horwich proposes L' as an index of *error* of a distribution (1982, p. 127). Unfortunately, his discussion does not take note of the distinction between proper and improper scores. Since the decision-theoretic conclusion Horwich seeks depends upon convexity of the payoff function, a property shared by all proper scores (see Savage 1971, p. 575), there is no need to think of one score as *the* error function—any proper score supports the decision-theoretic claim that cost-free evidence is (weakly) desirable. (For additional results see Lindley [1981].)

1973b, 1974; and Murphy and Epstein 1967; Murphy and Winkler 1977, for a selection), we have the following index of miscalibration. To repeat, the weatherman gives daily forecasts over an R-fold partition of meteorological events. Thus, each of his predictions is a vector

$$\vec{p}_i = (p_{i1},\ldots,p_{iR})\ (i = 1,\ldots,N).$$

Coherence entails:

$$\sum_{j=1}^{R} p_{ij} = 1.$$

Of the N vector-valued predictions $\vec{p}_i$ there are T distinct ones ($T \leq N$), identified as $\vec{p}^{\,t}$ ($t = 1,\ldots,T$), with n_t instances of forecast $\vec{p}^{\,t}$, where

$$\sum_{t=1}^{T} n_t = N.$$

Our concern with calibration requires a contrast between each $\vec{p}^{\,t}$ vector-valued forecast and the corresponding vector of observed relative frequencies: $\vec{rf}^{\,t} = rf^t_1,\ldots,rf^t_R)$, where rf^t_1 is the relative frequency of event$_j$ on those days when forecast $\vec{p}^{\,t}$ is announced. Weighting the squared-difference between $\vec{p}^{\,t}$ and $\vec{rf}^{\,t}$ by the number of occasions when $\vec{p}^{\,t}$ is used (n_t), we arrive at the Brier-calibration score for N R-fold forecasts:

$$\text{Brier-calibration} = (1/N)\sum_{t=1}^{T} n_t(\vec{p}_t - \vec{rf}^{\,t})(\vec{p}^{\,t} - \vec{rf}^{\,t})'.$$

As shown by Murphy (1973b), the Brier-score (B) can be decomposed into a sum of three indices, one of which is Brier-calibration.[19]

Can Brier-calibration be used to gauge short-range reliability of forecasts? Unfortunately, whereas Brier-score is proper, this index of cali-

[19]The remaining two summands (in Murphy's terminology) are: (i) the "uncertainty" of the events, and (ii) the negative of the "resolution" of the predictions.

The "uncertainty" is the Brier-score (B) for the vector of observed frequencies. That is, let $\vec{F} = (rf_1,\ldots,rf_R)$ be the vector of overall relative frequencies for the R events over the N days. Then the "uncertainty" is the Brier-score for the repeated forecast $\vec{F}$. This is the minimum Brier-score that can be achieved by repetition of a single prediction vector (for the given N days). Also, it corresponds to the inverse of the Fisher information in the R-cell multinomial distribution with multinomial weights $(rf_1,\ldots,rf_R)$.

The "resolution" of the forecasts is an index of the "distance" between the vector of observed relative frequencies ($\vec{F}$) and the vectors $\vec{rf}^{\,t}$. Specifically, "resolution" is the quantity:

$$(1/N)\sum_{t=1}^{T} n_t(\vec{F} - \vec{rf}^{\,t})(\vec{F} - \vec{rf}^{\,t})'.$$

Simply put, "resolution" is an indicator of the forecaster's ability to discriminate different relative frequencies according to his forecasts, regardless of what those forecasts are.

bration is not! Suppose a weatherman is to make daily forecasts for the Chicago vicinity and knows he is to be scored for performance on a calibration index of reliability after one year. For simplicity, imagine he is required to provide the binary vector prediction for ("rain," "no rain") ($R = 2$, $N = 365$). To perform well on the test for calibration, a good strategy is to take advantage of the practical certainty that the overall rate for precipitation in Chicago over the next 365 days will be very nearly the yearly average, approximately 20 percent. That is, regardless what the weatherman thinks about the day to day weather, a simple strategy—announce a "20 percent chance of precipitation"—is excellent for a superb calibration score. Of course, this strategy will be expected to fare noticeably worse on the Brier-score (with about an index of .16) than what is usual for Chicago weather forecasters (who score about .13—see Murphy and Winkler 1977, p. 44).[20]

To verify that Brier-calibration is an improper score, it suffices to give one context in which the forecaster's expected calibration score is improved by distorting his forecast. The opportunity to improve one's calibration score by misreporting is evident at the end of the fixed (365 day) forecast period. Even if the weatherman has been honest in his predictions for the first eleven months, he can do better not continuing with this policy for all twelve months. For example, suppose that (after eleven months) he notes that it has rained only 47 percent of those days for which he has announced a "50 percent chance of precipitation." The very next day that he is sure it will rain is a day he should announce a "50 percent chance of precipitation," thereby expecting to improve his overall calibration.[21] (See, also, De Groot's comments to Lindley *et al.* (1979, pp. 172–73).)

The incentive for distortion grows worse if we coarsen the calibration index by projecting the vector-valued forecast onto individual components (what Murphy and Epstein call "bias in the large" [1967]). In other words, if we calculate calibration of predictions for "rain" separately from the

[20]In terms of the decomposition discussed in footnote 19, the "one-note" forecast does well on the calibration component to Brier-score; it gets a zero on "resolution" and loses all its Brier-points on the fixed "uncertainty" component. The usual weather forecasts lose some Brier-points on calibration but more than make them up with their impressive "resolution."

[21]It is more difficult to verify that the "one note" forecast is better than honest reporting. However, a short sequence of forecasts for a beta-distribution, e.g., binomial sampling with a "uniform" prior and $N \leqq 7$, is one case where mere repetition of the "prior" forecast ($p_1 = \ldots = p_7 = .5$) has a better expected score than does honest reporting. Still more difficult is calculation of the optimal (sequential) strategy for forecasting. Even the elementary problem of forecasting from a beta-distribution involves rather complicated dynamic programming for its solution. In short sequences ($N \leq 7$) I conjecture that the "one note" strategy—repeat the "prior"—is optimal.

calibration of predictions for "no rain," then it may be strategic for the weatherman to report *incoherent* forecasts. For instance, imagine that after eleven months the weatherman's calibration curve shows overconfidence on his 50 percent forecasts for "rain," but an underconfidence on his 70 percent forecasts for "no rain," then he improves his calibration for both sorts of forecasts by reporting the (incoherent) "50 percent chance rain, 70 percent chance no rain" on the next occasion he is sure that it will "rain."

It is clear that, from a subjectivist point of view, short-range calibration is worse than useless as an index of reliable forecasting.

5. Conclusions. What have we learned about calibration? In general:

1. A coherent forecaster cannot recalibrate over an algebra of propositions, on pain of incoherence.
2. A coherent forecaster need not accept calibration as a norm when the sequence of predictions to be evaluated by calibration is a sequence of probabilistically dependent events.

Moreover, even where these two constraints are respected, as when forecasts are updated with consistent feedback:

3. long-range calibration is almost certain for a subjectivist;
4. whereas short-range calibration is an improper scoring rule.

Calibration in the long run is otiose, and in the short run is an inducement to hedge announced forecasts (or an inducement to feign a violation of the total evidence requirement by appeal to frequencies in a broad "reference class" of similar predictions). What is left of the intuition that observed frequencies provide a standard for assessing expertise in (probabilistic) forecasting? Of course, nothing shown here conflicts with the sound intuition that my reaction to your forecast is contingent upon what I know of your past performance. But, the desire for a universal, objective index of accuracy of probability appraisers seems a will-o'-the-wisp. The lesson to be learned from subjectivism is that rationality is a loose-fitting garment. When it comes to appraising forecast skill, why need we agree on the merits of a weatherman given the shared evidence of his calibration for the past N predictions? Subjectivism makes no parallel requirement for other hypotheses of mutual interest. Nor does the recent literature on "consensus" hint at the plausibility of such an assumption (see French forthcoming and Lindley forthcoming).

Our discussion of proper scoring rules suggests we might profitably shift our concern. Instead of seeking a universal index of reliable forecasting, we might consider whether there is some context-dependent procedure for ensuring that forecasts are reliably reported. In principle, all

proper scoring rules elicit the same forecast distribution from a given forecaster. But in practice (forecasts are "rounded," etc.) different proper scores are useful in focusing the forecaster on distinct aspects of his forecast distribution.

A decision maker is in a position to identify which aspects of his own personal probability are important to the decision at hand. In other words, the decision is sensitive to particular aspects of the decision maker's probability, e.g., the "tails" of his personal probability. Suppose the decision maker is explicit about how he plans to use the forecaster's predictions. For instance, the decision maker might agree to use the forecaster's probability as his own in this decision. Then a judicious choice of proper scoring rule will induce the forecaster to take care to report his opinions honestly, and to be precise in just those aspects important to the decision maker.[22]

None of this serves to justify calibration as an index of reliability. Long-range calibration is empty. Short-range calibration is counterproductive. If the forecaster knows his opinions are gauged by calibration (in the short run), he has incentive to distort his opinions. This hamstrings the listener who, then, lacks an understanding of the responses thus evoked. Is it not wise to avoid the temptation to reward the "skill" of aligning short runs of frequencies and announced forecasts? If not, well-calibrated predictors may reveal much more skill than you suspect.

APPENDIX

The proof of Theorem 3.1 is a straightforward application of Feller's (1966) Theorem 7.8.2—basically the law of large numbers for martingales. (Dawid [1982] uses Theorem 7.9.3 of the 1971 edition of Feller's book.) The point in outlining this proof is to emphasize the difference between Pratt's (1962) result and what Dawid (1982) shows. The difference makes a difference when personal probability is finitely, but not countably, additive.

First is Feller's Theorem 7.8.2.

Let $\{X_n\}$ be a sequence of *absolutely fair* random variables, i.e., $E[X_n|X_1,\ldots,X_{n-1}] = 0 = E[X_1]$ $(n = 2,\ldots)$. Define $S_n = X_1 + \ldots + X_n$. If $b_1 < b_2 < \ldots \to \infty$ and $\Sigma\, b_k^{-2}E[X_k^2] < \infty$, then with probability 1,

$$b_n^{-1}S_n \to 0,$$

and the variables:

$$Y_n = \sum_{k=1}^{n} b_k^{-1}X_k$$

converge. □

[22]Unfortunately, the decision maker must be on guard for the forecaster who takes an interest in the decision beyond his reward under the score. Then the forecaster's interest contaminates the score. His forecast is a "vote" in the decision-making process. The important works of Gibbard (1973) and Zeckhauser (1973) provide unsettling results about the inevitability of manipulable voting schemes. What is the decision-maker to do to compensate for the forecaster's potential interests?

Recall that the forecaster makes a sequence of predictions for the events e_i ($i = 1,\ldots$), where there is feedback f_i between the ith and $i + 1$st prediction. The feedback is (with background information) logically sufficient for the indicator

$$I_i = \begin{cases} 1, \text{ if } e_i \\ 0, \text{ otherwise.} \end{cases}$$

Thus, $p_1 = P(e_1)$ is the first prediction. $p_2 = P(e_2|f_1)$ is the second prediction. In general, $p_n = P(e_n|f_1,\ldots,f_{n-1})$ is the nth such prediction.

Let Let $X_i = \xi_i(p_i - I_i)$. Then $E[X_1] = 0$ as $p_1 = P(e_1) = E[I_1]$ and ξ_1 is a constant (=0 or =1). To verify that $E[X_n|X_1,\ldots,X_{n-1}] = 0$, consider two cases.

Case 1 (for Dawid's result): the feedback f_i is limited to a finite set ($i = 1,\ldots$). Then $E[X_n|X_1,\ldots,X_{n-1}] = 0$ because $I_1, \ldots, I_{n-1}$, hence $X_1, \ldots, X_{n-1}$, are a function of $f_1, \ldots, f_{n-1}$ and these (feedback) constitute a *finite* partition (and ξ_n is measurable after f_{n-1}).

Case 2 (for Pratt's result): some (all) feedback has an infinite sample space. Then one must add an assumption of disintegrability in the margin of the feedback variable(s) with infinite sample space(s). For instance, if f_1 can assume one of the denumerably many values f_{1i} ($i = 1,\ldots$), then to establish that $E[X_2|X_1] = 0$ it is necessary to require that:

$$E[X_2|f_{1i}] = 0 \ (i = 1,\ldots) \text{ entails } E[X_2|X_1] = 0, \qquad (*)$$

where, as above, I_1 hence X_1 is a function of f_{1i}.

Last, let $b_i = \nu_i^{-1}$ if $\nu_i > 0$, and $b_i = 0$ otherwise. Then (see Dawid 1982, p. 609) it is easy to verify that

$$\sum_{i=1}^{\infty} b_i^{\ 2} E[X^2] \leq \pi^2/24 < \infty,$$

which completes the assumptions needed for Feller's Theorem 7.8.2.

In case the forecaster's personal probability is not countably additive, two limitations exist. First, the nonfinitary version of convergence, $b_n^{-1}S_n \to 0$, needs to be rewritten in its finitary form (see Dubins 1974). (For discussion of finitely additive strong-laws of convergence, see R. Chen 1977.) Second, in *Case* 2 the entailment (*) (equivalent to disintegrability, see Dubins 1975) cannot always hold since its adoption across all denumerable partitions is equivalent to countable additivity (see Schervish *et al.* 1984).

Last, subject to these restrictions, the result continues to hold when feedback is delayed by several predictions, so long as the lag-times are not increasing too rapidly.

REFERENCES

Alpert, M., and Raiffa, H. (1982), "A progress report on the training of probability assessors", in *Judgment under Uncertainty: Heuristics and Biases,* D. Kahneman, P. Slovic, and A. Tversky, (eds.). Cambridge: Cambridge University Press, pp. 294–305. Hereafter, "*Judgment under Uncertainty.*"

Blackwell, D. and Girshick, M. (1954), *Theory of Games and Statistical Decisions.* London and New York: John Wiley.

Brier, G. W. (1950), "Verification of Forecasts Expressed in Terms of Probability", *Monthly Weather Review 78*: 1–3.

Bross, I. D. J. (1953), *Design for Decision.* New York: Macmillan.

Chen, R. (1977), "On Almost Sure Convergence in a Finitely Additive Setting", *Z. Wahrscheinlichkeitstheorie 37*: 341–56.

Dawid, A. P. (1982), "The Well Calibrated Bayesian", *Journal of the American Statistical Association 77*: 605–10; discussion, 610–13.

De Groot, M., and Eriksson, E. (forthcoming), "Probability forecasting, stochastic dominance and the Lorenz curve", in *Proceedings of the Second International Meeting on Bayesian Statistics,* Valencia, Spain, 1983.

De Groot, M., and Fienberg, S. E. (1981), "Assessing Probability Assessors: Calibration and Refinement", *Technical Report 105*, Dept. of Statistics. Pittsburgh: Carnegie-Mellon University.

De Groot, M., and Fienberg, S. (1982), "The Comparison and Evaluation of Forecasters", *Technical Report 244*, Department of Statistics. Pittsburgh: Carnegie-Mellon University.

Dubins, L. (1974), "On Lebesgue-like Extensions of Finitely Additive Measures", *Annals of Probability 2*: 456–63.

——— (1975), "Finitely Additive Conditional Probabilities, Conglomerability and Disintegrations", *Annals of Probability 3*: 89–99.

Feller, W. (1966), *An Introduction to Probability Theory and its Applications*. Vol. 2. London and New York: John Wiley.

Finetti, B. de (1972), *Probability, Induction and Statistics*. London and New York: John Wiley.

———. (1974), *Theory of Probability*. Vol. 1. London and New York: John Wiley.

French, S. (forthcoming), "Group consensus probability distributions: a critical survey", in *Proceedings of the Second International Meeting on Bayesian Statistics*, Valencia, Spain, 1983.

Gibbard, A. (1973), "Manipulation of Voting Schemes: A General Result", *Econometrica 41*: 587–601.

Hoerl, A. E., and Fallin, H. K. (1974), "Reliability of Subjective Evaluations in a High Incentive Situation," *Journal of the Royal Statistical Society A 127*: 227–30.

Horwich, P. (1982), *Probability and Evidence*, Cambridge: Cambridge University Press.

Kadane, J. B., and Lichtenstein, S. (1982), "A Subjectivist View of Calibration", *Technical Report 233*, Dept. of Statistics. Pittsburgh: Carnegie-Mellon University.

Kyburg, H. E. (1974), *The Logical Foundations of Statistical Inference*. Dordrecht: D. Reidel.

——— (1978), "Subjective Probability: Considerations, Reflections, and Problems", *Journal Philosophical Logic 7*: 157–80.

Levi, I. (1980), *The Enterprise of Knowledge*, Cambridge: The MIT Press.

——— (1981), "Direct Inference and Confirmational Conditionalization", *Philosophy of Science 48*: 532–52.

Lichtenstein, S., and Fischhoff, B. (1977), "Do Those Who Know More also Know More about How Much They Know?" *Organizational Behavior and Human Performance 20*: 159–83.

Lichtenstein, S.; Fischhoff, B.; and Phillips, L. (1982), "Calibration of probabilities: The state of the art to 1980", in *Judgment under Uncertainty*, D. Kahneman, P. Slovic, and A. Tversky (eds.). Cambridge: Cambridge University Press, pp. 306–34.

Lindley, D. V. (1981), "Scoring rules and the Inevitability of Probability", unpublished report, ORC 81-1, Operations Research Center. Berkeley: University of California.

———. (forthcoming), "Reconciliation of discrete probability distributions", in *Proceedings of the Second International Meeting on Bayesian Statistics*, Valencia, Spain, 1983.

Lindley, D. V.; Tversky, A.; and Brown, R. V. (1979), "On the Reconcilliation of Probability Assessments", with discussion, *Journal of the Royal Statistical Society A 142*: 146–80.

Murphy, A. H. (1973a), "Hedging and Skill Scores for Probability Forecasts", *Journal of Applied Meteorology 12*: 215–23.

——— (1973b), "A New Vector Partition of the Probability Score", *Journal of Applied Meteorology 12*: 595–600.

———. (1974), "A Sample Skill Score for Probability Forecasts", *Monthly Weather Review 102*: 48–55.

Murphy, A. H., and Epstein, E. S. (1967), "Verification of Probabilistic Predictions: A Brief Review", *Journal of Applied Meteorology 6*: 748–55.

Murphy, A. H., and Winkler, R. L. (1977), "Reliability of Subjective Probability Forecasts of Precipitation and Temperature", *Applied Statistics 26*: 41–47.

Pratt, J., and Schlaifer, R. (forthcoming), "Repetitive assessment of judgmental proba-

bility distributions: a case study", in *Proceedings of the Second International Meeting on Bayesian Statistics,* Valencia, Spain, 1983.

Putnam, H. (1981), *Reason, Truth and History*. Cambridge: Cambridge University Press.

Rao, C. R. (1980), "Diversity and Dissimilarity Coefficients: A unified approach," *Technical Report 80-10,* Institute for Statistics and Applications, Dept. of Mathematics and Statistics, University of Pittsburgh.

Sanders, F. (1958), "The evaluation of subjective probability forecasts", Dept. of Meteorology, Contract AF 19(604)-1305, Scientific Report 5. Cambridge: MIT.

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

———. (1971), "Elicitation of Personal Probabilities and Expectations", *Journal of the American Statistical Association 66*: 783–801.

Schervish, M. J. (1983), "A General Method for Comparing Probability Assessors", *Technical Report 275,* Dept. of Statistics. Pittsburgh: Carnegie-Mellon University.

Schervish, M.: Seidenfeld, T.; and Kadane, J. (1984), "The Extent of Non-conglomerability of Finitely Additive Probabilities", *Z. Wahrscheinlichkeitstheorie 66*: 205–26.

Seidenfeld, T. (1978), "Direct Inference and Inverse Inference", *Journal of Philosophy 75*: 709–30.

Seidenfeld, T., and Schervish, M. (1983), "A Conflict Between Finite Additivity and Avoiding Dutch Book", *Philosophy of Science 50*: 398–412.

Shimony, A. (1955), "Coherence and the Axioms of Confirmation", *Journal of Symbolic Logic 20*: 1–28.

Spielman, S. (1976), "Exchangeability and the Certainty of Objective Randomness," *Journal of Philosophical Logic 5*: 399–406.

Winkler, R. L. (1967), "The Assessment of Prior Distributions in Bayesian Analysis", *Journal of the American Statistical Association 62*: 776–800.

Zeckhauser, R. (1973), "Voting Systems, Honest Preferences and Pareto Optimality", *American Political Science Review 67*: 934–46.