BOSTON STUDIES IN THE PHILOSOPHY OF SCIENCE

EDITED BY ROBERT S. COHEN AND MARX W. WARTOFSKY

VOLUME 76

PHYSICS, PHILOSOPHY AND PSYCHOANALYSIS

Essays in Honor of Adolf Grünbaum

Edited by

R. S. COHEN
Boston University

and

L. LAUDAN
Virginia Polytechnic Institute

D. REIDEL PUBLISHING COMPANY
A MEMBER OF THE KLUWER ACADEMIC PUBLISHERS GROUP
DORDRECHT / BOSTON / LANCASTER

BAS C. VAN FRAASSEN

CALIBRATION: A FREQUENCY JUSTIFICATION FOR PERSONAL PROBABILITY *

If a physical theory states that the probability of some event, under certain conditions, is thus or so, we naturally take that to be a statement of objective fact, descriptive of the way the world is. And we expect that fact, if it is indeed the case, to be reflected in frequencies of occurrence among the described events. What is called the frequency interpretation of probability intends something more: namely, that such a probabilistic theory is really *only* about actual frequencies of occurrence.[1]

But the language of probability has uncontestably another use as well: it serves to formulate and express our opinion and the extent of our avowed ignorance concerning matters of fact. This use invites the epithets 'subjective' or 'personal' because it is keyed to the state of the user. When I say that it seems likely to me that it will rain today, or that rain seems as likely as (more likely than, twice as likely as) not, I express my very own opinion and judgment, I express some aspect of my own expectations for today. Any satisfactory view about probability must explicate this second use as well.

Here adherents to the frequency interpretation have fared very badly. And adherents of subjectivist or Bayesian views have done very well, on two counts. First, they have made an effort to show that within their own framework they can recapitulate the explanatory and explicatory successes of their objectivist rivals. Secondly, they have demonstrated that observance of the probability calculus in the expression of personal opinion or degree of belief is required, on their interpretation, by very minimal criteria of rationality ('coherence'). The paradigm example of the first is de Finetti's theorem in his 'Foresight: Its Logical Laws, Its Subjective Sources'; of the second, the well-known Dutch Book Theorem.[2] And finally, there appears to be a consensus in the literature that frequentists have never succeeded in meeting the major criticisms of their views as applied to this second use of probability language.

In this paper I shall attempt to redress the balance somewhat. I shall outline how the use of probability language to express personal opinion about a single event can be understood in a way that avoids the major problems with which frequentists have struggled. And I shall attempt to demonstrate

R. S. Cohen and L. Laudan (eds.), Physics, Philosophy and Psychoanalysis, 295–319.

that observance of the probability calculus in such expression of opinion is equivalent to satisfaction of a basic frequentist criterion of rationality (*frequency coherence*). Based on the idea of *scoring*, also a subject investigated by de Finetti and other Bayesians, this will be a frequency analogue of the Dutch Book Theorem.

1. THE PHENOMENON: PERSONAL PROBABILITY JUDGMENTS

As a form of speech, expressions of personal opinion are often easy to recognize. "It seems likely to me that it will rain" can not be equated with any precise probability evaluation, but "likely" is here surely synonymous with "very probable." And "He is twice as likely to win the race as is his brother" is a very exact statement of odds, which we equate in turn with a probability ratio. When the weather forecast on the radio says, finally, that the chance of precipitation equals 0.6, that sounds at once very precise and very objective, but it is an announcement of the metereologist's professional opinion, reached after conscientious consultation of the data. To say that the opinion is professional, does not even imply that all the professional colleagues he respects would have to reach the same estimate when given the same data, though it does imply a large measure of agreement among them.

How shall we understand this activity? We can perceive it in two ways, not perhaps mutually exclusive: as *expressing* attitudes or as *asserting* autobiographical facts. To bring out the difference, think of the somewhat parallel case of promising. Yesterday I said, "I promise to give you a horse." But I did not give you anything, and today you accuse me of the heinous immorality of breaking a promise. No, I reply, I am not guilty of that at all, but only of the much lesser offense of lying. All that happened was that yesterday I stated falsely that I was promising to give you a horse.

It is easy to see what is wrong with this story. In saying, "I promise . . .", I must (normally?) be taken to be doing something more than implying or stating an autobiographical fact. In just the same way, if I say, "It seems likely to me that . . . ", I may be implying or stating a fact about my own attitude or judgment; but I am first and foremost doing something else: expressing that attitude or judgment.

Attitudes, once expressed, are evaluated in two ways. The first question is one which it should, in principle, be possible to answer right away: is this attitude *reasonable*? The second concerns the future: is this attitude *vindicated*? Again an imperfect parallel may help: a practical decision to devote the evening to attending a certain play. Was this decision a reasonable

one? That depends on the reviews you have read, the amount of money and time you have, the time and the alternatives contemplated at that time. Was it vindicated? That depends on factors not settled for you at the time of decision: how good the performance turns out to be, how much pleasure or insight you gained from it, and also on what else happened that evening that you missed or could have prevented or influenced if you had not gone to the play.

A morass in which frequentists have often sunk is their search for objective criteria of how reasonable a judgment is, in the light of available information. The most ambitious and most successful attempt along these lines is that of Kyburg. I will not say that he *is* stuck in a morass: his program of defining the right reference class and a recipe for determining the correct epistemic probabilities on the basis of available statistical information, may be successful. But we cannot yet say that it is. The Bayesian approach appears to eliminate this enterprise, and its problems, entirely. And still the subjectivist Bayesian is not silent on the question of reasonableness. How is that possible?

Looking again at the parallel of practical or moral decisions, we see one minimal criterion of reasonableness that connects it with vindication. A decision is unreasonable if vindication is *a priori* precluded. The Bayesian equates a probabilistic expression of opinion with an announcement of betting odds the person is willing to accept. Vindication consists clearly in gaining, or at least not losing, as a consequence of such bets. The Dutch Book Theorem says that such vindication is *a priori* precluded if and only if the probability calculus is violated. Thus the possibility of vindication is taken as a requirement of reasonableness.

This general insight and strategy are open to all contestants. Let the frequentist equate probabilistic expression of opinion with something else; and let him investigate the conditions under which such vindication is not *a priori* excluded.[3]

2. THE THEORY AND ITS PROBLEMS

The phenomenon to be addressed is the constant stream of judgment expressed in (vague) probabilistic language. A theory will propose models of what is going on, in which phenomena of this sort can fit. Because we, as philosophers, are interested in epistemology rather than psychology, we look to such theories only to find out two things: understanding of what this activity *could be*, and of the conditions under which this activity is *rational*. An answer to the first will suggest one to the second, for rationality consists largely in the suitability of chosen means to intended ends. What the activity

is should determine its criteria of success. We will evaluate its rationality by seeing whether its aim is pursued in an optimal fashion, first with respect to its own criteria of success and secondly in view of other aims of the larger projects of which it is part.

John Venn, in his *Logic of Chance*, was perhaps the first to formulate explicitly the frequency interpretation as an answer to the first question. The activity of judgment, expressed in such utterances as "It seems to me as likely as not that it will rain today," "It seems 95% probable to me that it will snow today" is assigned two main underlying factors. The first is a *selection of a reference class* – a classification of the subject – and the second an *estimate of relative frequency* in that class – in these examples, frequency of rain or of snow. This sketches the very simplest model of the activity which is suggested by the idea that probability talk is centrally and essentially concerned with frequencies. In the Appendix, I shall discuss this further.

The basic objections to this theory were already – and perhaps best – formulated by John Maynard Keynes in his *Treatise on Probability* (especially Ch. VIII, Sections 7–13). They take the form of three questions. The first is: how is the reference class selected? The second: how or where does the person obtain his estimates of frequencies? And the most important: why should personal probabilities, arrived at in this fashion, either obey, or be rationally required to obey, the probability calculus?

We may take the first two questions to be a request for elaboration of the theory. Can we construct models in which all probabilistic judgments, including those concerning statistical frequencies, appear as the outcome of such a process? And in such models, what is the exact mechanism of reference class selection, et cetera? It is noteworthy that the most extensive and sophisticated attempt to construct such models, namely that of Henry Kyburg's *Logical Foundations of Statistical Inference*, is also an attempt to do so in the most constrained manner possible. In contrast, John Venn explicitly allowed for an element of subjective choice and volition in the selection of reference classes, differing from occasion to occasion.[4]

These first two questions, however, do not strike me as going to the heart of the matter at all. Why should we ask Reichenbach, for instance, for a recipe for arriving at a judgment (in the light of our own background beliefs and information) about which horse will win this specific race, when we certainly have no right to ask Ramsey or de Finetti how to arrive at a specific bet on this particular occasion? A presupposition that Kyburg gives the appearance of accepting, and Venn apparently rejected,[4] is that the judgment

will have been arrived at in a rational manner, exactly if the input (background beliefs and information) determines via the dictates of rational deliberation, a uniquely right answer – the rationally compelled one. The alternative view, which I urge as the correct one, is that requirements of rationality can only go so far, and that what is rational is what stays within their bounds; thus allowing for an element of subjectivity and personal volition within rational choice. Rationality is only bridled irrationality.

The heart of the matter appears in Keynes' third question. Whether or not our judgments are reasonable should be determinable at the time we make them. But such underlying factors as statistical estimates and reference class selection are *hidden variables*, they do not belong to the surface phenomenon of judgment, at least in general, and are not (entirely) accessible to introspection either. (Consider the famous case of the chicken sexers, or any other sort of expertise in professional judgment where we speak of talent as well as of book learning.) The one paradigm rule of thumb for a preliminary evaluation of the reasonableness of judgment, which can indeed be applied at the time and without acceptance of any interpretation, is to see whether the axioms of probability are not violated. Let the frequentist either justify this rule or show why it should be rejected or restricted.

The frequentist cannot answer this challenge by pointing out that finite proportions in classes (or suitably chosen relative frequencies in sequences) obey those axioms. For the choice of reference class plays a crucial role as well. Suppose I am asked two questions about today: will there be any precipitation? Will there be any snow? And imagine that for the first question I consult the almanac, which says that here in Toronto approximately one in five days is marked by precipitation. The second question I answer after I have looked outside and taken account of the fact that today is a cold, overcast December day. Then I announce my probabilities: 1/5 for the first, 1/3 for the second. I chose different reference classes; now I have given a lower probability to the first proposition although it is entailed by the second, a violation of probability theory. Does it not seem that the frequentist must show why it is necessary to avoid this, and that he can only do it by formulating and defending intricate rules for the choice of reference classes?

But as I explained in the preceding section, there is a general strategy for answering this third question of Keynes. We can give a frequentist explication of the criteria of success for such judgments – *vindication*; then set down as a minimal requirement of rationality that the judgments not be such as to preclude *a priori* the very possibility of vindication; and finally, demonstrate that this requirement entails non-violation of the probability calculus.

Vindication I shall explicate in terms of *calibration*, a measure of how reliable one's judgments have been as indicators of actual frequencies. Possibility of vindication I shall then explicate as *potential calibration*. And the required demonstration will take the form of two adequacy theorems and the sketch of a third.

This is an alternative to the well-known strategy of laying down rules for the choice of reference classes. Quite apart from the morass of complexities which has beset *that* strategy, it leaves an obvious open question: why is it rational to follow *those* rules in selecting a reference class? Those rules also need justification, so they may still force us back to what I here propose: an analysis of the possibility of vindication, for the judgments which result. Hence I advocate the outlined alternative strategy.

3. VINDICATION: SCORING AND THE CALIBRATION LEMMA

After a metereologist announces, in the morning, a chance of 0.8 for rain, during the day it then rains or does not rain. In the first case, the meteorologist may look proud, but in the second he need not look ashamed – obviously what happens on that day does not make his forecast correct or incorrect. But he has announced these probabilities for a year – how good was his forecasting performance? The first problem to solve is that of devising measures to 'score' his performance. The second is to show, of some such measure, that it makes sense, that is, that it measures success with respect to the aim of his enterprise.

The first problem was given a solution, in 1950, which became generally accepted. Weather forecasters are evaluated by the *Brier Score*.[5] When given feedback on their cumulative Brier score, they also improve that score – which is lovely if it really measures their success, and regrettable if it does not. As analyzed afterward, the score actually combines two criteria. The first is *informativeness* or extremeness: the score tends to improve if the announced probabilities are closer to *one* and *zero*. The second is called *calibration*; its basic idea is that the forecasts fit the series of actual events perfectly, exactly if it rained on 60% of the days on which he said the probability of rain was 0.6, and so on for the other stated numerical values.

It is of course very typical to see this combination of two criteria, of just that sort. Of a traditional, non-statistical scientific theory, philosophers of every persuasion demand (in their various terminologies) both informativeness and truth, in certain respects. The two aims are in desperate tension, for the more informative we make our theories (the more audacious we are, the

bolder our conjectures) the less sure we can be that they are true, the greater the chance they will be false (in the intended respect).[6] Calibration plays here the conceptual role that truth, or empirical adequacy, plays in other contexts of discussion.

Now it will be clear that calibration is meant to be a measure of how reliable the forecasts are, cumulatively, as indicators of actual frequencies of occurrence. Just what the frequentist would pose intuitively as the aim of the forecaster's activity.[7] But is the basic idea that motivates the proposed measure a good one? Can we say, from our chosen point of view, that this clever idea of perfect calibration marks correct execution of the judgmental activity? If that chosen point of view is the frequency interpretation, certainly.

This we can establish by means of a simple demonstration. Suppose the forecaster acts exactly as frequentists describe. Each morning he classifies the day x as belonging to a reference class $\beta(x, \textit{rain})$. The classifications open to him here form a logical partition, that is, he has one and only one such reference class for any day with which he is presented. Suppose also that for each class Y of days that he ever uses as a reference class, he has an estimate $\alpha(\textit{rain} \mid Y)$ of the relative frequency of rain in Y. So on the morning of day x he announces the number $\alpha(\textit{rain} \mid \beta(x, \textit{rain}))$ as his probability for rain on that day.

Now he could fare badly, even if he correctly classifies each day (x belongs to $\beta(x, \textit{rain})$ in each case), and even if he has perfectly correct estimates of the frequency of rain in each of these classes. For the total set D of days with which he is presented may be an unrepresentative sample of days in general. That would be just plain unlucky for him. But if we assume that the world does not make him unlucky in that way, then the following little lemma shows that the correctness of his proportion estimates and reference class selection guarantees perfect calibration.

To state the lemma, let the correct proportions (equalling by assumption the estimated ones) be represented by an additive set function m defined on D. As usual define the conditionalization $m(A \mid B)$ as $m(A \cap B) \div m(B)$, where the denominator is not *zero*.

(3.1) CALIBRATION LEMMA. *If X is a finite partition of D, in the domain of m, and for each x in D the function P_x is defined by*

$$P_x(A) = m(A \mid B^x)$$

where B^x is the member of X to which x belongs, then

(3.2) $m(A \mid \{x : P_x(A) = r\}) = r$

wherever defined.

Note that if m is the proportion estimate, and X the set of reference classes for the question, then $P_x(A)$ is the forecast probability for A.

To prove this lemma, denote as $A(r)$ the set $\{x : P_x(A) = r\}$ of cases in which the announced probability of A was the number r. That set is also exactly the union of the sets B in X – the reference classes – for which the estimated proportion $m(A \mid B)$ equals r:

$$\begin{aligned} (3.3) \quad A(r) &= \cup \{B \in X : m(A \mid B) = r\} \\ &= B_1 \cup \ldots \cup B_k \text{ (say)}, \end{aligned}$$

a union of disjoint members of the partition. Hence:

$$\begin{aligned} (3.4) \quad m(A \cap A(r)) &= m(\cup \{A \cap B_i : i = 1, \ldots, k\}) \\ &= \sum_{i=1}^{k} m(B_i) m(A \mid B_i) \\ &= r \sum_{i=1}^{k} m(B_i) \\ &= rm(B_1 \cup \ldots \cup B_k) \\ &= rm(A(r)). \end{aligned}$$

Hence also $m(A \mid A(r)) = m(A \cap A(r)) \div m(A(r)) = r$, provided of course the denominator is not zero.

The argument extends at once to countable partitions if m is sigma-additive, but that seems a bit irrelevant for personal probabilities or rain forecasts. We may conclude in any case that the basic idea of perfect calibration is exactly the idea of complete correctness to be associated with the frequency interpretation: a selection of reference classes and estimate of proportions that happen to be exactly right for the presented sample.

4. REASONABLENESS: POTENTIAL APPROACH TO CALIBRATION

Let us now proceed slightly more abstractly: I am given a field or Boolean algebra F of attributes and a domain D of individuals, and asked to express a judgment concerning whether x has A, for various attributes A and various entities x in that domain. Let Q be a finite set of such propositions [x *has* A]

and let function P, defined on the whole family of these propositions, be used to represent my judgments. Call P a *scheme* for D and F. I shall assume that P assigns real numbers as 'grades' of personal probability.

The first notion we must define, for such a set Q, is the proportion of truths in it. But which are the true propositions? That depends on the state of the world, which must be represented too – by a model M. (The obvious form for such a model is a couple $\langle D, loc \rangle$ where D is the domain and *loc* some function that determines what attributes in F the members of D have, i.e., their 'location' in the possibility space determined by F. But that is a technical detail.) Each proposition is *true* or *false* in each model, and the Boolean operations on propositions cohere with the usual 'truth-table' assignments of truth values in a model. Denote by 'TRUE(M)' the set of propositions (for D and F) true in the model M. Then define the proportion of truths:

(4.1) $\%MQ = \#(\text{TRUE}(M) \cap Q) \div \#Q$

where # denotes the set's cardinality.

The next obvious step is to define the subset $Q(r)$ of propositions to which P assigns value r, and then to call P perfectly calibrated on Q exactly if $r = \%MQ(r)$ for each such assigned value. But because we are now dealing with questions that may relate to more than one attribute, that procedure is too rough and ready. Suppose for example I am asked about each of 100 days: will it rain? will it rain or snow?, will it rain or snow or hail? If the actual proportions were 0, 1/2, 6/10, and my announced probabilities were the same on each day, namely 3/10, 1/2, 3/10, the calibration would be perfect. (For Q contains 300 propositions, to 200 of which I have assigned 3/10; but of that 200, all of the first hundred (the ones of form "it rains on day x") are false while 60 of the remaining one hundred are true, and 60/200 equals 3/10.) This perfect calibration on the subsets $Q(r)$ hides the irrationality of assigning a lower value to one proposition than to a second which *implies* the first. But that irrationality would become readily apparent if we subdivided further in the obvious way, in terms of the attributes as well as the numbers assigned.

(4.2) $Q_P(A, r) = \{E \in Q : (\exists z)(E = [z \text{ has } A]) \text{ and } P(E) = r\}$.

When no confusion threatens, abbreviate $Q_P(A, r)$ to $Q(A, r)$. Then we call *P perfectly calibrated on Q with respect to model M* exactly if

(4.3) $r = \%MQ(A, r)$ for each value r and attribute A for which $Q(A, r)$ is not empty.

Such perfect calibration may, however, be precluded for trivial reasons.

If, for example, Q contains only a single member, then P must assign it either *zero* or *one* if (4.3) is to hold. Moreover, given that Q is finite, a look at (4.1) shows that P cannot be perfectly calibrated on Q at all unless P assigns only proportions of the finite number $\#Q$. Surely it cannot be a precondition of vindication that our personal probabilities come in rational fractions! Reflect especially on the fact that we do not generally know beforehand how many questions we shall be asked. Our first need here is for a measure of approximation, or distance from perfect calibration. The obvious measure to come to mind here (especially to readers of Brier's article) is the length of the vector $\langle r_i - \%MQ(A, r_i)\rangle$ where $r_1, \ldots, r_n$ are the numbers (in some order) which P assigns to members of Q. But because I shall be concerned with the measure only with respect to the *possibility* of its decrease toward zero, we can without loss of *finesse*, use a much cruder one.

(4.4) P is *calibrated to within* distance q, *on* set Q, *with respect to* model M, exactly if q is the supremum of the numbers $|r - \%MQ(A, r)|$ such that $Q(A, r)$ is not empty.

To be perfectly calibrated is then to be calibrated to within distance zero. Now this may be impossible to achieve, for stated reasons, even if Q is increased indefinitely. But with such increase we may hope for ever better approximation.

It is too early, though, to announce this hope as furnishing a criterion for reasonableness. For suppose that I first state my probability for rain as 1/6 and then you ask me about one thousand tosses of a fair die for the probability of ace and I say 1/6 each time. On the total set of 1001 questions, my personal probability will probably be quite well calibrated, but that reveals nothing about the reasonableness of my initial judgment about rain. To see the problem in acute form, let this first judgment be replaced by two: adding to it also the judgment that the probability of there being no rain equals 1/6 as well. Calibration on the total set of 1002 propositions will be quite good, whereas there is something drastically wrong with my probabilities for the first two.

So the possibility of ever better calibration which we require, must be on extensions of the initial set of propositions which are in a relevant sense

like the original ones. A frequentist would say that optimally, the additional questions raised should be about the same attributes for entities for which the person selects the same reference classes. That selection being a 'hidden variable' of his judgment, however, we must make do with a relation of likeness reflected entirely in the personal probability function P, i.e., in the actual expression of the judgments.

(4.5) Entities x and y are *P-alike* exactly if $P[x \textit{ has } A] = P[y \textit{ has } A]$ for each attribute A.

(4.6) Q' is a *P-alike extension of* Q if and only if $Q \subseteq Q'$, P is defined for every member of Q', and if $[z \textit{ has } A]$ is in Q' then there is an entity y such that y and z are P-alike, and for each attribute B, $[z \textit{ has } B]$ is in Q' if and only if $[y \textit{ has } B]$ is in Q.

Thus a typical P-alike extension of $Q = \{[y \textit{ has } A], [y \textit{ has } B]\}$ looks like $Q' = \{[y \textit{ has } A], [y \textit{ has } B], [z_1 \textit{ has } A], [z_1 \textit{ has } B], \ldots, [z_n \textit{ has } A], [z_n \textit{ has } B]\}$, where $z_1, \ldots, z_n$ are all P-alike to y. Having introduced the relevant notion of likeness, we can now define potential calibration in two steps.

(4.7) Let P be a scheme for D and F and P' a scheme for D' and F'. Then P' is an *extension* of P exactly if $D \subseteq D'$, $F \subseteq F'$ and $P[y \textit{ has } A] = P'[y \textit{ has } A]$ for each y in D and A in F.

(4.8) P is *potentially calibrated* on finite set Q of propositions on which it is defined exactly if for every real positive number q there exists an extension P' of P and P'-alike extension Q' of Q such that P' is calibrated to within q on Q', on some model.

As minimal criterion of rationality, from a frequentist point of view, I state the requirement that our body of judgments should be representable by at least one scheme which is potentially calibrated on every finite set of propositions for which it is defined. Note that since "calibrated to within q" has been defined with reference to proportions, and hence only for finite sets of propositions, the P-alike extensions that play a role in determining potential calibration on Q are also all finite.

5. FIRST ADEQUACY THEOREM

In this section and the next I shall address what seems at first sight to be a weaker criterion than I announced in the preceding paragraph:

(5.1) Scheme P is frequency coherent exactly if it is potentially calibrated on every finite set of propositions of form $Q = \{[x \textit{ has } A] : A \text{ in } FQ\}$ for which it is defined.

Note that here all members of Q are about the same, single subject. The scope and limits of this requirement will be discussed in Section 7.

(5.2) THEOREM. *If P is a scheme for D and F, and for each element x of D the function P_x defined by*

(5.2*) $P_x(A) = P[x \textit{ has } A]$

is a probability function on F, then P is frequency coherent.

To prove this theorem, we proceed in two stages. First of all, assuming P, D, F to be as described, consider the set $Q = \{[y \textit{ has } A_1], \ldots, [y \textit{ has } A_k]\}$. The attributes $A_1, \ldots, A_k$ generate a finite sub-algebra F^* of F. Let $B_1, \ldots, B_m$ be the atoms of F^*. Think of these atoms as boxes, and the other elements of F^* (which are finite joins of these atoms) as composite boxes. Place n_j items in box B_j, for $j = 1, \ldots, m$ with the total $n = n_1 + \ldots + n_m$. Select any positive number r you like; you can then choose those 'occupation numbers' for the boxes so that $n_j/n = P_x(B_j) \pm 1/r$. The reason is of course that the $P_x(B_j)$ are non-negative numbers that sum to *one*, by the hypothesis that P_x is a probability function. If now A is in F^*, say $A = B_1 \cup B_2 \cup B_3$, then $P_x(A)$ is determined by the additivity of P_x and the occupation number for A similarly:

$$(5.3) \quad \begin{aligned}(n_1 + n_2 + n_3)/n &= P_x(B_1) + P_x(B_2) + P_x(B_3) \pm 3/r \\ &= P_x(A) \pm 3/r\end{aligned}$$

In general, the divergence can be no more than m/r. And because the number m is fixed as the number of atoms in F^*, we can set m/r less than or equal to any pre-selected positive number q by appropriate choice of r. This reasoning establishes the unsurprising fact that a probability function on a finite domain can be arbitrarily closely approximated by proportion in an urn.

As second stage, we turn this demonstration into the construction of a model which shows the potential calibration of P on Q. To the original domain we add $n - 1$ new entities. We extend P to P' by setting P' equal to P where both are defined, and all the new entities P'-alike to y itself. Now

we take a model M in which the set consisting of y itself plus the new entities is distributed in proportions n_j among the atoms B_j of F^*. Finally we consider the calibration of P' on the larger set $Q' = \{[z \textit{ has } A_i] : i = 1, \ldots, k,$ and $z = y$ or z is a new entity$\}$ and find that P' is calibrated to within q on Q' with respect to model M. Hence we conclude, by generalizing on this construction, that P is potentially calibrated on Q itself.

6. SECOND ADEQUACY THEOREM

As converse to the first result, we find that obedience of the probability calculus is also a necessary condition for frequency coherence.

(6.1) THEOREM. *If P is a scheme for D and F, and is frequency coherent, then each function P_x defined by (5.2*), for x in D, is a probability function on F.*

The axioms of probability theory (for personal probability I consider only finitary constraints) are

(I) $0 = p(\Lambda) \leqslant p(A) \leqslant p(K) = 1$

(II) $p(A \cup B) + p(A \cap B) = p(A) + p(B)$

where Λ and K are the minimal and maximal elements of Boolean algebra F and $\cup, \cap$ its join and meet operations.

Assuming now that P, D, F are as described in the antecedent of the theorem, it is clear first of all that $P_y(\Lambda)$ must equal to zero. For $[y \textit{ has } \Lambda]$ is the impossible proposition, and so that proportion of truths in any subset of $Q' = \{[z_1 \textit{ has } \Lambda], \ldots, [z_n \textit{ has } \Lambda]\}$ equals zero. Hence no extension P' of P will be calibrated on Q' to within $q > 0$ unless $P'[z_i \textit{ has } \Lambda]$ has an absolute value which is less than or equal to q. Thus $P[y \textit{ has } \Lambda]$ must have an absolute value less than every positive number, if P is potentially calibrated on $\{[y \textit{ has } \Lambda]\}$. Similarly $P[y \textit{ has } K] = 1$ if P is potentially calibrated on $\{[y \textit{ has } K]\}$.

It is just as easy to see that $P[y \textit{ has } A]$ must be in the interval $[0, 1]$. For if the value assigned is a distance q outside that interval, then no extension P' of P can be calibrated to within less than q on any P'-alike extension of $\{[y \textit{ has } A]\}$ – simply because all the relevant proportions are within it. Finally, we consider the four member set:

(6.2) $Q = [y \textit{ has } A \cup B], [y \textit{ has } A \cap B], [y \textit{ has } A], [y \textit{ has } B]$

Let us suppose that we have a violation of Axiom II:

(6.3) $P_y(A \cup B) + P_y(A \cap B) = P_y(A) + P_y(B) + d$

where d may be either positive or negative. We now extend P to a scheme P' for a larger domain D' in which all new entities are P'-alike to y. And we consider the P'-alike extension Q' of Q in which the same propositions occur with not only y but also these new entities as subjects. Let us abbreviate:

$$\begin{aligned} a_2 &= P_y(A \cup B) & b_1 &= \%MQ'(A \cup B, a_1) \\ a_2 &= P_y(A \cap B) & b_2 &= \%MQ'(A \cap B, a_2) \\ a_3 &= P_y(A) & b_3 &= \%MQ'(A, a_3) \\ a_4 &= P_y(B) & b_4 &= \%MQ'(A, a_4) \end{aligned}$$

where M is some appropriate model.

Because the new entities are all P-alike to y, $Q'(A', r)$ will be empty for all cases not listed in the above table. Thus for example $Q'(A \cup B, a_1)$ is the set of all propositions of form $[z \textit{ has } A \cup B]$ in Q', there are, let us say, m of these (and hence Q' has 4 m members exactly) of which m_1 are true, in which case $b_1 = m_1/m$. If we similarly set $b_i = m_i/m$ for $i = 2, 3, 4$ then it is clear that $m_1 + m_2 = m_3 + m_4$, so $b_1 + b_2 = b_3 + b_4$. We have $a_1 + a_2 = a_3 + a_4 + d$ and $b_1 + b_2 = b_3 + b_4$, and therefore:

(6.4) $(a_1 - b_1) + (a_2 - b_2) - (a_3 - b_3) - (a_4 - b_4) = d$

from which we conclude

(6.5.) $|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + |a_4 - b_4| \geqslant d$

which means that P' is not calibrated on q' to within less than d.

This argument being general with respect to extensions P' of P, P'-alike extensions Q' of Q, and relevant models M, we conclude that calibration to within less than d is impossible for these extensions, and so P is not potentially calibrated on Q.

7. ADEQUACY OF THE FREQUENCY COHERENCE CONCEPT

We have now established that a scheme P is frequency coherent if and only if each of its relativizations P_x is a probability function. But the reader may now have doubts about the significance of the notions used. Frequency coherence, as defined, relates only to calibration on sets of propositions that are all about the same subject. What about more diverse sets? This initial doubt, at least, can be put to rest.

(7.1) THEOREM. *P is frequency coherent if and only if P is potentially calibrated on all finite sets of propositions on which it is defined.*

The proof, which I shall sketch, relies on the simple lemma:

(7.2) LEMMA. *If P is calibrated to within q on disjoint sets $Q_1, \ldots, Q_n$ then P is also calibrated to within q on their union.*

For suppose Q is the union of those disjoint sets $Q_1, \ldots, Q_n$. The $Q(A, r) = Q_1(A, r) \cup \ldots \cup Q_n(A, r)$. Hence the proportion of M-truths in $Q(A, r)$ can neither be higher than all the numbers $\%MQ_i(A, r)$ nor lower than all of them. Hence the distance between that proportion and r cannot be larger than the supremum of all the numbers $|r - \%MQ_i(r)|$.

In the models, as we have conceived them so far, the questions whether $[x \textit{has } A]$, $[y \textit{ has } B]$ are true are totally independent. Hence we will be able to carry out the construction utilized in the first and second adequacy theorem simultaneously for any finite set of entities $y_1, \ldots, y_n$ in D. With respect to the set

$$Q' = \{[y_1 \textit{ has } A_1^1], [y_1 \textit{ has } A_2^1], \ldots, [y_n \textit{ has } A_{k_n}^n]\},$$

we can then find an appropriate model M such that the relevant extension P' is calibrated to within q on each subset

$$Q_j' = \{[y_j \textit{ has } A_1^j], \ldots, [y_j \textit{ has } A_{k_j}^j]\}$$

and therefore also on their union, i.e., Q' itself, by the above lemma.

But perhaps this is 'stonewalling'. For the uneasiness may lie exactly in the idea that there may be connections or relations among the entities. In that case, the questions whether x has A and whether y has B are *not* independent. Especially logic-minded readers, who want to see probabilities attached to all propositions expressed in a first-order predicate language, will be inclined to feel that the discussion so far has ignored relations among entities in the domain.

When Tarski reduced the problem of truth to the definition of *satisfaction*, he was showing, in effect, how questions about several entities can always be thought of as being about a single entity. For example, the following are equivalent:

(7.2) $x \textit{ has } A$ and $y \textit{ has } B$ and $x \textit{ bears } R \textit{ to } y$

(7.3) $\langle x, y \rangle \textit{ has } A \otimes K$
$\langle x, y \rangle \textit{ has } K \otimes B$
$\langle x, y \rangle \textit{ has } R$

for an appropriately chosen product construction. The usual definition of satisfaction relates countably infinite sequences to open sentences. But it is quite possible to do the same job for, on the one hand, finite sequences, and on the other, the calculus of relations represented by sets of such sequences. In the second appendix I shall describe this construction somewhat more fully. Here I shall only state the conclusion that *if* there are significant relations among the entities in a domain, we should represent the person's judgments not simply by means of a scheme for that domain and a family of attributes pertaining to its members – but also by schemes for powers, or unions of powers, of that domain and pertinent relational attributes. All the schemes used to represent his judgments need to be frequency coherent; and that reflection should remove the uneasiness expressed above.

8. CONCLUSION: IS THERE A FUTURE FOR THE FREQUENCY INTERPRETATION?

Can we understand the activity of judgment, expressed in (vague) probability language, in a way that accords with the frequency interpretation of probability? I think we can, in two ways. The first is via the contention that the very aim of our judgment is to be a reliable indicator of actual frequencies of occurrence. The second is via a reflection on how that aim could be achieved, without essential recourse to deliberation about anything except the correct classification of the subjects and estimates of relative proportions among the classes involved.

As the central problem for this attempt I have selected Keynes' third question: how can the frequency interpretation justify our observance of the rules of the probability calculus, as intelligible and rational? It is clear that even with correct estimates of statistical frequencies, the selection of different reference classes on different occasions could easily lead to violations of those rules. Selection of the same reference class for all questions, on the other hand, would rob our judgments of all informative content.

My solution consisted in describing the expression of judgment as the expression of an epistemic attitude, and to discuss the proper evaluation of such an attitude under two headings: vindication and reasonableness. As a basic criterion of reasonableness (without any suggestion that it is the only criterion), I pointed to the requirement that vindication should not be *a priori* precluded. Now the main task at this point, for any interpretation of probability, is to explicate exactly what is vindication for a body of judgments. I argued that from the frequentist point of view, the notion of

calibration, as it appears in the Brier score, is the core criterion of vindication.

After having refined this notion so as to allow for at least a crude measure of approximation, and to explicate the relevant sense of possibility when we consider whether a person's judgments have potentially good calibration, I could then formulate the correlate basic criterion of reasonableness. This was a special concept of *potential calibration* which I called *frequency coherence*. And it was possible to prove that satisfaction of this criterion is equivalent to non-violation of the probability calculus. Hence Keynes' challenge has been met.

Now I believe that this has far-reaching consequences for the frequentist program as a whole. I insisted, in my short discussion of frequency schemes (i.e., models of judgment formation) that we should reject the idea that we must provide a *recipe* – i.e., set of determinate rules – for the selection of reference classes and formation of frequency estimates. This was on the more general grounds that we should not identify rationality with being compelled by requirements of rationality, but rather with being within their bounds, allowed by them. Rationality is only bridled irrationality.

The demonstration that potential vindication requires obedience to the probability calculus can now take over much of the job that recipes for reference class selection were meant to do. For suppose we choose reference classes for some basic questions, and form corresponding judgments. The probability calculus will then constrain our further judgments to a large extent – and to that extent, we can be totally uninterested in a recipe for what reference classes are or should be chosen in those further cases. Suppose that you choose reference classes for *rain all day* and *dry all day* and announce your personal probabilities as 0.2 for today's having the first attribute and 0.3 for its having the second. Now I ask you about its having the attribute *rain all day or dry all day*. Why should you stop to consult a recipe for choosing a reference class? You know now that whatever one you choose, you will be irrational unless you come up with the answer 0.5. Hence if anyone is interested in building such frequency interpretation models for judgment formation, he should, I think, be counselled that he can now, *without loss to his program*, make the probability calculus part of the constraints on the selection of reference classes. For the use of that calculus has been justified on independent but frequentist grounds.

An unsympathetic reader may at this point ask why we should bother with the frequency interpretation at all. Certainly I am much more anxious that contemplation of 'objective' probabilities should not lead to a belief in propensities – anxious, that is, to maintain an empiricist view of probabilistic

scientific theories – than I am to deny 'subjective' probability a status *sui generis*. But there is a philosophical question: why is the same name 'probability' appropriate for both? The wonder can presumably be removed only by a plausible explanation which entails either that the question is mistaken, and there is no connection at all, or else that there is a very intimate connection. Since the question can perhaps best be focussed on the special fact that the same axiomatic theory proves to be wonderfully useful in the explication of both uses of 'probability', we should especially ask why that should be. I hope to have shown that the frequency interpretation can remove this wonder by exhibiting an intimate connection that implies that the same, familiar axioms should cover both uses.[8]

Princeton University

APPENDIX I. FREQUENCY SCHEMES

In Kyburg's work, the literature contains an impressive, large-scale attempt to give a model for judgment which includes judgments concerning statistical frequencies (relative proportions in classes) and single case probabilities ('epistemic probabilities') based on these statistical judgments plus rule-governed selection of the right reference classes, via a generalization of the concept of the 'statistical syllogism'. As a result it is now difficult to stand back and canvass in an abstract way how frequentists *could*, in principle, go about constructing their models. Such a survey would nevertheless be of value, even if we came to see Kyburg's work as entirely succeeding in its aims, for it would be valuable to know whether the aims could be achieved some other way.

Given a domain of entities D and a Boolean algebra F of attributes (perhaps identified with subsets of a larger domain that includes D) I call a *scheme* any map P of the *propositions* $[x \textit{ has } A]$ with x in D and A in F, into real numbers. This scheme is meant to represent the surface phenomena of judgment after initial regimentation into probabilistic form. The notion of *frequency scheme* is much vaguer: a structure, suggested by the frequency interpretation, one part of which is such a scheme (i.e., a theoretical model for the phenomena of judgment). I have not indicated what entity the proposition $[x \textit{ has } A]$ is; the reader may choose a convenient identification, for example, with the ordered pair $\langle x, A\rangle$.

Suppose we ask the subject on a given occasion whether entity x has attribute A. I propose in general that his judgment is determined by four

factors. The first is a partially defined scheme σ for D and an extension F' of F – his *initial scheme*. The second is his *estimate* α which is a binary function partially defined on F'; "$\alpha(A\,|\,B) = r$" is read as "the proportion of entities that have A among those that have B equals r." Note that α has nothing to do with domain D *per se*, at least at this general level of discussion. The third is his *selector* β, a function that selects for each x in D and each A in F a class $\beta(x, A)$ of attributes in the algebra F, and perhaps for some in $F' - F$ as well. Note that I have generalized the choice of a reference class to selection of a class of reference classes, for reasons made clear below. And the fourth is his *strategy* Σ which determines a numerical grade (*personal probability*) for each proposition on the basis of the foregoing. Let $M = \langle \sigma, \alpha, \beta, \Sigma\rangle$ be called a *frequency scheme*, and abbreviate

$$(I.1) \qquad P_M\,[x \textit{ has } A] = \Sigma(\sigma, \alpha, \beta)\,[x \textit{ has } A]$$

which is the *scheme* of frequency scheme M. We may at once impose the requirement that Σ be entirely determined by σ where defined, that is

$$(I.2) \qquad P_M\,[x \textit{ has } A] = \sigma\,[x \textit{ has } A] \text{ when defined}$$

The remainder of the structure represents the procedure of deliberation whereby the initial scheme is extended to other propositions in accordance with frequentist intuitions.

Now I will give some examples of what frequency schemes can be like. The first is the simplest. The initial scheme represents only something like initial full belief (or 'taking as evidence'), it just assigns zeroes and ones to some propositions. The selector β now acts as follows: for the couple $\langle x, A\rangle$ it selects a single attribute B such that (i) σ assigns 1 to $[x \textit{ has } B]$ and (ii) $\alpha(A\,|\,B)$ is defined. Denote the attribute selected, in general, as $\beta_{x,A}$. Finally, Σ then simply assigns that estimated proposition. Thus we have, for $M = \langle \sigma, \alpha, \beta, \Sigma\rangle$:

$$(I.3) \qquad P_M\,[x \textit{ has } A] = \alpha(A\,|\,\beta_{x,A}).$$

This follows closely Venn's original idea that we classify the subject (with no account taken of doubts about the classification) and announce the statistical frequency (assumed known) in that reference class.

But this does not seem very realistic to me. Does it not seem more plausible that we base our opinion in part on classifications of the subject, for which we have only partial certainty? So we can envisage a slightly more elaborate frequency scheme in which σ assigns some numbers between *zero*

and *one* as well. Here let β select as $\beta(x, A)$ a *partition* of attributes $B_1, \ldots, B_k$ for which σ is defined, that is, $\sigma(B_j \cap B_j) = 0$, $\sigma(B_1 \cup \ldots \cup B_k) = 1$, with $\alpha(A \mid B_i)$ defined for each. Then Σ should act so as to yield, for $M = \langle \sigma, \alpha, \beta, \Sigma \rangle$:

(I.4) $P_M [x \text{ has } A] = \Sigma \{\sigma [x \text{ has } B] \ \alpha(A \mid B) : B \in \beta(x, A)\}$

where this capital sigma is the summation sign.

But now, as $\Sigma(\sigma, \alpha, \beta)$ extends σ, the new propositions to which probabilities are assigned, can also begin to play a role in deliberation. So we can describe a third type of frequency scheme. In that larger algebra F', we may introduce a partial ordering. This has nothing to do with the Boolean operations *per se*, but it may have something to do with the subject x of the question at issue, so call it "x-precedes." In this type of scheme, σ is defined for $[x \text{ has } A]$ only if nothing x-precedes A. In that case principle (I.2) applies. Next we look at the case in which something x-precedes A; then β may select a partition of attributes that x-precede A, with the same conditions fulfilled for σ and α, so that (I.4) can apply. (Note that (I.3) is just a special case of (I.4) if σ assigns only zeroes and ones.) Finally, we come to the case where something x-precedes A but β does not act so that (I.4) can apply; then β must still select a partition $\beta(x, A)$ of attributes which x-precede A, and the following principle should be applicable:

(I.5) $P_M [x \text{ has } A] = \Sigma \{P_M [x \text{ has } B] \ \alpha(A \mid B) : B \in \beta(x, A)\}$.

The use of the partial ordering *x-precedes* allows these principles to govern the action of the strategy Σ without circularity. To the extent that P_M is defined on propositions $[x \text{ has } B]$ for attributes B that *x-precede* A, it takes over the role of initial scheme σ in the constraints on $\beta(x, A)$ and determination of $P_M [x \text{ has } A]$.

At this point we might even speculate again that restriction of σ to the assignment of zeroes and ones only, might not unduly impoverish the stock of frequency schemes of this third type. If the attributes A are sophisticated enough, a proposition $[x \text{ has } A]$ might well be the exact information that the proportion of Bs among Cs equals 0.75, say. In that case, α could be built up simultaneously with Σ. These two reflections are in the direction of what Kyburg's constructions are meant tc hieve. As among Bayesians, we can see a divergence of inclination toward 'global models' and 'local models' ('small worlds') respectively, open to frequentists as well. I have not discussed severe constraints on the selector here; for that I refer back to the last section of the body of this paper.

APPENDIX II. RELATIONS AND PRODUCT CONSTRUCTIONS

In Section 7 I discussed calibration of a scheme for sets of propositions about several individuals, and the difficulties that could occur due to relations among these. For example, x might be Christmas day and y Christmas morning, so that rain on y and dry weather throughout x are not logically independent. I shall here describe in some more detail the kind of product construction in which questions about several individuals are reduced to ones about a single entity, in a way directly relevant to this paper.

For definiteness I shall take the algebra F of attributes to be a field of subsets of a given set K (the maximal element of F). A *model* in which propositions receive truth values is then a couple $M = \langle loc, D \rangle$, where *loc* maps D into K, and $[x \text{ has } A]$ is true in M exactly if $loc(x)$ is a member of set A.

To take account of relational attributes, we focus on domain D^∞, which is the class of all finite sequences of members of D. We let K be itself a set K_0^∞ and F a field of subsets thereof. Intuitively we identify the binary relation R with all the sequences $e = \langle e(1), \ldots, e(n) \rangle$ in K such that $e(1)$ bears R to $e(2)$. Thus R is identified with a set which contains all finite elongations of its members. In the present case we say that R nevertheless has degree 2, because there is a subset of K_0^2 which 'determines' R. Stated precisely:

(II.1) If Y is a subset of X^∞ then Y^+ is the set of all members of X^∞ which have some initial segment that is in Y; and the *degree* of Y (if any) is the least positive integer m for which there exists a set Y^* such that for all e in X^∞, e is in Y if and only if $\langle e(1), \ldots, e(m) \rangle$ is in Y^*.

(II.2) RESTRICTION. *Each attribute in F has a finite degree.*

It follows at once that if A is in F, then $A = A^+$ (the operation$^+$ understood contextually here with reference to K_0^∞). This restriction is compatible with the Boolean character of F as a field of sets, but keeps it from being a sigma-field. (For example, the degree of $A \cap B$ is the maximum of the degrees of A, B if any.) Note that the degrees of K and of Λ equal 1, since e is in K (respectively, Λ), if and only if $\langle e(1) \rangle$ is in K_0^1 (respectively, Λ), where we denote as X^n the set of all n-tuples of members of X.

A *model* must be restricted so as to observe the structural relations among the sequences:

(II.3) $M = \langle loc, D^{\infty}\rangle$ is a *model* for D and F exactly if loc maps D into K_0 and for x in D^{∞}, of length n, $loc\,(x) = \langle loc\,(x(1)), \ldots, loc\,(x(n))\rangle$.

We define the following operations on K and its subsets:

(II.4) If $A_1, \ldots, A_n$ have degrees $m(1), \ldots, m(n)$ respectively, then $A_1 \otimes \ldots \otimes A_n$ is the set whose members are all sequences e in K such that $\langle e(1), \ldots, e(m(1))\rangle$ is in $A_1, \ldots, \langle e(m(n-1)+1), \ldots, e(m(n-1)+m(n))\rangle$ is in A_n.

(II.5) $e(m+n)b = \langle e(1), \ldots, e(m), b(1), \ldots, b(n)\rangle$ and undefined if the lengths of e, b are less than m, n respectively.

(II.6) Where m, n are the degrees of A, B respectively,
$A \overline{\wedge} B = \{b(m+n)d : b \text{ in } A \text{ and } d \text{ in } B\}^{+}$
$A \underline{\vee} B = \{b(m+n)d : b \text{ in } A \text{ or } d \text{ in } B\}^{+}$.

I shall call $\overline{\wedge}$ and $\underline{\vee}$ the *directed meet* and *directed join*. Note that the degree of $A_1 \otimes \ldots \otimes A_n$ equals the sum of the degrees of $A_1, \ldots, A_n$, and similarly for $A_1 \overline{\wedge} A_2$, $A_1 \underline{\vee} A_2$. We impose on F also:

(II.7) RESTRICTION. *If $A_1, \ldots, A_n$ are in F so are $A_1 \otimes \ldots \otimes A_n$, $A_1 \overline{\wedge} A_2$, $A_1 \underline{\vee} A_2$, and each set K_0^n, $n = 1, 2, \ldots$.*

This is again compatible with the Boolean character of F and with (II.2).

It is clear that the directed meet and join are not commutative, but on the level of truths of propositions commutation is effectively restored. The $(m+n)$ operation makes sense for any sequences, hence can be used on D as well. Then we see

(II.4) $[x(m+n)y \text{ has } A \underline{\vee} B]$ is true in model $M = \langle loc, D\rangle$ iff $\langle loc\,(x(1)), \ldots, loc\,(x(m))\rangle$ is in A or $\langle loc\,(y(1)), \ldots, loc\,(y(n))\rangle$ is in B, hence iff $[x \text{ has } A]$ or $[y \text{ has } B]$ is true in M

where it was assumed that A, B have degrees m and n respectively, and x, y appropriate lengths. Thus we see that if *we identify a proposition with the set of models in which it is true*, then

(II.5) $[x(m+n)y \text{ has } A \underline{\vee} B] = [x \text{ has } A] \cup [y \text{ has } B]$,

and similarly for directed meet and intersection.

Let us now inspect the adequacy proofs for the special case of such a product construction. There are just two points that must especially be made. It may seem at first that we need to modify the notion of P-alike, by stipulating that not only $P[x \text{ has } A] = P[y \text{ has } A]$ for all A in F, but also that $P[\langle x(k), \ldots, x(k+m)\rangle \text{ has } A] = P[\langle y(k), \ldots, y(k+m)\rangle \text{ has } A]$ for all A in F as long as $k+m$ is not too long. Actually no such emendation is needed, because $[\langle x(k), \ldots, x(k+m)\rangle \text{ has } A]$ is the same proposition as $[x \text{ has } K_0{}^{k-1} \otimes A \otimes K_0{}^{r}]$ where r equals the length of x minus $(k+m)$. The second point relates to the justification of the additivity principle in the second Adequacy proof, the only place where we deal explicitly with an initial set containing more than one proposition. Consider:

$$P([x \text{ has } A] \cup [y \text{ has } B]) + P([x \text{ has } A] \cap [y \text{ has } B]) = P([x \text{ has } A]) + P([y \text{ has } B]).$$

Let m and n be the degrees of A, B respectively. Then the four propositions can all be seen to be identical with propositions about the single entity $x(m+n)y$:

$$[x \text{ has } A] \cup [y \text{ has } B] = [x(m+n)y \text{ has } A \underline{\vee} B]$$
$$[x \text{ has } A] \cap [y \text{ has } B] = [x(m+n)y \text{ has } A \overline{\wedge} B]$$
$$[x \text{ has } A] = [x(m+n)y \text{ has } A \otimes K_0{}^{n}]$$
$$[y \text{ has } B] = [x(m+n)y \text{ has } K_0{}^{m} \otimes B].$$

After this the proof can proceed as before.

NOTES

* Through his writings and as my teacher, dissertation supervisor, and friend, Adolf Grünbaum has been my main guide into philosophy of science ever since I read his 1955 article on the foundations of special relativity, which I came across as a undergraduate. I dedicate this paper to him, with sincere gratitude and warm affection. Support for this research by the National Science Foundation is gratefully acknowledged. A preliminary version of this paper was circulated in September 1979. I have learned that results that appear to be similar to the theorems in this paper were stated in a public lecture by Abner Shimony in 1978. I also wish to thank J. Hellige and W. Edwards, of the University of Southern California, for helpful discussions.

[1] The equation is not a simple one; see my (1979).

[2] *Annales de l'Institut Henri Poincaré*, vol. 7 (1937), in English translation in Kyburg and Smokler (1964).

[3] Both Reichenbach and Salmon have discussed vindication (I owe the term to Salmon) of predictions in connection with probability and inductive strategy, and have been concerned to analyze the condition of *possible* vindication. Hence my strategy here continues *one* venerable strand in frequentist thinking.

[4] Venn (1888), p. 213.

[5] See Brier (1950); there is now a large body of literature on scoring in general. For the decomposition into calibration and extremeness, see Dickey (1974), and Murphy (1972). See also de Finetti (1965), Pickhardt and Wallace (1974), Shuford *et al.* (1966), Winkler and Murphy (1968); and see further Note 8 below.

[6] In my own view of theories the truth requirement is one of empirical adequacy (truth about what is both actual and observable) only. Information has several objective dimensions, such as logical strength and what I call empirical strength, but also plays an essential role in such pragmatic virtues as being explanatory (informative in 'relevant' respects). See my (1980), (1981), (1982).

[7] Reichenbach formulated a crude measure of vindication which he used at several places in his (1949), including in his discussion of "single case probability" (which he called a "pseudo-concept," that "must be replaced by a substitute constructed in terms of class probabilities.") That is, if a person assents to all propositions about individual events of sort B, when he believes the relative frequency of B to be $\geqslant r$, then he will be right in proportion $\geqslant r$ of the cases, if that belief is correct. By taking $r > 1/2$ he will thus be right more often than not. This refers to a choice of the same reference class for each question about an individual having attribute B. Reichenbach then points out that if we switch to a smaller reference class, in which the proportion of B is higher, the proportion of success in our predictions will also increase. He did not, as far as I know, investigate what proportion of success is possible in the general case in which the questions are about different attributes, and the reference classes chosen may vary, even for the same attributes from individual to individual. But although measurement by a division of this type (assent at level $\geqslant r$) is crude, it is the *sort* of measure of vindication that is needed here.

[8] The reader may well have wondered how Bayesians can or should approach the question of 'correct' scoring procedures. A good indication is found in the results proved by Shuford *et al.* (1966) who suggest as a basic criterion that a scoring procedure is admissible exactly if anyone can maximize his expected score if and only if he correctly reports his personal probabilities. (The expected score is of course the score's expectation value calculated by his own personal probability.)

REFERENCES

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

de Finetti, B. 1965. 'Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item,' *British Journal of Mathematical and Statistical Psychology* **18**, 87–123.

Dickey, J. M. 1974. 'Comments on Suppes,' *Journal of the Royal Statistical Society* **B36**, 179–180.

Keynes, J. M. 1921. *Treatise on Probability*. London: Macmillan.

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

Kyburg, H. E., Jr. and H. E. Smokler (eds.), 1964. *Studies in Subjective Probability*. New York: Wiley.

Murphy, A. H. 1972. 'Scalar and Vector Partitions of the Probability Score,' *Journal of Applied Metereology* **11**, 273–282.

Pickhardt, R. C. and J. B. Wallace. 1974. 'A Study of the Performance of Subjective Probability Assessors,' *Decision Sciences* **5**, 347–363.

Reichenbach, H. 1949. *The Theory of Probability*. 2nd ed. Berkeley: University of California Press.

Salmon, W. 1967. *The Foundations of Statistical Inference*. Pittsburgh: University of Pittsburgh Press.

Shuford, E. H., A Albert, and H. E. Massen. 1966. 'Admissible Probability Measurement Procedures,' *Psychometrika* **31**, 125–145.

van Fraassen, B. C. 1979. 'Foundations of Probability Theory: A Modal Frequency Interpretation.' In G. Toraldo di Francia (ed.), *Problems in the Foundations of Physics*. Amsterdam: North-Holland.

van Fraassen, B. C. 1980. *The Scientific Image*. Oxford: Clarendon Press.

van Fraassen, B. C. 1981. 'Theory Construction and Experiment: An Empiricist View.' In P. Asquith and R. Giere (eds.), *PSA 1980*, vol. 2. East Lansing, Michigan: Philosophy of Science Association.

van Fraassen, B. C. 1982. 'Glymour on Evidence and Explanation.' In J. Earman (ed.) *Minnesota Studies in the Philosophy of Science*, vol. 10. Minneapolis: University of Minnesota Press, forthcoming.

Venn, J. 1888. *The Logic of Chance* (1886). London: Macmillan.

Winkler, R. L. and A. H. Murphy 1968. ' "Good" Probability Assessors,' *Journal of Applied Metereology* 7, 751–758.