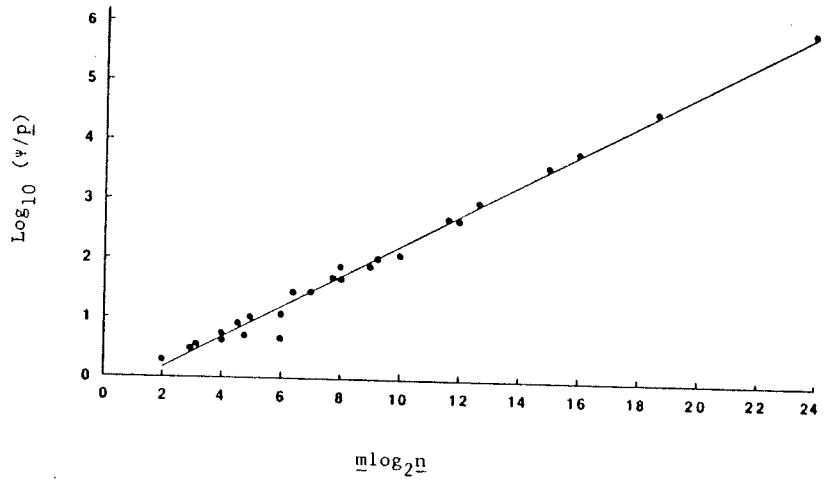

Figure 1. Log relative overestimation plotted against the number of bits ($\log_{10}(\Psi/p)$ against $m \log_2 n$). p = compound probabilities; m = number of stages; n = number of alternatives.

Our interpretation is supported by similar experiments with subjects aged 9+ and 10+ years, in which, apart from gross relative overestimation of Ψ, no trends are discernible with variations in the values of m and n. This too suggests that the multiplicative element is not primitive.

The method we have used involves an indirect evaluation of Ψ. A more direct evaluation could be obtained by asking the subject to choose between different types of array. The utility of the choice, however, might then become an important factor.

This experiment elucidates the apparent tendency, in a variety of multi-stage choice situations, for the decision maker to misjudge the likelihood of his success, and therefore to adopt an inappropriate strategy which he will later regret.

Of historical interest in this connection is the fact that the most subtle thinkers of ancient Greece, though greatly intrigued by the idea of the possible, especially in Stoic philosophy, never grasped combinatorial analysis, which had to wait until the sixteenth century for its development. Aristotle himself evidently had only a small appreciation of the concept of probability. Whatever intuition of the subject he and others might have had was submerged by long established habits of thought.

The relative overestimation of compound probabilities which the experiment has revealed may be a phenomenon of considerable generality in decision and choice. If so, it merits a special designation. We propose to name it the "inertial Ψ effect."

25. Conservatism in human information processing

Ward Edwards

. . . An abundance of research has shown that human beings are conservative processors of fallible information. Such experiments compare human behavior with the outputs of Bayes's theorem, the formally optimal rule about how opinions (that is, probabilities) should be revised on the basis of new information. It turns out that opinion change is very orderly, and usually proportional to numbers calculated from Bayes's theorem – but it is insufficient in amount. A convenient first approximation to the data would say that it takes anywhere from two to five observations to do one observation's worth of work in inducing a subject to change his opinions. A number of experiments have been aimed at an explanation for this phenomenon. They show that a major, probably the major, cause of conservatism is human misaggregation of the data. That is, men perceive each datum accurately and are well aware of its individual diagnostic meaning, but are unable to combine its diagnostic meaning well with the diagnostic meaning of other data when revising their opinions. . . .

Probabilities quantify uncertainty. A probability, according to Bayesians like ourselves, is simply a number between zero and one that represents the extent to which a somewhat idealized person believes a statement to be true. The reason the person is somewhat idealized is that the sum of his probabilities for two mutually exclusive events must equal his probability that either of the events will occur. The additivity property has such demanding consequences that few real persons are able to conform to all of them. Since such probabilities describe the person who holds the opinion more than the event the opinion is about, they are called personal probabilities (see Savage, 1954).

Bayes's theorem is a trivial consequence of the additivity property, uncontroversial and agreed to by all probabilists, Bayesian and other. One

Excerpts from a paper that appeared in B. Kleinmuntz (Ed.), *Formal Representation of Human Judgment*. New York: John Wiley and Sons, Inc., 1968. Reprinted by permission.

way of writing it is as follows. If $P(H_A|D)$ is the posterior probability that hypothesis A has after datum D has been observed, $P(H_A)$ is its prior probability before datum D is observed, $P(D|H_A)$ is the probability that datum D will be observed if H_A is true, and $P(D)$ is the unconditional probability of datum D, then

$$P(H_A|D) = \frac{P(D|H_A)\,P(H_A)}{P(D)} \tag{1}$$

$P(D)$ is best thought of as a normalizing constant, intended to make the posterior probabilities add up to one over the exhaustive set of mutually exclusive hypotheses being considered. If it must be calculated, it can be as follows:

$$P(D) = \sum_i P(D|H_i)\,P(H_i)$$

But more often $P(D)$ is eliminated rather than calculated. One convenient way of eliminating it is to transform Bayes's theorem into its odds-likelihood ratio form. Consider another hypothesis, H_B, mutually exclusive of H_A, and modify your opinion about it on the basis of the same datum that changed your opinion about H_A. Bayes's theorem says

$$P(H_B|D) = \frac{P(D|H_B)\,P(H_B)}{P(D)} \tag{2}$$

Now divide Equation 1 by Equation 2; the result is

$$\frac{P(H_A|D)}{P(H_B|D)} = \frac{P(D|H_A)}{P(D|H_B)} \cdot \frac{P(H_A)}{P(H_B)}$$

or

$$\Omega_1 = L \cdot \Omega_0, \tag{3}$$

where Ω_1 is the posterior odds in favor of H_A over H_B, Ω_0 is the prior odds, and L is a quantity familiar to statisticians as a likelihood ratio. Equation 3 is as appropriate a version of Bayes's theorem as Equation 1, and often considerably more useful especially for experiments involving two hypotheses.

Bayesian statisticians argue that Bayes's theorem is a formally optimal rule about how to revise opinions in the light of evidence, that revision of opinion in the light of evidence is exactly what statistical inference consists of, and that therefore statistical inference should be structured around Bayes's theorem – with many consequent differences from classical statistical practice. For an elementary exposition of these ideas written for experimenting psychologists, see Edwards, Lindman, and Savage (1963). But we are not statisticians, or at any rate none of us are wearing our

statistician's dunce caps today. Instead, as psychologists, we are interested in comparing the ideal behavior specified by Bayes's theorem with actual human performance.

To give you some feeling for what follows, let us try an experiment with you as subject. This bookbag contains 1,000 poker chips. I started out with two such bags, one containing 700 red and 300 blue chips, the other containing 300 red and 700 blue. I flipped a fair coin to determine which one to use. Thus, if your opinions are like mine, your probability at the moment that this is the predominantly red bookbag is 0.5. Now, you sample, randomly, with replacement after each chip. In 12 samples, you get 8 reds and 4 blues. Now, on the basis of everything you know, what is the probability that this is the predominantly red bag? Clearly it is higher than 0.5. Please don't continue reading till you have written down your estimate.

If you are like a typical subject, your estimate fell in the range from 0.7 to 0.8 – though the statement frequently made in the preceding paragraphs that men are conservative information processors may have biased your answer upward. If we went through the appropriate calculation, though, the answer would be 0.97. Very seldom indeed does a person not previously exposed to the conservatism finding come up with an estimate that high, even if he is relatively familiar with Bayes's theorem.

In about 1960, William L. Hays, a graduate student named Lawrence D. Phillips, and I were interested in finding discrepancies between human performance and that specified by Bayes's theorem. The simple example of the previous paragraph didn't occur to us; instead we were sure that we would need to use a fairly complex situation in order to get non-Bayesian behavior. So we used a hypothetical computerized radar system. There were 12 possible observations, 4 possible hypotheses, and so subjects had to understand and use a display of 48 different values of $P(D|H)$. Subjects worked under two conditions. In one, the subject saw a single stimulus, a dot in a sector of a radar scope; he then revised his prior probabilities over the four hypotheses on the basis of the datum by setting four levers to his posterior probability estimates, then reset the levers to 0 in preparation for the next stimulus. The second stimulus consisted of the old dot plus a new one; the subject set his levers to report the cumulative impact of both dots. And so on, until 15 dots had accumulated. In the second condition, the stimuli were shuffled, and the subject in effect started afresh with each new stimulus. To the surprise of the experimenters the prediction of Bayes's theorem that this difference in conditions should make no difference to behavior was borne out. Moreover, there was yet another condition in which each new dot was displayed alone, but the subjects were allowed to preserve their estimates from one stimulus to the next rather than resetting levers to zero after each estimate. Again, the variation in conditions made little difference to behavior.

The positive findings of the Phillips-Hays-Edwards experiment were three in number. First, subjects were overwhelmingly conservative. Secondly, they were least conservative on the first dot, becoming more so with more dots. Finally, the sums of their probability estimates, which were not constrained, in general added up to more than 1, and increased as the subjects progressed through successive stimuli in an ordered sequence. Apparently the subjects found it easier to determine which hypothesis was favored by a stimulus, and so to increase the probability of that hypothesis, than to decide from which other hypotheses probability should be withdrawn in order to give it to the favored one.

We were notably dilatory in publishing this original conservatism experiment. Though the data were complete by 1962, the Phillips-Hays-Edwards paper didn't make it into print until 1966 (Phillips et al., 1966).

The magnitude and consistency of the conservatism finding startled us. It seemed appropriate to try much simpler tasks. So, without much faith, Phillips and I tried a pretest similar in character to the bookbag and poker chip example you tried above. To our surprise, it worked very well. Most of the current research comparing human behavior with Bayes's theorem can be traced to that pretest and the subsequent experiment.

If the proportion of red chips in the bookbag is p, then the probability of getting r red chips and $(n - r)$ blue chips in n samples with replacement in a particular order is $p^r(1 - p)^{n-r}$. So in a typical bookbag and poker chip experiment, if H_A is that the proportion of red chips is p_A and H_B is that that proportion is p_B, then the likelihood ratio is

$$L = \frac{p_A^r(1 - p_A)^{n-r}}{p_B^r(1 - p_B)^{n-r}} \tag{4}$$

Note that while Equation 4 was derived from considering the actual sequence of reds and blues in the sample, it could equally well have been derived from considering r reds and $(n - r)$ blues in any order; the binomial coefficient that represents the number of different ways one can obtain r reds in n draws appears in both numerator and denominator and thus cancels out of the likelihood ratio. This is an illustration of the likelihood principle of Bayesian statistics (see Edwards, Lindman, & Savage, 1963), which in effect says that a Bayesian need consider only the probability of the actual observation he has made, not the probabilities of other observations that he might have made but did not. This principle has sweeping impact on all statistical and nonstatistical applications of Bayes's theorem; it is the most important technical tool of Bayesian thinking.

In the special case in which $p_A = 1 - p_B$ (the symmetric binomial case), the likelihood ratio reduces to

$$L = \left(\frac{p_A}{1 - p_A}\right)^{2r - n} \tag{5}$$

Note that $2r - n = r - (n - r)$ is the difference between the number of reds

and the number of blues in the sample; only that difference, and not the total number of observations, is relevant to inference in this symmetric case. Statistical tradition labels that difference successes minus failures, or $s - f$; $s - f$ is the usual independent variable of bookbag and poker chip experiments. To understand the rationale for the usual dependent variables, substitute Equation 5 into Equation 3, take logarithms and rearrange terms. The result is

$$\log L = (2r - n)\log \frac{p_A}{1 - p_A} = \log \Omega_1 - \log \Omega_0$$

If the subject is perfectly Bayesian, the log likelihood ratio that can be inferred by subtracting the log of the prior odds from the log of the posterior odds should be proportional to $s - f$, the independent variable. It is appropriate to plot the subject's inferred log likelihood ratio, thus calculated from his posterior odds (which in turn were calculated from his posterior probabilities if he was estimating probabilities) and the objectively appropriate prior odds, against $s - f$.

Most of the bookbag and poker chip experiments in the Michigan laboratory have used a display consisting of 48 numbered locations each containing a pushbutton, a red light, and a green light. When the button at a location is pushed, one of the lights goes on and stays on; subjects are told that this is equivalent to a sample with replacement of a chip of the corresponding color from the bookbag. The subjects are told that the program that controls the lights was prepared by sampling from a bookbag. Actually, for most experiments that program is rather carefully prepared so that the displayed sequence is appropriately representative of the bookbag, and in particular so that in each experiment samples of size n favor the untrue hypothesis appropriately often for the value of p_A being used, for all values of n.

Phillips and I (1966) investigated the effect of p_A, using sequences of 20 chips and p_A values of 0.55, 0.7, and 0.85. Subjects estimated posterior probabilities by distributing 100 white wooden discs over two troughs. Typical results of such experiments are presented in Figure 1, for the 0.7 bag with various prior probabilities. Three findings, illustrated in Figure 1, appeared for all subjects. First, the inferred log likelihood ratios were roughly proportional to $s - f$. Second, the prior probabilities were appropriately used; that is, the best fitting line through the data points passes through the origin. Third, subjects were conservative; the best fitting line was flatter than the line representing optimal Bayesian performance. The finding of near-linearity of inferred log likelihood ratios with $s - f$ (or, equivalently, with Bayesian log likelihood ratios) suggests yet another dependent variable: the ratio of the slope of the best-fitting line through the subject's estimates to the slope of the Bayesian line. Peterson, Schneider, and Miller (1965) have named that ratio the accuracy ratio; they also found it more or less constant with $s - f$.

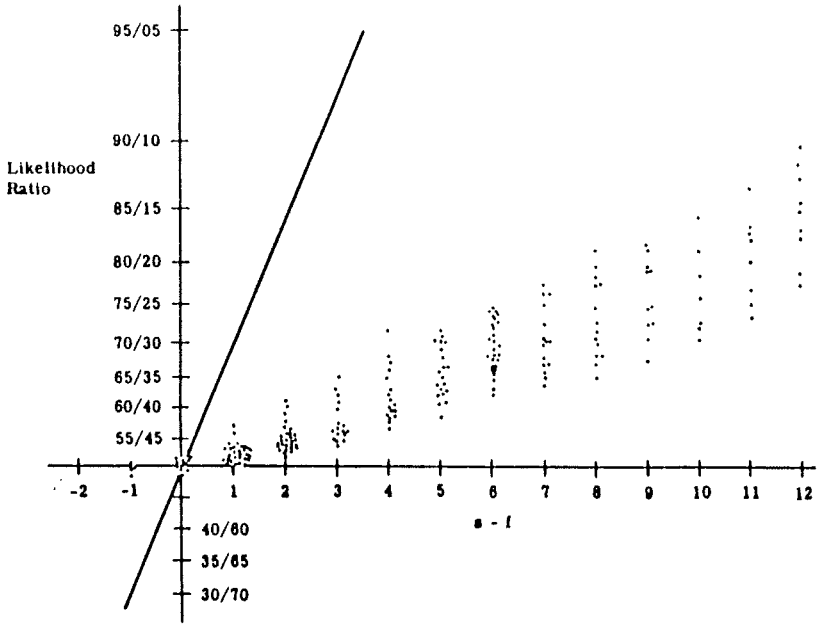

Figure 1. A single subject's estimates for p_A of 0.7, expressed in inferred log likelihood ratios as a function of the difference between the number of successes and the number of failures in the sample.

Figure 2 shows accuracy ratios for the Phillips-Edwards data for the three values of p_A. For the least diagnostic information, the subjects were more extreme than Bayes's theorem. (Dale has found the same thing; see W. Edwards, 1965.) But for information having reasonably high diagnostic value, subjects were conservative, and the accuracy ratio was nicely constant with $s - f$. Note that as diagnosticity increases, conservatism increases also. This is a standard finding of such experiments; any procedure that increases diagnosticity of the individual observation (of one chip or several) also increases conservatism. (See for example Peterson, Schneider, & Miller, 1965.)

Phillips and I, after obtaining these results, speculated that one reason for conservatism might be that subjects, knowing that the probability scale is bounded and observing that evidence might go on mounting up and up, were holding their estimates down. The obvious remedy, if so, is to use an unbounded response mode, like odds. So we ran a four-group study. The control group estimated probabilities by distributing 100 discs over two troughs, as before. The verbal odds group simply made verbal estimates of odds; we always take odds as numbers equal to or greater than one, and therefore always accompany odds statements by statements of which

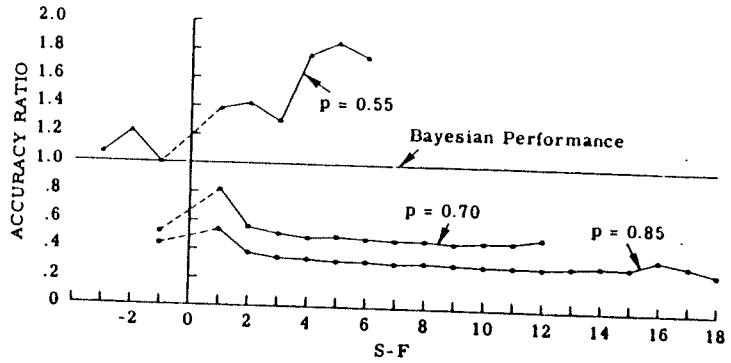

Figure 2. Accuracy ratios for three values of p_A over various sample compositions.

hypothesis is favored by the odds. The odds on a log scale group made their estimates by moving a pointer along an odds scale which contained four log cycles, so that odds anywhere from 1:1 to 10,000:1 could be estimated. The fourth group used the odds on a log scale device also, but the numbers entered opposite the scale markings were probabilities rather than odds (thus 0.5 rather than 1:1, 0.67 rather than 2:1, 0.80 rather than 4:1, etc.). It was called the probability on a log odds scale group. The findings were that all groups were quite conservative. The probability group was most so, probability on a log odds scale was next worst, and the two odds groups were about comparable, with odds on a log scale slightly superior.

This finding simply underlines a fact that has become increasingly clear in the course of Bayesian work. Probability is a rather poor measure of uncertainty, except in situations in which repartitioning or other direct use of the additivity property is necessary. Either odds or log odds is better. Odds is most intuitive for naive subjects, and can most easily be linked to simple acts (e.g., choices among bets); the fact that the gambling industry structures all its statements and displays around odds rather than probability is both recognition of and perhaps cause of the greater intuitive value of odds. Log odds, uniquely among the more-or-less common metrics for uncertainty, has the property that in that metric evidence is additive. If opinion is measured in log odds, the amount of change in opinion produced by a piece of evidence is independent of where the opinion was to start with. This elegant property makes log odds uniquely convenient for Bayesian experiments.

The Phillips-Edwards data can be well fit by a simple modification of Bayes's theorem:

$$\Omega_1 = L^c \Omega_0$$

The constant c, the power to which each likelihood ratio is raised before

processing it by means of Bayes's theorem, is the accuracy ratio. Unfortunately, it is dependent on important independent variables, including diagnosticity of the data and response metric. Still, the fact that so simple a descriptive model fits so well must be explained by any theory of conservatism. . . .

. . . A Probabilistic Information Processing system, or *PIP*, . . . is an idea about how to design man-machine systems that must process information for the purpose of reaching a conclusion about what state the world is in. Examples of settings in which such information processing must be done include medical diagnosis, military command (in which a commander may need to determine whether or not he is under attack, and if so, what his opponent's plan is), and business management (for example, in the case of a businessman deciding whether or not to manufacture a new product). The idea of *PIP* is much too complicated to explain in detail here. For recent expositions of it, see Edwards, Lindman, and Phillips (1965), or W. Edwards (1966). The essence of it is that the task of diagnostic information processing can be divided into two classes of subtasks. One class of subtasks consists of the judgment of the diagnostic impact of an individual datum on a single hypothesis or pair of hypotheses. For the verbal, qualitative kinds of data and hypotheses that characterize many real diagnostic settings, this seems to be a task necessarily done by men, the more expert the better. But the second class of subtasks is the aggregation of these separate diagnostic impacts across data and across hypotheses into a picture of how all the hypotheses currently stand in the light of all available data. This aggregation task is readily mechanized by means of Bayes's theorem, if the diagnostic impacts of the individual data are judged in the form of $P(D|H)$ values or likelihood ratios. (In most situations, though not all, judgments of likelihood ratios are clearly preferable, for formal reasons, to judgment of $P(D|H)$.)

About fifteen collaborators and I were interested in finding out whether PIP works or not. So we designed an imaginary but elaborate world of 1975. In that world we listed six hypotheses that subjects were to consider, specified three data sources (the Ballistic Missile Early Warning System, a reconnaissance satellite system, and the intelligence system) that provided data bearing on these hypotheses, and designed four information processing systems to process the data. The four systems were named PIP, POP, PEP, and PUP. In PIP, the subjects estimated five likelihood ratios per datum. One of the six hypotheses was "Peace will continue to prevail" and the other five were various possible wars; the five pairings of a war with peace specified the five likelihood ratios to be estimated. The other three information processing systems all had in common that the subject estimated posterior odds or probabilities or similar posterior quantities; thus in PIP the computer aggregated the data by means of Bayes's theorem, while in all three other systems the subjects had to aggregate the data in their heads. To help them do this, the subjects in POP, PEP, and PUP had

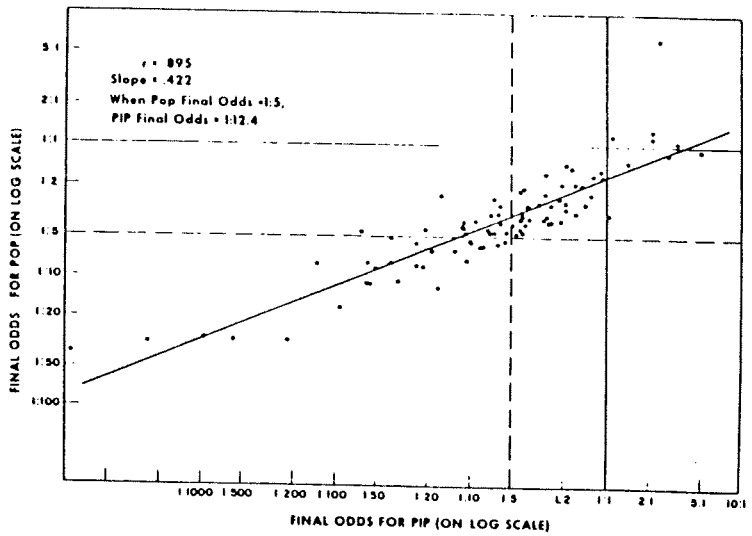

Figure 3. Final odds in favor of war for POP vs. PIP plotted on log scales.

heir estimates after the nth datum available when they considered the $n + 1$)th datum, so they only needed to modify those estimates affected y the datum.

There were a total of 18 scenarios, with 60 data items per scenario. All ata items except for those from the Ballistic Missile Early Warning System ere in the form of short paragraphs. The 34 subjects were exhaustively rained in the characteristics of the world, the hypotheses, the three data ources, and the information processing system each was to operate.

Since PIP was clearly best and POP was next best, I shall present only he comparison between them. (PUP was third best, and PEP, the nearest e could get to how such information processing is done now, was worst.) igure 3 shows the final odds, after the 60th datum in each scenario, in avor of each war as compared with peace for PIP and for POP. The two ost important things to note about the figure are that the two groups gree very well qualitatively (the correlation between them is 0.895), but ey disagree quantitatively. PIP is much more sensitive to data than POP; he same scenario that will lead PIP to be very sure of peace or of some war ill lead POP to be much less sure. To put it another way, PIP is much less onservative than POP – presumably because in POP, the subjects must ggregate the data, while in PIP, the subjects judge the diagnostic impact f each datum separately and Bayes's theorem does the aggregating.

You should note also that both axes on Figure 3 are logarithmically paced. If you translate the difference in efficiency back into odds, the ramatic difference between PIP and POP becomes apparent. For example, lculating from the regression line, if a scenario led PIP to give 99:1 odds

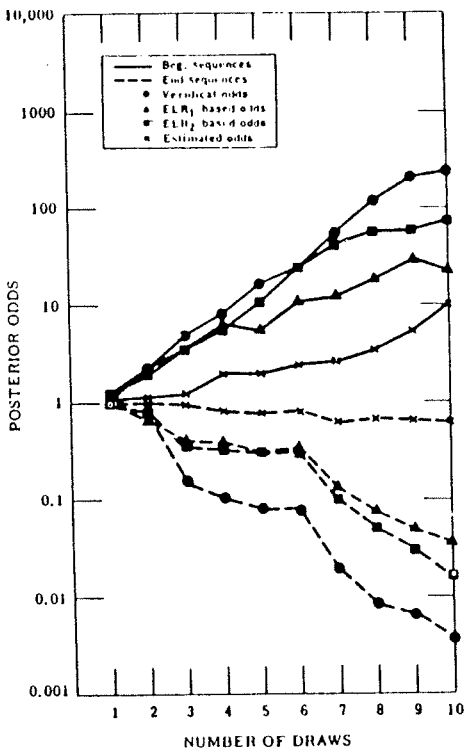

Figure 4. Median posterior odds, across subjects, in favor of the beginning bookbag as a function of number of draws.

in favor of some war over peace, POP would give only 4:1 odds in favor of that war over peace.

The misperception hypothesis cannot possibly explain this discrepancy between PIP and POP. The PIP subjects estimate the diagnostic impact of each datum separately; the POP subjects must aggregate in their heads - and do so quite conservatively. Since no model of the data-generating process is available, it is impossible to say what the right posterior odds are. But the difference between PIP and POP is clearly caused by a difference in the aggregation process.

Larry Phillips, one of the collaborators in this experiment, was concerned about the fact that no model of the data-generating process was available and so it was not possible to say with certainty whether PIP or POP was more nearly right. So for his Ph.D. thesis he compared PIP with POP in a situation in which a model of the data-generating process was available, it was meaningful to ask for a likelihood ratio estimate for a single datum, and the POP procedure produced conservative estimates.

His subjects were the editors of the University of Michigan student newspaper. He took each editor's editorials for a semester, counted the first two letters and the last two letters of each word of each editorial, and thus for each editor prepared a bookbag full of beginning bigrams and a bookbag full of ending bigrams. For the PIP task, he took certain bigrams, and asked an editor to estimate (for his own bookbags only) the likelihood ratio, taken with the beginning-bag hypothesis in the numerator and the ending-bag hypothesis in the denominator, associated with each bigram. For the POP task, he prepared a sequence of bigrams sampled from one of the bookbags, and asked the editor, as he worked through the sequence, to estimate the posterior odds that it was the beginning, not the ending, bag being sampled from. Much care was devoted to preliminary training of the editors, and likelihood ratio estimates were collected twice, once before and once after posterior odds estimates.

A problem in data analysis arose because all judgments, for both PIP and POP, were biased in favor of the beginning bag. This is probably because it is much easier, for example, to think of words that begin with *re* than to think of words that end in *re*, even though *re* is more common as an ending than as a beginning; we are accustomed to tagging words by their beginnings, not endings, when we, for example, look them up in a dictionary. However, it is possible to correct for such biases. Figure 4 shows the results after such a correction. The veridical odds, calculated from the actual bigram counts, are most extreme. Next come the odds calculated from the second set of likelihood ratio estimates. Next come the odds calculated from the first set of likelihood ratio estimates. And, closest to the middle and therefore most conservative, are the directly estimated posterior odds. If we believe these data (and I do), though PIP is considerably less conservative than POP, it is still too conservative - but PIP estimates improve with practice.