Reply

Bayes's Theorem and Reliability: A Reply to Levin*

DAVID SHERRY Northern Arizona University

§1 Misusing Bayes's Theorem

Psychological research appears to show that people are bad at inductive reasoning. In a well known experiment due to Kahneman and Tversky,[1] subjects are told that a cab was involved in an accident in a city in which 15% of the cabs are blue and 85% are green, and that a witness testified that the cab was blue. Kahneman and Tversky ask subjects to estimate the probability that the testimony is correct given that the witness can reliably identify blue and green cabs 80% of the time. The majority say the probability that the offending cab was blue, given the testimony, is around .8. Using Bayes's theorem, the experimenters calculate the probability to be .41. Levin is not the first to argue that this experiment does not show that people are bad at inductive reasoning. But he is the first to argue that the experiment fails because the Bayesian calculation involves a mistaken analysis of reliability. Further, he contends that given the correct analysis of reliability, the probability that the witness's testimony is correct is indeed .8.

The court tests a witness's reliability by showing him a series of blue and green cabs in conditions similar to those in which the accident occurred. As Levin puts it, the test shows the witness is "right 80% of the time about when a cab is a Blue and when it is a Green" (63). According to the Bayesians the witness's reliability is

(1) P(witness testifies blue/cab was blue) = .8.

In evaluating the testimony, however, Bayesians must determine

(2) P(cab was blue/witness testifies blue),which requires Bayes's theorem as well as the prior probability that the cab was blue (.15)[2], the prior probability that the cab was green (.85), and the probability that the witness testified the cab was blue even though it was green (.2). By Bayes's theorem, (2) is

$$\frac{P(B)P(W_B/B)}{P(B)P(W_B/B)+P(\sim B)P(W_B/\sim B)} = \frac{(.15)(.8)}{(.15)(.8)+(.85)(.2)} = .41.$$

Levin contends the proper analysis of the witness's reliability is not the Bayesians' (1), but their (2):

© *Informal Logic* Vol. 25, No. 2 (2005): pp. 167-177.

(2) P(cab was blue/witness testifies blue),

which, he claims, has already been stipulated to be .8. Moreover, since Levin agrees that (2) gives the probability that the witness's testimony is correct, the stipulation indicates how reliable this particular testimony is, and there is no need of Bayes's theorem. While Levin grants that

> ... a phrase such as "the probability that Witness was right about the cab being blue" is doubtless somewhat vague, and there may be no such thing as *the* idea it conveys (64),

he defends the choice of (2) on the grounds that it is "what is normally meant" (*ibid.*). He also argues that Bayesian reasoning yields the absurdity that a marksman's "chances of a bull's eye ... will improve if the rest of his regiment improves their aim" (65). Against Levin I urge that phrases like "the probability that witness was right about the cab being blue" have different senses, depending on context, and that, in fact, the Bayesian analysis helps us to appreciate the different senses (§2). Next I show that Levin is unable to reduce the Bayesian analysis to absurdity (§3). Finally I suggest an account of the experimental result which excuses without justifying subjects who think the proportions of green cabs and blue cabs, i.e., the base rates, are irrelevant (§4).

§2 Reliability

What is meant by saying, as Levin does, "Mark Marksman's reliability with a rifle is .55"? A natural way to explicate this statement is to say that for every 100 shots M takes, 55 strike the bull's eye. That is,

(3) P(M hits the bull's eye) = .55.

(3) is a categorical probability. Like a statement of batting average, (3) reflects the ratio of successes to attempts. It is acceptable, but misleading to express (3) as a conditional probability, say,

P(M hits the bull's eye/M attempts to hit the bull's eye) = .55.

A conditional probability assigns a probability to one event on the supposition that another event has occurred. For example, our estimate of M's reliability may change if we learn that M is shooting in specific circumstances, say from 25 yards or 50 yards.

P(M hits the bull's eye/M shoots from 25 yards)

may be different from

P(M hits the bull's eye)

and

P(M hits the bull's eye/M shoots from 50 yards).

While learning that M was attempting to hit the bull's eye could, in odd circumstances, influence our estimate of the probability of M hitting the bull's eye, it is not, like learning that M is shooting from 25 yards, a consideration that would ordinarily be relevant to determining a marksman's reliability. Just as a baseball statistician presupposes that the batter is attempting to get a hit,[3] so someone measuring the reliability of a marksman presupposes that the marksman is attempting to hit the bull's eye. Likewise, someone measuring the reliability of our witness would presuppose that he is attempting to identify a cab's color. §4 explains why a categorical probability is inadequate to explicate the reliability of the witness in the cab experiment. For now, I presume that the witness's reliability should be represented as a conditional probability, as both Levin and the Bayesians agree.

When he's being tested by the court, the witness is in a position like a marksman who shoots sometimes from 25 yards and sometimes from 50; the witness testifies sometimes in the presence of a blue cab and sometimes in the presence of a green cab. Which conditional probability, $P(B/W_B)$ or $P(W_B/B)$, is appropriate to express the witness's reliability as tested by the court? If we are measuring the marksman's reliability at different distances, then the conditional probabilities are

P(M hits the bull's eye/M shoots from 25 yards)

and

P(M hits the bull's eye/M shoots from 50 yards).

Reliability at 25 yards (50 yards) is properly measured by the ratio of bull's eyes from 25 yards (50 yards) to total attempts from 25 yards (50 yards). The inverses are ruled out because they are measured by the ratio of bull's eyes scored from 25 yards (50 yards) to total bull's eyes scored from either distance, and the frequency of a bull's eye from 50 yards presumably has no bearing on the marksman's reliability from 25.

In analogy with the representation of a marksman's reliability, the condition under which the witness's testimony takes place ought to follow the slash (/). That is, the measure of the witness's reliability ought to be

(1) P(Witness testifies blue/cab was blue).

Let us call the quantity which the court measures "reliability$_W$." It measures the witness's success at identifying a cab's color when it is known what color cab he is attempting to identify.

Levin objects to the preceding analysis.

> "Reliability" should be explicated so as to preserve the apparent truism that someone equally reliable at two tasks—such as ... identifying cabs of different colors—is equally likely to succeed at both. This principle is violated by the "Bayesian" analysis I have criticized. (65)

The truism is violated, Levin continues, because it entails that

> ... the cab is more likely to have been Green if witness says Green than to have been Blue if Witness says Blue. (66)

The Bayesian analysis does indeed entail that $P(G/W_G)>>P(B/W_B)$, and this does mean that the witness would be more likely to be correct if he had testified "green" rather than "blue." But these consequences do not violate the truism; P(G/

W_G)>>P(B/W_B) does not say that the witness is far more likely to succeed at identifying a Green cab than at identifying a Blue cab.

The jurors know that the witness testified the cab was blue, and they also know how good the witness is at identifying green cabs and identifying blue cabs, i.e., how reliable$_W$ the witness is. What neither they nor the witness know is which task—identifying a green cab or identifying a blue cab—the witness was undertaking the night of the accident. The jurors know which task the witness was attempting that night if and only if they know the color of the errant cab. Keeping this equivalence in mind is valuable for understanding what P(G/W_G) and P(B/W_B) mean. In view of the equivalence, the jurors are being asked to estimate the likelihood that the witness was engaged in one task rather than another the night of the accident. Could a reasonable juror estimate the probability that the witness was engaged in one task rather than another *prior* to hearing the witness's testimony? Of course. For instance, in the absence of *any* evidence, the juror could judge that it was as likely as not that the witness was attempting to identity a blue cab. Likewise, in the presence of evidence other than the witness's testimony, a juror could assign unequal probabilities. Suppose he hears that blue cabs are more accident prone than green taxis, or that blue cabs frequent that part of town more than green cabs. In such circumstances, the juror would presumably judge it more likely than not that the witness was attempting to identify a blue cab. Such a judgment, a prior probability, guides the jury in deciding how much weight to place upon the witness's testimony. If, prior to the testimony, the juror thinks it likely the witness was attempting to identify a blue cab, then the testimony will carry more weight than if the juror thinks it likely the witness was attempting to identify a green cab. The probabilities computed by Bayes's theorem, P(G/W_G) and P(B/W_B), indicate the weight the jurors place on the testimony in view of the prior probabilities they assign to each of the tasks the witness could have attempted. Thus, contrary to Levin's suggestion, P(G/W_G)>>P(B/W_B) means that in view of the evidence—including the testimony and any evidence that informs the prior probability—it is more likely that the witness was attempting to identify a green cab if he says "green" than that he was trying to identify a blue cab if he says "blue." P(G/W_G) and P(B/W_B) are also measures of reliability; let us call it "reliability$_T$." Reliability$_T$ tells jurors how reliable the testimony is when they don't know whether the witness was attempting to identify a green cab or a blue cab when he testified. Reliability$_T$ differs from reliability$_W$, the reliability which the court tests.The court does not test the reliability of the *particular* testimony; rather, the court tests how reliably the witness identifies blue cabs and how reliably he identifies green cabs. The notation of conditional probability helps us to keep this distinction before our minds, just as the quantifier notation enables us to keep the distinction between "all actions aim at some end" and "there is some end at which all actions aim" before our minds.

Levin blurs together the two senses of "reliability" when he writes

> I suggest the culprit here is much more easily identified: it lies in initially explicating "the probability of Witness being right about the cab being blue" as P(w/h) [i.e., our P(W_B/B)], and, generally, in taking the probability of someone's being right about a world-state to be the probability of his saying that that state obtains given that it does. (64)

There are two senses in which the witness can be right about the cab being blue. On the one hand, "the cab" can refer to the cab which the witness saw involved in an accident. The probability of the witness being right about *that* cab being blue is P(B/W_B), not P(W_B/B) as Levin states. That assessment is complicated because it must be made without the court knowing the color of the cab the witness is trying to identify; thus it is a measure of reliability$_T$. On the other hand, "the cab" can refer to one of the cabs in the test performed by the court. In the Bayesian analysis the probability of the witness being right about one of *those* cabs being blue *is* P(W_B/B), the probability of someone's saying that a world-state obtains given that it does. When that assessment is made the court *does* know the color of the cab the witness is trying to identify.

Levin is perhaps aware of the distinction being drawn here. In his penultimate paragraph he writes

> What we *are* discussing, when Bayes' Theorem comes into play, is the cab's likely color when we do *not* know the probability that a cab is the color Witness says it is. Background information, including base rates, then becomes pertinent. (66)

But surely this is a description of the case at hand. We don't know the probability the cab is the color the witness says it is because we don't know which task the witness is attempting when he testifies. Levin urges that in such cases "we use the descriptor "reliability" of the conditional probability we are calculating [P(B/W_B)], not the converse conditional probability [P(W_B/B)]" (*ibid*). Yet doing so ignores a straightforward sense in which a witness can be reliable, a sense analogous to a marksman's being reliable.

Subsequently Levin distinguishes cases for which Bayes's theorem is relevant, from the cab problem. The former, he says,

> ...involve an indicator of unknown trustworthiness. We know the odds that a subject with clogged arteries will feel fatigue, and the odds that a subject with normal arteries will feel fatigue. What we would *like* to know is the specificity of fatigue, the probability that someone feeling fatigue has clogged arteries. In such cases we should not say we know how well fatigue predicts clogged arteries. Did we know that, further information would be superfluous. (*Ibid.*)

This case is quite analogous to the cab problem. Even if we know the odds that a subject with clogged arteries will feel fatigue, we can not say how well fatigue predicts clogged arteries (P(clogged arteries/fatigue)) without a prior probability that the patient has clogged arteries. By the same token, even if we know the odds that a witness who is shown a cab of a certain color will identify the cab as having that color, we can not say how well his identification predicts a cab of that color without a prior probability that the cab he was trying to identify was that color. The court assumes, perhaps unjustifiably, that the actual attempt to identify the offending cab's color (which is reliable$_T$) is sufficiently like the staged attempts (which are reliable$_W$) that the latter can serve as a model of the former. But there are two sorts of staged attempts, and even though they have the same degrees of success, the fact that it is unknown which the witness was attempting when he testified 'blue" means the court cannot employ the results of its test without further evidence.

§3 Absurdity?

Levin claims that applying the Bayesian analysis to an analogous case leads to absurdity.

> If the odds of Marksman's hitting the bull's eye the next time he tries are calculated as the standard analysis calculates the odds of the cab Witness saw being Blue, a few reasonable assumptions show that he will almost surely miss. For if m is "Marksman shoots" and s is "A bull's eye is scored," let "Marksman hits the bull's eye 55% of the time" be interpreted as P(m/s)=.55, i.e. that Marksman is the shooter 55 times out of every 100 times a bull's eye is scored. Finally, suppose the rest of Marksman's regiment are such poor shots that the regimental average is .2. In other words, the background probability that a shot will hit the bull's eye, or P(s), is .2. The probability that Marksman will hit the bull's eye the next time he shoots at it, explicated as P(s/m) is then ... a feeble .23. (65)

He has correctly calculated *some* probability by the Bayesian method, and I grant that if he has calculated the probability of Marksman's success it follows that his aim would improve if the rest of the regiment improved its aim. But presuming we understand reliability in the manner of the Bayesians, Levin has not calculated the probability that Marksman scores a bull's eye.

The failure to calculate the probability that Marksman scores a bull's eye is obscured by Levin's use of the passive construction, "a bull's eye is scored." The phrase is ambiguous between "a bull's eye is scored by Marksman" and "a bull's eye is scored by some member of the regiment." The former is suggested by "the odds of hitting the bull's eye the next time *he* tries." But the latter must be what is meant, not only because the background probability, P(s), is the *regimental* average, but also because P(Marksman shoots/a bull's eye is scored by Marksman) must be 1, not .55. Hence, the probability Levin calculates is

(5) P(a bull's eye is scored by some member of the regiment/
Marksman shoots).

For a Bayesian, (5) addresses a different issue from the probability of Marksman's success, viz., the probability that someone or other hits the target given that Marksman is also shooting. This issue might concern someone interested in the psychological effect of shooting alongside a reliable marksman, but it is not a measure of Marksman's reliability in either of the senses discussed in §2. Outside of the psychologist's concern, there is no need for a conditional probability to represent the probability that Marksman scores a bull's eye. He is not firing from different distances or under different conditions, and so a categorical probability is appropriate. But plainly the absurdity can't be gotten from the categorical probability. If we strain usage and treat the probability that Marksman scores a bull's eye as

P(a bull's eye is scored by Marksman/Marksman shoots),

the paradox would still not be derivable, since the aim of the rest of the regiment is out of the picture.

Just as P(s/m) is not, for a Bayesian, the probability that Marksman scores a bull's eye, neither is the measure of reliability which Levin attributes to the Bayesians,

(6) P(Marksman shoots/a bull's eye is scored by some member of the regiment),

a measure of Marksman's reliability.[4] Consider what question (6) answers. The regiment is firing away, and, bad shots that they are, the curiosity of the range sergeant is aroused when a member of regiment finally scores a bull's eye. Knowing that Marksman is the only good shot in the bunch he hypothesizes: Marksman must have shot. Then he wonders, "What is the probability that Marksman was the shooter given that a bull's eye was scored by some member of the regiment?" That is the question which (6) answers, *not* "how reliable a shot is Marksman?" (6) is tailor made for Bayes's theorem, but we would need to know the probability that a given shot was Marksman's, P(m), and the probability a given shot was from someone else, P(~m), to apply Bayes's theorem. Thus no Bayesian would claim that P(m/s)=.55 on the grounds that Marksman hits the bull's eye 55% of the time.

§4 Categorical Probability

If the preceding problem calls for a categorical probability, then why not represent the witness's reliability with a categorical probability? Representing a degree of reliability by means of a categorical probability presumes that at each trial the subject is attempting the *same* task; in the cab experiment this task is identifying cab color. With respect to *this* task, the witness is indeed 80% reliable. But the matter is not so simple as it appears. The witness can demonstrate 80% reliability at identifying cab color in different ways. As in the cab experiment, he can be 80% reliable at identifying cab color and *equally reliable* at identifying green cabs and blue cabs. However, he can also be 80% reliable at identifying cab color and yet better at identifying one color cab then another. Let's consider these cases separately.

In the lingo of probability, correctly identifying cab color is a *compound event*, for there are distinct ways it can happen:

Either the cab presented is green (G) and the witness says "green" (W_G) *or* the cab presented is blue (B) and the subject says "blue" (W_B).

Let C be the event of correctly identifying cab color. Since cab colors are (in this case) mutually exclusive and exhaustive,

$P(C) = P(G\&W_G)+P(B\&W_B)$.

By the general conjunction rule,

$P(C) = P(G)P(W_G/G)+P(B)P(W_B/B)$,

Where $P(W_G/G)$ and $P(W_B/B)$ are measures of reliability$_W$. If they are both equal to p, then $P(C) = P(G)p+P(B)p = p(P(G)+P(B)) = p$, because G and B are mutually exclusive and exhaustive and so $P(G)+P(B) = 1$. Thus, if the witness is equally reliable at identifying green cabs and blue cabs, his reliability at identifying cab color will *not* be affected by the proportions of green cabs and blue cabs he is shown, or the likelihood he is shown a green cab or a blue cab.

When the witness is *not* equally reliable at identifying blue cabs and green cabs, it is not so easy to determine his reliability at identifying cab color. Suppose the witness is only 50% successful at identifying blue cabs but 85% successful at identifying green cabs. In this case, the witness's reliability at identifying cab color will depend upon the relative proportions of blue and green cabs presented in the test. If he is shown 15% blue cabs and 85% green cabs, then

$P(C) = (.15)(.5)+(.85)(.85)H \approx .80$.

But if he is shown equal numbers of blue cabs and green cabs,

$P(C) = (.5)(.5)+(.5)(.85) = .675$.

In general, by adjusting the proportions of blue cabs and green cabs shown to the witness it is possible to obtain any probability between his different degrees of reliability at identifying blue cabs and green cabs. This is important even though in the cab experiment the witness is equally reliable at identifying blue cabs and green cabs.

The possibility of manipulating the measure of the witness' reliability in the preceding manner demonstrates that the propensity to identify cab color correctly is not a genuine characteristic of the witness. A genuine characteristic should be sufficiently stable that a change in the test conditions with no apparent causal influence on the witness's perceptual apparatus does not affect the measure of the reliability. Underlying the pseudo-propensity are two *genuine* propensities, the propensity to respond "green" in the presence of a green cab and the propensity to respond "blue" in the presence of a blue cab. Their measures can be manipulated, but only by manipulating the propensity itself, for example, by changing lighting conditions. If jurors (or psychological subjects) understood the witness's reliability simply as a proportion of successful attempt to identify cab color rather than as a complex involving distinct propensities to identify different colors, then they wouldn't have appreciated that a pair of propensities is necessary to solve the problem. Even less would they have appreciated that the witness's reliability$_W$ has to be represented by a pair of conditional probabilities. Without a pair of conditional probabilities there is no way to connect a single testimony (W_B) with the different conditions under which that testimony can occur, W_B/B and W_B/G. By representing the situation as the witness's exercising a single ability to identify cab color, subjects pass over the fact that no one knows which of two propensities was exercised by the witness. Thus, subjects recognize no need to consider prior probabilities and ignore the base rates.

The possibility that subjects might have construed the witness's reliability as a reflection of a single pseudo-propensity is perhaps behind Kahneman and Tversky's modification to the original cab experiment. The earlier version informed subjects that

> When presented with a sample of cabs (half of which were.Blue and half of which were Green) the witness made correct identifications in 80% of the cases and erred in 20% of the cases. (1977, 174)

But the later version stated

> The court ... concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. (1982,156).

The earlier version, but not the later, is compatible with the witness's being more reliable at identifying one color cab than another. For example, the subject could be 60% reliable at identifying blue cabs but 100% reliable at identifying green cabs. Moreover, the 80% figure in the earlier version depends upon equal numbers of blue and green cabs being shown, while there is no such dependence in the later version. Thus, 80% in the earlier version has all the marks of a categorical probability, which, as we've seen, is apt to hide from jurors the complexity involved in estimating the reliability of the witness's testimony.

The modified experiment gives the same result in spite of the attempt to make it clear that two conditional probabilities rather than a single categorical probability are to be considered. Yet, it would not be surprising if subjects continued to treat the witness's reliability as an indication of his reliability at identifying cab color rather than as an elliptical statement for a pair of measures of reliability. They are tempted to do this by the fact that the subject is equally reliable at either task. I suspect that this would be less likely to happen if the subjects were told that the witness had different rates of success at identifying blue cabs and green cabs; for different rates would prevent the subjects from jumping to the pseudo-propensity to identify cab color correctly and so make it more likely that they pay attention to base rates and other relevant evidence.[5]

The cab experiment is like an elementary, but tricky problem in arithmetic. A car travels a 10 mile route at 40 miles per hour and returns by the same route at 60 miles per hour. What is the average speed? It is tempting to answer "50 miles per hour," but doing so ignores the subtlety that the car spends 5 minutes longer on the first leg of the trip then on the second leg; thus 20 miles is covered in 25 minutes, which is an average of 48 miles per hour. Half the distance is covered at 40 mph but it's not true that half the time is spent traveling 40 mph. The differing time intervals that have to be recognized to solve this problem are analogous to the distinct, underlying propensities which are necessary to solve the cab problem. In both cases, stating the problem in a way that makes the underlying conditions explicit would, I suspect, improve the subjects' responses.

We're left to wonder whether Levin would have denied that Bayes's theorem applies to the cab problem if Kahneman and Tversky had stipulated that the witness was better at identifying one color cab than the other.[6]

Appendix

Kahneman and Tversky give their subjects only one piece of information prior to the witness's testimony: 85% of the cabs in the city are Green and 15% are Blue. From this information the subject is to determine a prior probability, i.e., to judge how likely it is that the witness was trying to identify a Blue cab. 15% is one reasonable estimate of the probability that the witness was trying to identify a blue cab. Some such figure is crucial to the Bayesian method, but nothing in that method dictates that the prior probability must be the proportion of blue cabs in the city. Thus it would not be unreasonable for a juror to take that evidence into account by assigning a lower probability to blue than green but nonetheless a probability greater than .15. The proportion of green cabs is one piece of evidence, but its relevance, like the relevance of the witness's testimony, must be gauged in light of all the available evidence. This includes any of the subject's prior beliefs. Suppose that in the subject's experience small cab companies are more apt to have unqualified drivers; that would suggest that blue cabs are more accident prone. Subjective differences like these don't bother Bayesians, because they know that subjects with different prior probabilities will converge to the same posterior probabilities given *enough* evidence. The cab experiment presumes that subjects should determine their prior probability solely on the basis of the proportion of blue cabs. This feature of the experiment offends the spirit of Bayesianism by ignoring that probability judgments occur against a large, but indefinite, body of evidence, viz., our knowledge of the world. Hence, even if Levin is wrong to criticize the Bayesian analysis of reliability, he is justified in thinking that Kahneman and Tversky's experiment has no clear implication for our powers of inductive reasoning.

Notes

[*] Levin, M. 1999 "A Misuse of Bayes' Theorem," *Informal Logic 19*, 63-6.

[1] Tversky, A. and Kahneman, D. 1977 "Causal Thinking in Judgement under Uncertainty," in Basic Problems in Methodology and Linquistics, ed. Butts and Hintikka (Dordrecht: Reidel): 167-190.

[2] The appendix discusses the validity of this figure.

[3] It can happen that the ball strikes the bat accidentally, when, for example, the batter is trying to avoid being hit. If the ball lands fair, the statistician treats the event as an attempting to get a hit, though, strictly speaking, this is not true. Attempting to correct this inaccuracy would complicate the statistics unnecessarily.

[4] Moreover, even if P(m/s) = .55, the complementary likelihood, P(Marksman shoots/~a bull's eye is scored by some member of the regiment), need not be .45 as Levin requires. For it is the probability that Marksman was the shooter, given that no one in the regiment scored a bull's eye; *that* depends on what proportion of the shots Marksman takes. If Marksman does most the shooting, then, even if he's a reliable shot, he could still make the majority of misses.

[5] Base rate neglect disappears if the crucial information is made explicit. See Gigerenzer *et al.* 1989, *The Empire of Chance* (Cambridge: Cambridge University Press), 231.

[6] I received much helpful advice from the journal's referees.

David Sherry
Department of Philosophy
Northern Arizona University
Box 6011
Flagstaff AZ 86011-6011

david.sherry@nau.edu