Journal of Personality and Social Psychology
1992, Vol. 62, No. 5, 739-752

0022-3514/92/$3.00

Naked Statistical Evidence of Liability: Is Subjective Probability Enough?

Gary L. Wells
Iowa State University

Five studies tested the idea that people are reluctant to make proplaintiff liability decisions when the plaintiff's evidence is based on naked statistical evidence alone. Students (n = 740) and experienced trial judges (n = 111) averaged fewer than 10% affirmative decisions of liability when a case was based on naked statistical evidence but averaged over 65% affirmative decisions based on other forms of evidence even though the mathematical and subjective probabilities were the same for both types of evidence. Numerous hypotheses, including causal relevance, linkage to the specific case, and fairness to the defendant proved inadequate to explain the data. For evidence to affect decisions, the evidence must do more than affect people's perceptions about the probabilities associated with the ultimate fact; people seem to require that suppositions regarding the ultimate fact affect their perceptions of the truth or falsity of the evidence.

> Toronto (CP)—An application for child support has been dismissed despite a blood test showing it is 99.8% probable that the man being sued is the father of a four-year-old girl. ("99.8% Probability," 1986).

What is your reaction to this paternity suit? If you are like most psychologists, you are likely to pass it off as another example of people's poor understanding of probabilities. Indeed, innumeracy (Paulos, 1988) is widespread in the population, probability theory is not intuitive (Kahneman, Slovic, & Tversky, 1982), and courts of law are not immune to developing subjective probabilities that deviate profoundly from mathematical probabilities (e.g., Saks & Kidd, 1980; Thompson & Schumann, 1987). This article concerns cases in which the probability information is easily and intuitively processed, such that the mathematically correct probabilities and the subjective probabilities agree, and yet, the subjective probabilities do not seem to mediate people's verdict decisions. Hence, the current article differs from most previous studies of people's use of statistical evidence in legal decision making. Previous work has been concerned almost exclusively with the problem of people deriving subjective probabilities that deviate from the statistically correct answer (e.g., Saks & Kidd, 1980; Thompson & Schumann, 1987). In the current article, on the other hand, the emphasis is on an as yet unexplored problem wherein subjective probabilities are congruent with statistically correct answers but the statistical evidence does not have an impact on people's decisions.

There are three main purposes of this research, each of which is related to the other two. First, this research is designed to illustrate the incompleteness of current psychological theories regarding how people evaluate trial evidence. In particular, the reader should consider what these simple experiments indicate about the dominant decision model put forth in the psycholegal literature, namely the probability–threshold model. Although expressed in many forms (e.g., Bayesian, cascaded inference, Poisson stochastic) and subject to some caveats, the common assumption is that a trier of fact makes an affirmative decision such as guilt or liability when the subjective probability of guilt or liability exceeds a threshold probability (e.g., Carlson & Dulaney, 1988; Connolly, 1987; Dane, 1985; Fried, Kaplan, & Klein, 1975; Kagehiro & Stanton, 1985; M. F. Kaplan, 1977; Kerr et al., 1976; Marshall & Wise, 1975; Nagel, 1979; Ostrom, Werner, & Saks, 1978; Simon, 1970; Thomas & Hogue, 1976).

This is not to suggest that the probability–threshold model is blind, totally mechanistic, or without caveat. Illegally obtained evidence, for example, could drive subjective probabilities to very high, beyond threshold levels and yet not produce affirmative (liable or guilty) verdicts because the triers of fact might have fundamental objections to the process by which the evidence was obtained. As well, the probability–threshold model does not assume the threshold to be a constant value. Kerr (1978), for example, has shown that people will consider the seriousness of the offense and the punishment in deciding how much evidence would be needed to produce a guilty verdict. Nevertheless, the probability–threshold model is heavily interwoven into the psycholegal literature and represents one of the cornerstones of the judgment and decision process by which people make verdict decisions.

The second purpose of this work is to establish an empirical foundation for evaluating some hypotheses put forward by legal scholars regarding why people resist returning affirmative verdicts when the evidence is *naked statistics.* Naked statistical evidence is ill defined in the legal literature but typically refers to probabilities that are not case specific in the sense that the evidence was not created by the event in question but rather existed prior to or independently of the particular case being tried.

The third purpose of this article is to provide an alternative to

I thank Richard Lempert, Edward Wright, Teddy Warner, and the anonymous reviewers for their comments on a draft of this article.

Correspondence concerning this article should be addressed to Gary L. Wells, Department of Psychology, W112 Lagomarcino, Iowa State University, Ames, Iowa 50011-3180.

or a refinement of the probability–threshold account of the process by which people reach affirmative verdicts. This third purpose is subordinate to the first two purposes because the proposed refinement of the subjective probability–threshold model is clearly underdeveloped and overly simplistic. Nevertheless, some kind of modification to the probability–threshold model seems necessary and the one proposed here might be a reasonable starting point.

The Balance of Probability Problem

Legal scholars have debated a hypothetical liability suit that serves as fodder for the current experiments. In its general form, it is said that a bus company accounts for a certain portion of all business in the area (e.g., 80%). An accident occurs in which it is known that a bus was at fault, but the specific company is not known. Because the Blue Bus Company accounts for most of the business (say 80%), the balance of probability (in this case 80%) clearly favors the idea that a Blue Bus Company bus caused this accident. According to the general rule of civil litigation for suits of this type, a court should rule for the plaintiff (i.e., against the Blue Bus Company) if "after considering all the evidence in the case, the jurors believe that what is sought to be proved on that issue is more likely true than not true" or "more likely so than not so" (Nesson, 1985, p. 1364). Indeed, although there has been and continues to be some disagreement over the interpretation that civil proof represents anything over a .50 probability, the .50 probability conceptualization has gained wide acceptance (see McCormick, 1984; Simon, 1969).

The problem that has plagued the courts and legal scholars is that they at once endorse the balance of probability criterion for such suits and also refuse to rule in favor of the plaintiff when such evidence is presented (e.g., Guenther v. Armstrong Rubber Co., 1969; Smith v. Rapid Transit, 1945). Indeed, suits based on naked statistics of this sort are usually thrown out by a summary judgment. In other words, the Blue Bus Company (the defendant) wins because the case never reaches a jury even though the balance of probability says that a Blue Bus Company bus was at fault. The rationale of legal scholars for reconciling the balance of probability criterion in civil cases with the common refusal to accept naked statistical evidence of this sort can perhaps enlighten our understanding of the verdict process.

Legal scholars have struggled with this problem in creative ways, usually concluding that the plaintiff should not win the case on grounds of policy. In general, legal scholars argue that to rule in favor of the plaintiff would create other ills in the system of justice and cause more damage than good. Thompson (1989), for example, argued that findings for the plaintiff on the basis of mere base-rate statistics might lead to a strategy of suppression of particularistic proof by the party favored by the base rate. Hence, Thompson argued, the use of mere base-rate evidence should be prohibited on policy grounds. Indeed, trials serve purposes other than verdict accuracy, and the desirability of using mere base rates depends on the relative value of verdict accuracy concerns versus various policy concerns (Koehler & Shaviro, 1990). Among those policy concerns might be such notions as the "acceptability" of a verdict (Nesson, 1985), the need to not waste a jury's time on a case that cannot be won, the need for incentives to uncover more convincing evidence, and concerns about people being perceived as liable merely because they fall into a particular class or category. Policy concerns of this sort have been discussed eloquently by a number of legal scholars (e.g., Allen, 1986; Ball, 1961; Brook, 1985; Callen, 1982; Cohen, 1977, 1981, 1986; Kaye, 1979, 1982, 1986; Koehler & Shaviro, 1990; Lempert, 1977, 1986; Nesson, 1985, 1986; Tribe, 1971).

In this article, various possible bases for people's reluctance to rule against the Blue Bus Company are tested. Although these experiments use simple hypothetical cases, this preliminary inquiry proves to be highly informative. The robust effects observed with these simple problems, generalized across three uniquely different populations of respondents, cannot be easily dismissed. These data yield strong implications for our understanding of how people reach judgments of liability that seem to call for a revision of the dominant model in psychology for such decisions, namely the probability–threshold model.

The experiments described in this article are meant primarily to address the question of how and why people react in certain ways to statistically based evidence with implications for psychological models of such processes. Nevertheless, there might also be implications of the current work for legal scholars' conceptions of how the trial process should operate and the role of mathematics in the fact-finding process. There is widespread and lively interest among legal scholars in this type of problem and "nowhere is the concern for proof more central than in that body of scholarship which seeks to build on or criticize mathematical models as modes of proof or as a means of understanding trial processes" (Lempert, 1986, p. 440). The blue bus case and its analogous counterpart, the paradox of the gate-crasher, have been the staple of hypotheticals around which extensive legal scholarly debate on probabilistic evidence has taken place (e.g., Allen, 1986; Ball, 1961; Brook, 1985; Callen, 1982; Cohen, 1977, 1981, 1986; Finkelstein & Fairley, 1970; J. Kaplan, 1968; Kaye, 1979, 1982, 1986; Koehler & Shaviro, 1990; Nesson, 1985, 1986; Schum, 1986). Although the blue bus case is an overly simplistic and unusual hypothetical case, its value in this article is similar to the value it holds among legal scholars, namely, it drives to the heart of the questions of proof, sufficiency, and process.

In the four experiments that follow, various versions of the blue bus case were tested with different populations of subjects. Because no prior data have been reported using the blue bus case, the initial questions and objectives were quite basic. Are people reluctant to rule against the Blue Bus Company when the evidence is confined to naked probabilities? Do their subjective probabilities coincide with the mathematically correct odds? Are naked probabilities confined to mere base rates or are case-specific odds also forms of naked probability and hence resisted as evidence? Does other evidence that produces the same subjective probability yield the same verdict?

Subjective Versus Mathematical Probabilities

Tribe (1971), a leading legal scholar, has attempted to solve the blue bus problem by noting that verdicts are based on *subjective* probabilities rather than the actual mathematical odds.

Although the plaintiff's objective proof indicates an 80% likelihood that the defendant's bus caused the injury, Tribe notes, a juror is not bound to accept this probability, believing perhaps that the probability is 50%. This is perhaps the easiest argument to test empirically. In the first experiment, some subjects were asked to provide subjective probabilities for the blue bus case to assess Tribe's hypothesis.

Before entertaining Tribe's hypothesis, however, there is something even more basic that needs to be studied, specifically, whether people are in fact reluctant to rule against the Blue Bus Company when given the proportion-of-business evidence. Somewhat surprisingly, in spite of the healthy debate among legal scholars spanning many years, no one has yet collected data bearing on the question of whether people are in fact reluctant to rule against the Blue Bus Company. Instead, courts have commonly rendered directed verdicts, in effect throwing the plaintiff's case out of court (see, e.g., Guenther v. Armstrong Rubber Co., 1969; Smith v. Rapid Transit, Inc., 1945) on any of a series of policy bases. Hence, it is not clear how people would react to evidence of this sort.

In addition, there needs to be some kind of comparison case to make certain that people's reluctance to rule against the Blue Bus Company is not possibly attributable to their misunderstanding of the balance of probability rule or general reluctance to rule against defendants. For example, people might be using a threshold beyond .50, perhaps even beyond .80, for their liability threshold, which would account for their reluctance to rule against the defendant even when their subjective probabilities are .80. Indeed, there are three standards of proof used in trials, the balance of probability standard (used in most civil trials), the beyond a reasonable doubt standard (used in criminal trials), and clear and convincing evidence (an intermediate standard used in cases such as civil commitment and deportation). Because of the difficulties of quantifying people's thresholds for these standards of proof (e.g., different methods produce different results, see Dane, 1985), a decision was made to create a comparison case that yielded the same subjective probability but was not based on naked probability evidence. If people's reluctance in the blue bus case stems from the nature of the evidence rather than their threshold for proof, there should be a critical comparison case that yields the same subjective probability but affirmative (i.e., for the plaintiff) findings.

Experiment 1

The standard (*volume-of-traffic*) version of the blue bus case was presented as follows (adapted from Nesson, 1985):

> Mrs. Prob is suing the Blue Bus Company for having caused the death of her dog. At trial, the following evidence was given:
>
> Mrs. Prob testified that she was walking her dog on county road #37 when she heard a large vehicle behind her. She turned around and saw a bus swerving recklessly down the road. She jumped out of the way but the bus swerved and hit her dog, killing him instantly. The incident occurred at 11:40 A.M. The bus continued at a high speed down the road. Unfortunately, Mrs. Prob is color blind and thus does not know the color of the bus.
>
> A county transportation official took the stand, was sworn as a witness, and testified that there are only two bus companies that travel in the county; the Blue Bus Company and the Grey Bus Company. Each company uses the road to run empty buses back to their stations after dropping off their passengers. Therefore, one of these two bus companies had to be responsible for the death of Mrs. Prob's dog.
>
> A second county transportation official took the stand, was sworn as a witness, and reported that the Blue Bus Company owned 80% of all the buses and that 80% of the county road #37 bus traffic was from the Blue Bus Company. The Grey Bus Company owned 20% of all the buses and accounted for 20% of the traffic on that road. Mrs. Prob's attorney argued that the jury must find the Blue Bus Company liable for damages because on the balance of evidence, it was a Blue Bus Company bus that killed Mrs. Prob's dog.

The standard version was given to 80 undergraduate psychology students. Half were asked to render a verdict ("If you were a juror in this case, would you rule against the Blue Bus Company and force them to pay damages to Mrs. Prob?") and half were asked to estimate "the probability that a bus from the Blue Bus Company killed Mrs. Prob's dog."

A comparison version was created that was designed to produce the same mathematical and subjective probabilities as the standard version. In this case, however, an error-prone eyewitness account from a weigh-station attendant was used as the evidence. Note that the bus companies have an equal proportion of the traffic in this version:

> Mrs. Prob is suing the Blue Bus Company for having caused the death of her dog. At trial, the following evidence was given:
>
> Mrs. Prob testified that she was walking her dog on county road #37 when she heard a large vehicle behind her. She turned around and saw a bus swerving recklessly down the road. She jumped out of the way but the bus swerved and hit her dog, killing him instantly. The incident occurred at 11:40 A.M. The bus continued at a high speed down the road. Unfortunately, Mrs. Prob is color blind and thus does not know the color of the bus.
>
> A county transportation official took the stand, was sworn as a witness, and testified that there are only two bus companies that travel in the county; the Blue Bus Company and the Grey Bus Company, each of which has an equal share of the bus traffic on that road. Each company uses the road to run empty buses back to their stations after dropping off their passengers. Therefore, one of these two bus companies had to be responsible for the death of Mrs. Prob's dog.
>
> A second county transportation official took the stand, was sworn as a witness, and reported that he was on duty as the weigh attendant the day of the bus–dog incident. He explained that all vehicles with more than two axles (such as buses) must enter a weigh station and drive slowly over a set of scales. As they drive over, the weigh attendant notes their weight and jots down a two-word description of the vehicle on the log book. In the weigh attendant's log book for the day in question, he had entered "blue bus, 11:30 A.M." along with a weight. The dog was hit at 11:40 and the distance from the weigh station to the point where Mrs. Prob's dog was killed is about a 10 minute drive.
>
> The defense attorney for the Blue Bus Company recalled the weigh station attendant and entered evidence showing that his previous log book entries were correct only 80% of the time and wrong 20% of the time. This was proven by records seized after the alleged incident. These records showed that 20% of the time in which a blue bus was weighed the attendant wrote down "grey bus" and 20% of the time that a grey bus was weighed the attendant wrote down "blue bus."
>
> Mrs. Prob's attorney argued that the jury must find the Blue Bus Company liable for damages because in all likelihood it was a Blue Bus Company bus that killed Mrs. Prob's dog.

This version, hereafter called the *weigh-attendant version*, was given to a separate sample of 80 psychology students, half of

whom answered the question regarding liability and half of whom estimated the probability.

The results of this first experiment are shown in Figure 1. No differences emerged in the estimated probability that a Blue Bus Company bus caused the death of Mrs. Prob's dog ($p > .50$). The differences in verdict, however, were profound (8.2% vs. 67.1% liable verdicts), χ^2 (1, n = 40) = 30.7, $p < .01$. Furthermore, the subjective probabilities were extremely consistent across both versions, with 85% and 80% of the subjects reporting a .80 probability in the standard and weigh-attendant versions, respectively. The second most common response was a .50 probability (12.5% of respondents), and the remainder reported values between .50 and .80.

These data illustrate three main points. First, there does in fact appear to be a strong reluctance for people to use the volume-of-traffic evidence to rule against the Blue Bus Company. Second, it does not appear that Tribe's (1971) explanation suffices because subjective probabilities do indeed seem to follow reasonably well the mathematically correct probabilities. Third, it does not appear that subjects had an inappropriate threshold for making a judgment of liability because they found against the Blue Bus Company in the weigh-attendant case even though their subjective probabilities were no higher than they were in the standard version. Hence, an apparent anomaly begins to unfold. Why should two versions of evidence that yield equal subjective probabilities produce such profoundly different verdicts?

Experiment 2

Causal Relevance

Is it possible that people are reluctant to rule against the Blue Bus Company in the proportion-of-traffic version because the statistic lacks causal relevance? There are many psychological researchers who have argued that the use of statistical evidence is greater when there is causal relevance to the statistics (e.g., Ajzen, 1977; Borgida & Brekke, 1981; Tversky & Kahneman, 1980, 1982). Accordingly, a new version, hereafter called the *rate-of-accidents version,* was created by changing the third paragraph in the standard version to indicate that the Blue Bus Company was responsible for 80% of the accidents rather than 80% of the traffic. This seems to establish causal relevance (Tversky & Kahneman, 1980). The new paragraph read as follows:

> A second county transportation official took the stand and reported that the Blue Bus Company was responsible for 80% of all the accidents involving buses in the county and on thus particular road. The Grey Bus Company was responsible for 20% of all the accidents involving buses in the county and on this road.

Eighty psychology students read this version and 80 read the weigh-attendant version. This time, however, all were asked both to render a verdict and to estimate the probability (as opposed to answering only one of the two). Half of the subjects rendered a verdict first and half estimated the probability first.

Order of question had no effect and will not be discussed further. The results are shown in Figure 2. Once again, subjective probabilities did not differ between the two cases ($p > .50$), whereas robust differences emerged in verdicts, χ^2 (1, n = 160) = 52.4, $p < .001$). Clearly, it did not matter to subjects whether the probabilistic evidence against the Blue Bus Company was based on their rate of accidents or on their market share. Hence, causal relevance does not seem to be the issue in this case. This is not surprising. In other contexts in which causal relevance has been shown to play a role, the problem has been that base-rate information has failed to significantly affect subjective probabilities under conditions in which people must combine base rates with individuating information in a Bayesian fashion. No such problem exists here. The blue bus problem is not what is commonly called a Bayesian problem (see Appendix). The mathematically correct probability is transparently obvious, subjective probabilities are appropriately driven by the statistical evidence, and there is no other evidence (e.g., individuating evidence) to integrate. So, it is a different kind of problem than the type that has been shown to respond to causal relevance.

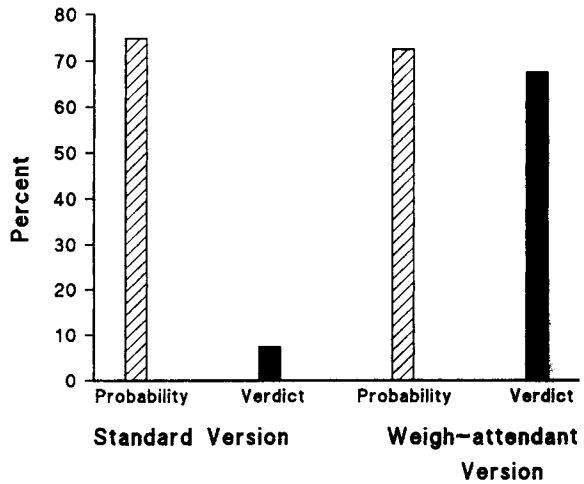

Figure 1. Probability estimates and percentages of liability verdicts for the standard case and the weigh-attendant version.

Fairness and Market Dislocation

In the American legal system, economics has become a powerful force in civil litigation. Indeed, of all the social sciences, none has become so well integrated and influential in civil case law as economics. Thus, it should come as no surprise that an economic argument has been associated with the courts' concerns about naked statistical evidence. Posner (1972) has made an economic argument in the blue bus case. He argues that one must not allow the volume-of-traffic evidence to result in a finding against the Blue Bus Company because it would impose too large a burden on the defendant. In particular, the argument is that, if a court held the Blue Bus Company liable in this case, it would have to hold the company liable in all similar cases even though it was responsible for only 80% of them. Posner argues that this would dislocate the market by disproportionately burdening larger companies and subsidizing their smaller competitors. The smaller companies would then have

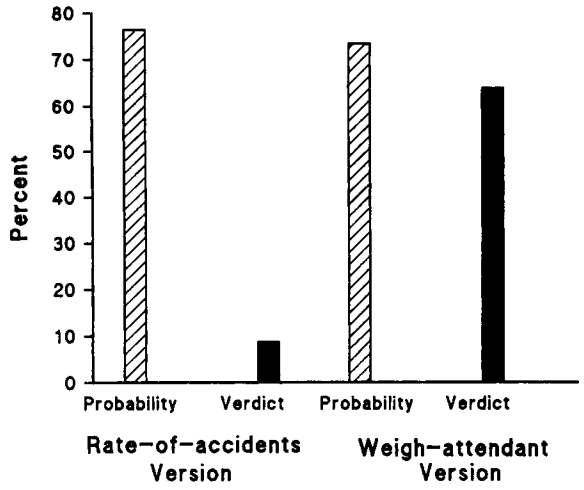

Figure 2. Probability estimates and percentages of liability verdicts for the rate-of-accidents and weigh-attendant versions.

little incentive to promote careful driving by its employees and accident rates would increase.

The rate-of-accidents version partially addresses Posner's (1972) argument. In particular, the rate-of-accidents version does not subsidize smaller competitors but rather subsidizes the company that has a safer driving record. Nevertheless, it remains true that the Blue Bus Company is being required to pay in 100% of the cases whereas it is responsible for only 80% of the accidents.[1]

Equity Theory (Adams, 1965; Walster, Walster, & Berscheid, 1978) tells us that people might have an intuitive understanding of the blue bus case as an equity problem. Of course, finding in favor of the defendant, as more than 70% of the subjects do in the proportion-of-traffic and rate-of-accidents versions, does not represent an equitable solution. After all, the plaintiff's loss remains an uncompensated cost. However, given the all-or-nothing rule (see footnote 2), this could be the subjects' best solution to the problem.

Note that, although the rate-of-accidents version does not have the property of distributional fairness or equity (in that the company must pay in 100% of such cases although it is responsible for only 80% of such accidents), the weigh-attendant version does have the property of distributional fairness. Specifically, in future cases of a similar type the eyewitness will not always identify the same bus company. As described to subjects, the weigh attendant mistakes blue for gray as often as he mistakes gray for blue. Because the witness is accurate at a rate exceeding chance, in the long run each company will lose cases at a rate proportional to their share of cases for which they were responsible. Thus, the fairness argument could account for why people are reluctant to rule against the defendant in the rate-of-accidents version but are not reluctant to rule against the defendant in the weigh-attendant version.

The distinction between the rate-of-accidents version and the weigh-attendant version, therefore, could be a distinction between base rates and incidence rates or what some legal scholars have referred to as a distinction between naked statistical evidence and case-specific or particularistic proof.[2] Indeed, a base rate or prior probability does not, by its very nature, have the property of distributional fairness or equity over the long run. Consistent application of a base-rate rule will always result in a finding against the company with the greatest base rate even when another company accounts for some frequency of the problem. Conversely, incidence rates refer to dynamic events that could, on a case-by-case basis, produce evidence that is either consistent with or opposed to the base rate.

Experiment 3

If people's reluctance to rule against the defendant in the standard blue bus case reflects an intuitive understanding of the distributional fairness problem, then it should be possible to change the statistical evidence to reflect case-specific evidence and thereby get people to rule against the defendant. A new version of the problem was created for this purpose. In the new version, the evidence remains purely statistical but is a case-specific likelihood rather than a base rate and, therefore, has the property of distributional fairness. This version, hereafter called the *tire-tracks version*, replaced the third paragraph of the standard version with the following:

> A second county transportation official took the stand and reported that he examined the dead dog and took prints of the tire tracks. These prints were then transferred onto paper and compared to all 10 of the 10 buses owned by the Blue Bus Company and the 10 owned by the Grey Bus Company. The tracks matched 80% of the Blue Bus Company's buses and matched only 20% of the Grey Bus Company's buses.

As with all versions of the problem, the mathematically correct probability is .80 in the tire-tracks version. However, distribu-

[1] One obvious solution is to make the defendant in this case pay 80% of the damages, a proportionate award approach. This solution is so simple and seemingly complete that a question arises as to why our legal system is so firmly committed to the all-or-nothing rule. The answer rests partly with legal tradition and partly with the message that such verdicts would send to the litigants and the general public (Nesson, 1985), and it is partly beyond logic (see Coons, 1964). Perhaps the easiest way to understand why proportionate damages are unwarranted is to note that *only one bus company* was in fact negligent. This is unlike proportionate damages awarded in the class action against manufacturers of DES who were in fact proportionately responsible for their share of the plaintiffs' injuries (see Brook, 1985).

[2] It is important to distinguish between statistical evidence derived from base rates and that derived from a conditional incident probability. In the case of base rates (or prior probabilities), the evidence exists independently of and prior to the event in question. For example, the Blue Bus Company owned and operated 80% of the buses before Mrs. Prob's dog was run over and this evidence would have existed even if the incident had not occurred. In the case of a conditional incident probability (case specific or particularistic proof), the evidence would not exist at all had the event in question not occurred. The tire-tracks evidence, for example, is particularistic evidence because, without the event in question, there would have been no squashed dog from which to lift tracks for comparison to the two companys' bus tires. It is in this sense that I disagree with some other commentators who have argued that the distinction between base rates and case-specific evidence is purely semantic or illusory (e.g., Koehler & Shaviro, 1990; Nesson, 1985; Saks & Kidd, 1980).

tional fairness prevails with the tire-tracks version; in future cases of a similar type, the company at fault has an 80% of chance of being caught. If the Grey Bus Company causes 20% of the accidents of this type, then they will lose 20% of the cases over the long run of cases. It will not always be the Blue Bus Company that loses. The company that is responsible for most of the neglect will lose most of the cases. The tire-tracks case is as close to the weigh-attendant case as possible from the standpoint of economics, statistics, fairness, and equity.

The tire-tracks version, the rate-of-accidents version, and the weigh-attendant version were given to a new set of subjects. This time, however, the subject population was expanded to include two critical groups. One was a sample of 120 master of business administration (MBA) students in a business statistics course at the University of Alberta, Edmonton, Canada. The other sample consisted of 61 practicing trial judges. The judges were part of a continuing education program in California in which I was giving a workshop on eyewitness testimony at Lake Tahoe in August 1986. The judges were given the task at the beginning of the workshop, before any discussion of eyewitness testimony. In addition to the MBA students and the judges, a sample of 120 undergraduate psychology students was given the task. For each of the three subject populations the three versions of the blue bus case were distributed randomly with the proviso that equal numbers of subjects receive each of the three versions. All subjects first estimated the probability that a Blue Bus Company bus was at fault and then indicated whether they would hold the Blue Bus Company liable.

Results are depicted in Figure 3. Statistical comparisons were not made between the three subject populations because they were given the task at different times and in different locations. The important contrasts are between versions of the evidence (rate of accidents vs. tire tracks vs. weigh attendant) within subject samples for both the subjective probability measure and the verdicts measure. There were no differences in subjective probabilities across cases for any of the three subject populations (all $ps > .30$). On the other hand, each subject population treated the three cases differently in their verdicts. On the verdict measure, the tire-tracks version and the rate-of-accidents version did not differ for any of the three subject populations, all χ^2 (1, ns = 120, 120, and 61) < 1.7, $ps > .20$, whereas the weigh-station attendant version differed from the other two versions for all three subject populations, $\chi^2(1, n = 120) = 144.1$, $p < .001$, $\chi^2(1, n = 120) = 154.8$, $p < .001$, and $\chi^2(1, n = 61) = 81.5$, $p < .001$, for the psychology students, MBA students, and judges, respectively.

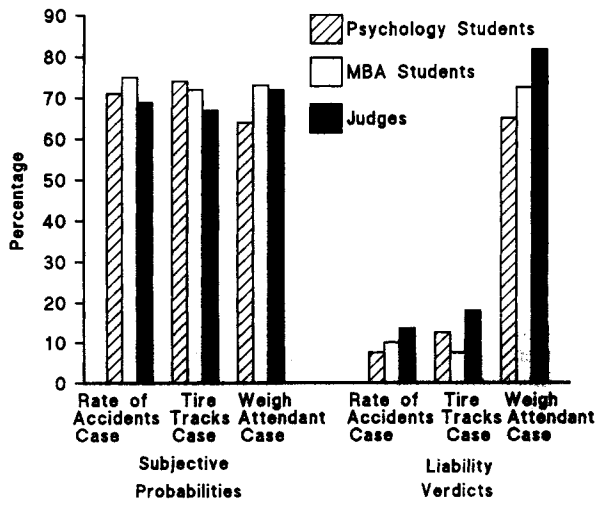

Figure 3. Subjective probabilities and percentage of liability verdicts for the rate-of-accidents, weigh-attendant, and tire-tracks versions. (MBA = master of business administration.)

These data are extremely valuable for ruling out several interpretations of people's reluctance to rule against the Blue Bus Company. First, these data indicate that Posner's (1972) economic argument does not represent a fundamental reason for people's reluctance to use naked statistical evidence. If the economic argument were the basis of such reluctance, the tire-tracks version should have relieved the subjects of the concern and resulted in verdict rates similar to that of the weigh-attendant version. For the same reason, the fairness or equity explanation seems not to be the critical factor because the tire-tracks version distributes outcomes over cases the same way the weigh-attendant version distributes those outcomes. And, although the nonparticularistic or base-rate nature of the rate-of-accidents version fits legal scholars' views of naked statistical evidence, the tire-tracks version is a particularistic, incidence-rate statistic in precisely the same way as the weigh-attendant version.

Clearly, there is something more involved than the mere dismissal of nonparticularistic proof or the concern that the Blue Bus Company will have to pay in 100% of these cases although being at fault in only 80% of the cases. Furthermore, there is considerable agreement in the verdict pattern across three quite different populations. And, one of these populations, practicing trial judges, is hardly naive about matters of proof. The judges provided rich and somewhat sophisticated justifications for their verdicts, as described in the next section.

Explanations by Judges

There were 19 judges in the rate-of-accidents case and 18 judges in the tire-tracks case who refused to rule against the Blue Bus Company. These 37 judges constitute a special group that had some interesting and rather inventive explanations for their judgments. Their explanations were easily placed into several categories. Each explanation is described and discussed in the following paragraphs.

Naked statistics. The most dominant category, mentioned in some form by 20 of the 37 judges, was to dismiss the evidence as naked statistics or mere probability. Unfortunately, this kind of explanation is not very informative and tends to beg the question. Surprisingly, this type of explanation was as common for the tire-tracks version (n = 9) as it was for the proportion-of-accidents version (n = 11) even though it is only the latter that should qualify as nonparticularistic or naked statistical evidence. In other words, the judges seem to have treated the incidence-rate evidence and the base-rate evidence precisely the same not only with regard to verdicts but also with regard to their verdict rationale.

In some cases the judges' dismissal of the evidence as naked

statistics was somewhat more sophisticated than mere name-calling. Some invoked an argument that has been made by Kaye (1979), Thompson (1989), and other legal scholars to the effect that the statistical evidence should be ruled insufficient to support a verdict for the plaintiff so as to create an incentive for the plaintiff to do more than establish background statistics. Indeed, Kaye argues that the very fact that no other evidence is presented is a form of data that allows one to infer that the defendant is probably not liable; otherwise there would have been more convincing proof. This absence of other evidence argument, which is sometimes called the *spoilation inference,* is a curious one as it fails to treat the classes of plaintiffs and defendants equivalently. As other legal scholars have noted "If the plaintiff can produce more evidence, then so can the defendant . . . the result, one would think, would be that the inference Kaye discusses would arise on both sides and cancel each other out" (Allen, 1986, pp. 412–413). The situation is somewhat different, of course, if one believes that either party is refusing to disclose relevant information in the discovery process, but that concern hardly seems unique to naked statistical evidence and there are other incentives, procedures, and remedies in place to produce evidence. Nevertheless, this absence of other evidence argument is revisited after the fourth experiment as its credibility seems substantially reduced by the data in Experiment 4.

Sample size. Far more interesting was the argument that the probability evidence was unconvincing because the observations were too few to be trustworthy. This kind of statement was mentioned by 3 judges in the rate-of-accidents version and 5 judges in the tire-tracks version. The most articulate example was by 1 judge in the tire-tracks version who said, "Having only 10 buses, a statistician would say that there are too few cases to base a stable estimate of the true probability." This is an interesting argument but, of course, a misunderstanding of the situation. The 10 buses owned by each company represent populations rather than samples. Accordingly, the number of buses is irrelevant. More relevant is the argument about sample size in the rate-of-accidents version of the problem. None of the judges articulated the sample-size issue particularly well, so the following is a paraphrase of the argument: "Although the defendant company's buses accounted for 80% of the accidents involving buses, there were probably few bus accidents. If the defendant's company was responsible for 4 of a total of 5, then the proportion doesn't seem very stable. If they were responsible for 80 of 100, then that would seem to be a reliable statistic." Although this is an excellent point, it is important to note that no one criticized the weigh-attendant case along these lines, even though the same type of argument can be made. For example, no judge said "Although the weigh attendant was accurate only 80% of the time, this could have been 4 out of 5 or it could have been 80 out of 100." It could be assumed, credibly of course, that bus weighings are a frequent event, probably numbering several hundred at the weigh station in question, whereas accidents would be based on a small sample. Nevertheless, concern about sample size does not seem to explain the different verdicts across cases because sample size is not an issue in the tire-tracks version because the sample included the entire population. From a statistical viewpoint the tire-tracks version provides the strongest evidence.

Time frame. One rather sophisticated argument came from 4 judges in the proportion-of-accidents version. They argued that the accident data should not be trusted because accident records could change over time. "How do we know that the 80% figure for accidents is true in the most recent 3 months or the most recent 3 weeks? Perhaps their record improved over time." Considered in isolation, this is an excellent point. But why is such a question not raised with regard to the weigh-attendant data? Why did no judge say something like "How do we know that the weigh attendant's errors are true of the most recent time frame? Perhaps his mistakes are not applicable to the most recent 3 months or 3 weeks on the job." In addition, the time-frame argument fails to explain their reluctance to rule against the defendant in the tire-tracks case.

Either–or. Perhaps the weakest argument was one that was invoked by 7 of the judges in the proportion-of-accidents version and by 8 of the judges in the tire-tracks version. Although the precise wording varied, the argument was an either–or argument, to wit "there were two possibilities; it was either the Blue or the Grey Bus Company. Thus, the real probability is 50/50 regardless of the evidence presented." Such statements are something of a shock to the statistically minded person. Presumably, people who agree with such statements will take even–odds bets on any sporting event (as there are only two teams and either one or the other will win) and would disregard any weather forecast in deciding whether to carry an umbrella. After all, it either will or will not rain, so the odds are 50/50 regardless of what the forecaster says. Once again, however, such a logic begs the question of why verdicts were so different in the weigh-attendant version. After all, there are still only 2 bus companies and the weigh attendant was either right or wrong. Interestingly, of the 15 judges who gave an either–or statement backing their verdict, 5 gave probability estimates of 80%.

Although the focus of the foregoing analysis was on the explanation of the 37 judges who ruled for the defendant, there were also explanations given by the 25 judges who ruled for the plaintiff. Most of these were in the weigh-attendant case (18) rather than the tire-tracks (4) or proportion-of-accidents (3) cases. These explanations fell into two categories as described in the following two paragraphs.

Balance of probability. All 4 judges in the tire-tracks case, all 3 in the proportion-of-accidents case, and 12 of the 18 in the weigh-attendant case made some kind of statement about the evidence being sufficient on the "weight of evidence" or the "balance of probability" or that the plaintiff's argument was "more likely true than not."

The 10-min coincidence. Eight of the 18 judges in the weigh-attendant version argued that there was a coincidence that could not be readily dismissed. Specifically, these judges argued that the time logged by the weight attendant (11:30) and the time Mrs. Prob's dog was killed (11:40) matches well the distance (a 10-min drive) from the weigh station to the scene of the bus–dog incident. Presumably, these judges are saying that this is an additional piece of circumstantial evidence. This fails to explain why subjective probabilities are the same in the weigh-attendant version as in the other versions. But it is even more important to recognize that this coincidence is illusory because it only tells us that the vehicle was a bus; it remains the

case that there is only an 80% chance that the bus was blue. In fact, no one questioned the proposition that it was a bus that killed Mrs. Prob's dog in any version of the evidence. Furthermore, if this apparently stipulated element was in question, it would seem that the tire-tracks version would be the version that most clearly puts this concern to rest.

Fact-to-Evidence Reasoning Versus Evidence-to-Fact Reasoning

Up to this point, numerous interpretations of people's reluctance to return affirmative verdicts in the statistical versions of the blue bus case have been ruled out. These include the idea that their subjective probabilities are too low, their thresholds are too high, failure to see causal relevance, and the unfairness of requiring the same company to pay in all such cases when they are not responsible for all such cases. In the process of ruling out these interpretations, the first two goals of this article have been met. Specifically, the completeness of the probability–threshold model as it is commonly understood in the literature has been questioned by these data, and some of the major hypotheses of scholars who have discussed the blue bus problem were subjected to empirical tests and found to not be adequate. The third goal, to suggest a modification or alternative to the probability–threshold model that can account for these data, was clearly stated as a tertiary goal. Nevertheless, an attempt is made here to provide at least a speculative account of how evidence can drive subjective probabilities without affecting verdicts.

It seems clear that the statistical versions of the blue bus case (including the particularistic, causally relevant, physical tire-tracks version) lack something that people require of evidence before they will render an affirmative verdict. Although it is tempting to argue that people accept the weigh-attendant version because they are overly willing to trust eyewitness accounts (e.g., see Wells & Loftus, 1984), such an explanation is inadequate on several counts. First, that would not be an explanation at all but rather a mere description of what people do; it continues to beg the question. Second, it fails to explain why such overbelief of eyewitnesses is not reflected in inflated subjective probabilities in the weigh-attendant version. Finally, it fails to account for the fact that affirmative verdicts are commonly reached in actual trials without eyewitnesses.

The hypothesis offered here is that *in order for evidence to have a significant impact on people's verdict preferences, one's hypothetical belief about the ultimate fact must affect one's belief about the evidence.* Notice how this criterion is not satisfied in the standard blue bus case. In the standard case, what one believes as the ultimate truth (i.e., that the bus was blue or gray) in no way affects one's belief of the evidence. For example, a person can entertain the hypothetical belief that the offending bus was in fact gray and continue to believe that the transportation official was correct that the Blue Bus Company owns and operates 80% of the buses. Similarly, a person can entertain a belief that the offending bus was gray rather than blue and continue to fully accept the official's statement that the tire tracks matched 80% of the blue buses and only 20% of the gray buses. This is not true of the weigh-attendant version, however. A person cannot believe both that the bus was gray and also believe that the weigh attendant was correct about the identity of the bus.

In other words, it is proposed that mere subjective probability is not sufficient to drive verdicts. Instead, *the evidence must be presented in a form that makes that evidence believable or not believable depending on what one assumes about the ultimate fact.* If one's assumptions about the ultimate fact do not require one to disbelieve the evidence, then the evidence need not affect one's belief about the ultimate fact. This type of reasoning could be called *fact-to-evidence reasoning* to distinguish it from our typical depiction of the process where people are presumed to reason only from evidence to ultimate fact.[3]

Issues of rationality or normative appropriateness notwithstanding, most people probably react differently to situations where their assumptions (or knowledge) of the ultimate fact affect their assessments of the evidence as opposed to situations where the ultimate fact does not affect their assessments of the evidence. Consider two ways in which a meteorologist might warn listeners of rain. In one case she or he says, "There is an 80% chance of rain today"; in another case she or he says, "Based on a set of readings that are 80% accurate, it will rain today." To the Pascallian thinker, these are equivalent statements.[4] But suppose it does not rain. Lack of rain does not lead one to think of the meteorologist as having been wrong in the former case. We need not disbelieve the meteorologist to reconcile the ultimate fact (no rain) with the forecast. In the latter case, however, one's knowledge of the ultimate fact tempts one to think of the forecaster as having been wrong. Psychologically, there seems to be a difference between saying that there is an 80% chance that something is true and saying that something is true based on evidence that is 80% reliable.

---

[3] It is assumed here that one need not have actual knowledge of the ultimate fact in order to make such assessments. One needs merely to mentally simulate (cf. Kahneman & Tversky, 1982; Wells, Taylor, & Turtle, 1987) the process by asking "what if . . ." questions to determine whether assumptions regarding the ultimate fact require a reevaluation of the evidence.

[4] One reviewer objected to the idea that all Pascallians would treat these as equivalent statements. The central objection seems to be that the former forecast may indicate only a prior based on previous occurrences for the day with no observations of other data such as a barometer or a weather satellite. The latter forecast, however, may indicate several readings combined. The reviewer's objection, however, is at least a rebuttable assertion. A Pascallian should be no more impressed with an 80% prediction based on the mere fact that it rains on this day at this location 80% of the time than on a combination of evidence that produces *that same probability.* Notice, for example, that to yield combined probability of 80% using more than one source, each of the individual probabilities would have to be either less than 80% or inconsistent in their direction of prediction. Hence, if the barometer indicates a 70% chance of rain, the base rate for rain that day would have to be around 63% to indicate an 80% overall probability. A probability by base rate alone of 80% is no less certain than one based on a base rate of 63% and a second, independent probability of 70%. Koehler and Shaviro (1990; see also Koehler, 1991) make a similar point in reference to the concept of "second-order uncertainty," or uncertainty about one's probabilistic estimates.

Experiment 4

The fourth experiment in this series attempted to test the idea that the evidence must be presented in such a way that one's assumption about whether it was a blue bus affects one's tendency to believe the evidence. Accordingly, the tire-tracks version was reframed in the critical paragraph as follows:

> A second county transportation official took the stand and reported that he examined the dead dog and took prints of the tire tracks. The prints were then transferred onto paper and matched to all of the 10 buses owned by the Blue Bus Company and the 10 owned by the Grey Bus Company. He testified that the technique used for matching is correct 80% of the time and, based on this technique, he believed that the bus that ran over Mrs. Prob's dog was a Blue Bus Company bus.

Notice that this version, hereafter called the *tire-tracks-belief* version, should have no different effect on people's subjective probabilities than the previous versions. Using fact-to-evidence reasoning, however, people must discount the transportation official's conclusion to believe that the bus was not blue.

The tire-tracks-belief version, the original tire-tracks version, and the weigh-attendant version were assigned randomly to 90 students and 45 judges. The students were enrolled in a psychology course and the judges were attendees at the Canadian Institute of Criminal Justice in Edmonton, Alberta, Canada. Each gave a subjective probability before giving his or her verdict. Figure 4 shows the subjective probabilities and verdicts for each subject group and each version of evidence.

Once again, no differences emerged in subjective probabilities, all ts (dfs = 89 and 44) < 1.2, $ps > .20$. The tire-tracks version, however, differed significantly from the tire-tracks-belief version in the verdicts for both students, χ^2 (1, n = 90) = 8.24, $p < .01$, and for judges, χ^2 (1, n = 45) = 8.88, $p < .01$. The tire-tracks-belief version and the weigh-attendant version did not differ for either students or judges on the verdict measure, both χ^2s (1, ns = 90 and 45 for students and judges, respectively) <1.3, $ps > .25$. Therefore, a mere reframing of the tire-tracks evidence was sufficient to increase affirmative liability decisions by over 5 times while leaving subjective probabilities at the same level.

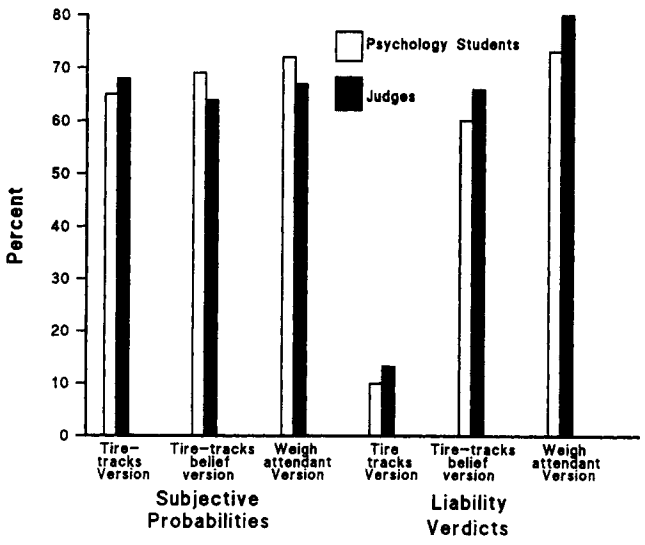

Figure 4. Subjective probabilities and percentage of liability verdicts for the tire-tracks, tire-tracks belief, and weigh-attendant versions.

Recall that some of the judges rationalized their reluctance to use statistical evidence on the argument that the absence of some better form of proof calls into question the plaintiff's version of the events: If it really happened this way, would there not be something more convincing to prove it? Indeed, juries are allowed to draw adverse inferences from missing evidence, and a rational person might very well do so. For example, the plaintiff's failure to produce bus schedule evidence to show that a blue bus was likely to be in the general vicinity might lead one to think that the plaintiff had collected such evidence and then withheld it because it was nonsupportive. The failure of the defense to come forward with such offsetting evidence notwithstanding, perhaps because of incompetent defense counsel, people might infer that the actual probability is not .80 but .80(X), where X represents some number less than 1.0 and reflects the missing evidence.

There are two reasons to believe that the adverse-inference-from-missing-evidence interpretation does not account for the data in these experiments. First, and most obvious, is the fact that such a process predicts that subjective probabilities will regress downward more in the naked statistics versions than in the weigh-attendant version. Clearly, this did not occur. Second the missing-evidence notion fails to explain why verdicts in the tire-tracks version should differ from those in the tire-tracks-belief version. The same evidence (tire tracks) exists in both versions and any missing evidence (such as bus schedules, mileage checks, dog splatterings on a bus chassis, etc.) are equally missing in both versions.

Experiment 5

The previous four experiments illustrate that people react to probabilistic evidence in different ways, depending on how that evidence is framed even under conditions in which subjective probabilities do not vary. Suppose, however, subjects were more explicitly instructed to follow a preponderance of evidence standard to the point where they were told that the law's norms *required* them to rule for the plaintiff if the chances are greater than 50%. When placed under such restrictive instructions, two kinds of effects might emerge. First, subjects might simply conform to the stated standard, accepting the mandate that they treat the decision as a game of numbers and act accordingly. This would tend to have the effect of eliminating differences between the nonimpactful versions (e.g., proportion-of-accidents evidence) and the impactful versions (e.g., weigh-station attendant evidence).

A second possibility is that subjects will continue to resist making proplaintiff decisions in nonimpactful versions but perhaps adjust their subjective probability estimates downward to 50% so as to be able to reach the preferred verdict (prodefense) while staying consistent with the stated requirement of the law's norms. Other possibilities exist as well, such as a combination of the above two effects as well as the possibility that

the instructions will have no effect at all. It seems unlikely, however, that such instructions will have no effect as they constitute a relatively unambiguous and simple demand to subjects about the rules for responding to the verdict and subjective probability questions. In effect, subjects would have to violate the instructions to give subjective probabilities beyond 50% while failing to rule for the plaintiff. How subjects resolve this dilemma can be informative about the process.

A fifth experiment was conducted using the proportion-of-accidents version of evidence combined with one of three instructions regarding the criterion for decision. At the least restrictive level, subjects read, "You are a juror in this case and the judge has instructed you to find for the plaintiff, Ms. Prob, if there is a preponderance of the evidence in her favor." At a second, more restrictive level, a sentence was added "By a preponderance of the evidence is meant that you should find for the plaintiff if it is more likely than not that a Blue Bus Company bus ran over Ms. Prob's dog." A third, more restrictive and explicit sentence replaced the latter sentence at the third level of instruction, to wit "By a preponderance of evidence is meant that you should rule for the plaintiff if the chances that Ms. Prob's dog was run over by a Blue Bus Company bus are greater than 50%, however slightly."

Ninety undergraduate students were assigned randomly to the three instruction conditions with the proviso of equal cell sizes. Each subject estimated a probability and made a verdict judgment. The results are shown in Table 1. Verdicts were significantly affected by instructions, $\chi^2(1, n = 90) = 26.3$, $p < .01$, in the direction that one would expect. Importantly, however, probability estimates were also affected, $F(2, 87) = 4.37$, $p < .02$. These results indicate that, although some subjects used the instructions in such a way as to increase their rate of affirmative decisions for the plaintiff, others chose to lower their reported subjective probabilities so as to maintain their decisional stance against accepting a verdict for the plaintiff. Of these two effects, the more interesting seems to be the decline in subjective probabilities with increasingly strict instructions. It seems that subjects were placed in a bind; they did not want to rule for the plaintiff but were told to do so if their subjective probabilities exceeded 50%. As a result, 14 of the 30 subjects (46.7%) who were given the strictest instructions (i.e., greater than 50%) gave a probability estimate of 50% and, in each of these 14 cases, refused to rule against the Blue Bus Company. When given the weaker instructions (preponderance of evidence), on the other hand, only 5 of the 30 subjects (16.7%) gave probability estimates as low as 50%.

This fifth experiment indicates that people probably would yield affirmative verdicts for a plaintiff from naked statistical evidence if the instructions to them were sufficiently explicit regarding the greater-than-50% criterion and delivered authoritatively. But it also appears that people are sufficiently uncomfortable with reaching affirmative verdicts on the basis of naked statistics that a reasonable number of them lower their subjective probabilities to not reach such a verdict.

Table 1
Subjective Probabilities and Verdicts for the Proportion-of-Accidents Evidence as Functions of Level of Instruction

Instructions	Probabilities	Verdicts (%)
Preponderance of evidence	.71	17
More likely than not	.61	37
Greater than 50%	.58	47

General Discussion

Overview

This research began with the question of whether people are reluctant to accept naked probabilities as evidence of liability. The answer is clearly yes, and this is equally true for experienced trial judges, business students, and psychology students. The problem is not that people's subjective probabilities fail to match the mathematical probabilities, as Tribe (1971) has suggested; subjective probabilities closely followed the mathematically correct probabilities across all versions of the blue bus case. Nor can the results of these studies be explained by suggesting that subjects misunderstood the balance of probability criterion for proof in civil trials; the weigh-attendant case and the tire-tracks-belief version yielded the same subjective probabilities as the other versions and yet produced affirmative verdicts at rates that were 5–10 times higher.

Causal relevance, which has an established role in people's underuse of statistical information in other contexts, does not appear to provide explanatory power in this context. The causal relevance of the rate-of-accidents version is clear and yet is treated by subjects as no different from the proportion-of-business version, which has no clear causal relevance. Furthermore, the tire-tracks version has causal relevance in that the bus at fault caused the tracks to be left. Additionally, the concept of causal relevance fails to explain why a mere reframing of the tire-tracks version (the tire-tracks-belief version) produced a 500% to 600% increase in affirmative verdict rates with no effect on subjective probabilities.

Concerns about equity, fairness, and economic dislocation (Posner, 1972) seem compelling and do indeed constitute good arguments against naked probabilities when those probabilities are based on mere base rates. But these arguments do not explain the pattern of verdicts across the various versions of the blue bus case. Unlike the volume-of-traffic version, to which the fairness argument can be applied, the tire-tracks version has the same property of distributional fairness as the weigh-attendant version. Specifically, the volume-of-traffic version requires the trier of fact to rule against the larger company in 100% of the future similar cases even though the larger company is responsible for less than 100% of the cases; the tire-tracks version, however, carries no such implication. Instead, the rate at which evidence will surface against the offending bus company will be proportional to that company's rate of offenses in future similar cases involving the tire-track evidence. And, again, the fairness argument fails to account for why the two versions of the tire-track evidence yielded such markedly different verdicts.

Table 2 is a matrix of the various versions of the evidence and the characteristics of these versions. This summarizes the main points. Although the probability for each version is .80 (and

Table 2
Characteristics of Evidence

Version of evidence	$p = .80$?	Causal relevance?	Particularistic proof and long-term equity?	Sample size or time-frame problem?	Passes fact-to-evidence reasoning test?
Proportion of business	Yes	No	No	Possibly	No
Rate of accidents	Yes	Yes	No	Possibly	No
Tire tracks	Yes	Yes	Yes	No	No
Weigh attendant	Yes	Yes	Yes	Possibly	Yes
Tire tracks–belief	Yes	Yes	Yes	No	Yes

subjective probabilities were near .80 and undifferentiated across versions of the evidence), the proportion-of-business version did not have causal relevance. Nevertheless, the other four versions had causal relevance, hence failing to explain differences in verdicts across those versions. Note also that, although the proportion-of-business and rate-of-accidents versions were not forms of particularistic proof and did not have the characteristic of long-run equity, the tire-tracks version had both of these characteristics, thus failing to explain why the tire-tracks version continued to not produce affirmative verdicts. As well, although arguments about sample size and time frame (or recent trends) are possible problems for the proportion-of-business version, the rate-of-accidents version, and the weigh-attendant version, neither is a problem for the tire-tracks version. Hence, the sample size and time-frame arguments fail to account for the verdict pattern across versions of the evidence.

Why are people so reluctant to use pure probability as evidence of liability but highly willing to make affirmative liability decisions under other conditions that yield the same subjective probabilities? It seems that reliance on probabilistic information is qualitatively distinct from reliance on someone else's belief or opinion, even if that person's belief or opinion is itself based merely on the probabilistic information. Consider the following hypothetical, adapted from Nesson (1985):

> Jill was given a brief glimpse of a playing card, too brief to be certain of its identity, but she thinks it was not a face card. Jill is now asked to decide whether or not it was a face card. She is at least 75% certain that it was not a face card, so she says it was not a face card.

Now, suppose the card is turned over and it is a queen of hearts. Would you say that she made a mistake? If she had convinced someone to place a bet on her advice and they lost money, would she feel guilty and perhaps even apologize for misleading them? Most people would.

Consider an alternative scenario:

> Jill observes someone drawing a playing card randomly from a full, well-shuffled deck. She never sees the card, but is asked to decide whether or not it is a face card. Knowing that fewer than 25% of playing cards in a full deck are face cards, Jill is at least 75% certain that it is not a face card, so she says it was not a face card.

Again, suppose that the card is now turned over and it is a queen of hearts. Would you say that Jill made a mistake? If she had convinced someone to place a bet on her advice and they lost money, would she feel guilty and apologize for misleading them? Probably not. Jill was not to blame, she made the correct decision; it was mere chance that produced the improbable outcome.

There is a sense in which Jill was right in both of these cases and wrong in both of these cases. She was right to go with the highest subjective probability; she was wrong in that the correct answer was not what she guessed. But the purpose of this comparison of scenarios is to illustrate the subjective differences that seem to exist. In the "glimpse" case, the knowledge of the ultimate fact (that it was in fact a face card) leads us to reevaluate the evidence. One cannot both believe the ultimate fact and also believe that Jill was correct. In the random draw case, however, one continues to find the evidence credible and needs not reevaluate the goodness of that evidence even when it is known that the ultimate fact is that it was a face card. So too, perhaps it is with decisions of liability that triers of fact are sensitive to differences between evidence that need not be reevaluated by knowledge of the ultimate fact and evidence that is affected by knowledge of the ultimate fact. Assuming, for instance, that it was a gray bus that ran over Mrs. Prob's dog does not affect our evaluation of the plaintiff's evidence in the proportion-of-business, rate-of-accidents, or tire-tracks cases. But it does affect how we think about the weigh-attendant evidence (the weigh attendant was wrong!) and the tire-tracks-belief evidence (the transportation official's conclusion was wrong).

What is being suggested here is that people require more of evidence than merely that it affect their views of the ultimate fact; their views of the ultimate fact must also affect their perceptions of the evidence. Perhaps eyewitness testimony is consistently persuasive because it can virtually always pass this bidirectional test. Similarly, fingerprint experts, when allowed to state conclusions (e.g., "I conclude that the prints lifted from the glass are those of the defendant"), are likely to pass this bidirectional test of good evidence.

Returning to the paternity suit case described at the beginning of this article, in which experts reported that blood tests showed a 99.8% probability that a defendant was in fact the father, one suspects that the plaintiff would have won the suit if the expert had reframed his testimony to say that "based on a blood test that is 99.8% accurate, I conclude that the defendant is the father" rather than "based on a blood test, there is a 99.8% probability that the defendant is the father." Although it is easy to recognize the statistical equivalence of these two statements, especially when the statements are placed side by side, there appears to be a robust psychological difference between

the former (statement of belief based on a high probability) and the latter (statement of a high probability).

On the other hand, the paternity suit was not a pure case of naked statistical evidence because the mother gave testimony that the man had fathered her child and the defendant took the stand and made consistent and confident statements of denial that he was the father. Indeed, an actual liability suit based on nothing more than naked statistical evidence would be unlikely to occur (but see Turner v. U.S., 1970, and Sindell V. Abbot Labs, 1980, for examples in which the admissibility of naked statistical evidence has been upheld). In the blue bus case, for example, it seems likely that the defense would have each of the Blue Bus Company drivers take the stand and deny that they had driven that road at that time. Both sides might present bus schedules, the weight recorded by the weigh attendant might be compared with the empty weight of the defendant's buses, and so on. As a result, judge and jury would end up with several pieces of information that, although perhaps not highly probative in isolation, would be combined with the naked statistics in making a final determination of liability. As well, the absence of such elements as a voir dire, direct and cross examination, live testimony, and deliberation render the methods used in the present experiments highly suspect in terms of generalization to actual trials. Hence, the present research is less informative about the kinds of outcomes that are to be expected in actual court cases than it is about the processes that govern the way people reason about evidence.

Implications for the Probability–Threshold Model

Most models of jury decision making proposed in psychology assume a subjective probability–threshold process.[5] Hence, it has been important to try to estimate the threshold value that people hold under various sets of conditions as well as develop models of how people combine evidence to reach a particular subjective probability (e.g., Carlson & Dulaney, 1988; Connolly, 1987; Dane, 1985; Fried et al., 1975; Kagehiro & Stanton, 1985; M. F. Kaplan, 1977; Kerr, 1978; Kerr et al., 1976; Marshall & Wise, 1975; Nagel, 1979; Ostrom, Werner, & Saks, 1978; Schum, 1975, 1977; Schum & Martin, 1982; Simon, 1970; Thomas & Hogue, 1976). Research based on subjective probabilities and threshold notions has been fruitful and informative. But the current set of experiments represents a challenge to the subjective probability–threshold model to either incorporate a caveat or explore alternative models that assume a different process.

Perhaps the easiest solution is to incorporate a caveat. Indeed, some caveats have already been incorporated into the probability–threshold model, such as when people reject the evidence on the basis of the process that generated it (as in forced confessions) or when the penalties are so severe that the juror refuses to rule against the defendant out of an objection to the laws that would call for such penalties. In these situations, it matters little that subjective probabilities exceed the threshold for an affirmative verdict. In these cases, Lempert's (1977) regret matrix or any simple model of sentiment or policy considerations can easily explain why the probability–threshold model should not drive verdicts.

In the case of naked and "apparently naked"[6] statistical evidence, however, a somewhat different and perhaps more peculiar caveat seems to apply. Specifically, it could be argued that people will allow their subjective probabilities to drive their verdict decisions only if the evidence on which those subjective probabilities are based is responsive to assumptions about the ultimate fact. Hence, mere base rates (short of 100%) and even case-specific likelihoods frequently will fail to produce affirmative verdicts even when such evidence drives subjective probabilities beyond threshold values.

Possible alternative models would probably look more social–psychological and less mathematical than the probability–threshold model. For example, one might propose a belief-heuristics model in which the trier of fact assesses the persuasive value of evidence according to how easy or difficult it would be to both believe the plaintiff's evidence and believe the defendant was not at fault. Here, ease would not be associated with some probability but perhaps with some assessment of logical connection. Alternatively, a model of anticipated justification might better capture people's actual decision process. The idea would be that people mentally simulate the possibility that the truth will be uncovered at a later time and, if they were wrong, would need to justify their decision by saying that they were misled. Notice that such justifications would be difficult in the naked and apparently naked statistical evidence cases because someone merely reported factual numbers and those numbers are still credible and accurate. But in the tire-tracks-belief and weigh-attendant cases, one could say that they were misled ("I believed him and he was wrong!"). Perhaps it is the case that people prefer to base their beliefs on the beliefs of another person, even when the other person's beliefs are based on a form of evidence that those people themselves would not directly use in deriving a belief.

Yet another possibility was introduced recently by Wasserman (1991), who argues that naked statistical evidence requires the trier to make an inference that, in effect, denies the defendant's freedom to depart from his or her past conduct or the conduct of his or her group. The volume-of-business evidence, for example, denies the Blue Bus Company's capacity for operating at a greater level of safety than its smaller competitor, and the proportion-of-accidents evidence denies the company's capacity for having exercised greater safety now than when their past conduct was recorded. Wasserman argues that triers are reluctant to use such evidence because it would infringe on the defendant's individuality and autonomy.

Whatever the best approach might be at this time (alternative models vs. modifications of the probability–threshold model), it is clear that there was no theoretical reason in the current psychological literature to expect that these versions of the evidence, which yield functionally equivalent subjective probabilities, would produce highly discrepant verdicts. Our under-

---

[5] A notable exception is Pennington and Hastie's (1986) story model.

[6] The term *apparently naked* is used here to refer to cases like the tire-tracks case, which people seem to treat the same way they treat the volume-of-business or proportion-of-accidents cases. Technically, however, the tire-tracks evidence is not what the courts and legal scholars have called naked statistics. The tire-tracks evidence is a case-specific, particularistic likelihood that would not exist had the event in question never occurred.

standing of how people evaluate evidence for purposes of rendering verdicts might benefit from exploring additional forms of evidence in which subjective probabilities and final verdicts do not agree.

References

Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 267–299). San Diego, CA: Academic Press.

Ajzen, I. (1977). Intuitive theories of events and effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35,* 303–314.

Allen, R. J. (1986). A reconceptualization of civil trials. *Boston University Law Review, 66,* 401–437.

Ball, V. C. (1961). The moment of truth: Probability theory and standards of proof. *Vanderbilt Law Review, 14,* 807–821.

Borgida, E., & Brekke, N. (1981). The base-rate fallacy in attribution and prediction. In J. H. Harvey, W. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 63–95). Hillsdale, NJ: Erlbaum.

Brook, J. (1985). The use of statistical evidence of identification in civil litigation: Well-worn hypotheticals, real cases, and controversy. *St. Louis University Law Journal, 29,* 293–352.

Callen, C. R. (1982). Notes on a grand illusion: Some limits on the use of Bayesian theory in evidence law. *Indiana Law Journal, 57,* 1–44.

Carlson, R. A., & Dulaney, D. E. (1988). Diagnostic reasoning with circumstantial evidence. *Cognitive Psychology, 20,* 463–492.

Cohen, L. J. (1977). *The probable and the provable.* Oxford, England: Clarendon Press.

Cohen, L. J. (1981). Subjective probability and the paradox of the gate crasher. *Arizona State Law Journal, 52,* 627–656.

Cohen, L. J. (1986). The role of evidential weight in criminal proof. *Boston University Law Review, 66,* 635–649.

Connolly, T. (1987). Decision theory, reasonable doubt, and the utility of erroneous acquittals. *Law and Human Behavior, 11,* 101–112.

Coons, P. (1964). Approaches to court-imposed compromise: The uses of doubt and reason. *Northwestern University Law Review, 58,* 750–796.

Dane, F. C. (1985). In search of reasonable doubt. *Law and Human Behavior, 9,* 141–158.

Finkelstein, M., & Fairley, W. (1970). A Bayesian approach to identification evidence. *Harvard Law Review, 83,* 489–517.

Fried, M., Kaplan, K. J., & Klein, K. W. (1975). Juror selection: An analysis of voir dire. In R. J. Simon (Ed.), *The jury system in America: A critical overview.* Beverly Hills, CA: Sage.

Guenther v. Armstrong Rubber Company, 406 F.2d. 1315–1318 (3rd Cir. 1969).

Kagehiro, D. K., & Stanton, W. C. (1985). Legal vs. quantified definitions of standards of proof. *Law and Human Behavior, 9,* 159–178.

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

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

Kaplan, J. (1968). Decision theory and the factfinding process. *Stanford Law Review, 20,* 1065–1084.

Kaplan, M. F. (1977). Judgment by juries. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes in applied settings.* San Diego, CA: Academic Press.

Kaye, D. (1979). The law of probability and the law of the land. *University of Chicago Law Review, 47,* 34–56.

Kaye, D. H. (1982). The limits of the preponderance of the evidence standard: Justifiably naked statistical evidence and multiple causation. *American Bar Foundation Research Journal, 2,* 487–515.

Kaye, D. H. (1986). *Comment: Boston University Law Review, 66,* 657–672.

Kerr, N. (1978). Severity of prescribed penalty and mock jurors' verdicts. *Journal of Personality and Social Psychology, 36,* 1431–1442.

Kerr, N. L., Atkin, R. S., Stasser, G., Meek, D., Holt, R. W., & Davis, J. H. (1976). Guilty beyond a reasonable doubt: Effects of concept definition and assigned decision rule on the judgments of mock jurors. *Journal of Personality and Social Psychology, 34,* 282–294.

Koehler, J. J. (1991). The probity–policy distinction in the statistical evidence debate. *Tulane Law Review, 66,* 141–150.

Koehler, J. J., & Shaviro, D. N. (1990). Veridical verdicts: Increasing verdict accuracy through the use of overtly probabilistic evidence and methods. *Cornell Law Review, 75,* 247–279.

Lempert, R. (1977). Modeling relevance. *Michigan Law Review, 75,* 1021–1057.

Lempert, R. (1986). The new evidence scholarship: Analyzing the process of proof. *Boston University Law Review, 66,* 439–477.

Marshall, C. R., & Wise, J. A. (1975). Juror decisions and the determination of guilt in capital punishment cases: A Bayesian perspective. In D. Wendt & C. Vleck (Eds.), *Utility, probability, and human decision making.* Dordrecht, The Netherlands: Reidel.

McCormick, C. (1984). *Handbook of the law of evidence* (3rd ed.). St. Paul, MN: West.

Nagel, S. (1979). Bringing the values of jurors in line with the law. *Judicature, 63,* 189–195.

Nesson, C. (1985). The evidence or the event? On judicial proof and the acceptability of verdicts. *Harvard Law Review, 98,* 1357–1392.

Nesson, C. (1986). Agent orange meets the blue bus: Factfinding at the frontier of knowledge. *Boston University Law Review, 66,* 521–539.

99.8% probability in blood test rejected on child support case. (1986, January 3). *Edmonton Journal,* p. A2.

Ostrom, T. M., Werner, C., & Saks, M. J. (1978). An integration theory analysis of jurors' presumptions of guilt or innocence. *Journal of Personality and Social Psychology, 36,* 436–450.

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences.* New York: Hill & Wang.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology, 51,* 242–258.

Posner, R. (1972). *Economic analysis of the law,* Boston: Little, Brown.

Saks, M. J., & Kidd, R. F. (1980). Human information processing and adjudication: Trial by heuristics. *Law and Society Review, 15,* 123–160.

Schum, D. A. (1975). The weighing of testimony in judicial proceedings from sources having reduced credibility. *Human Factors, 17,* 172–182.

Schum, D. A. (1977). The behavioral richness of cascaded inference models: Examples in jurisprudence. In N. Castellan, D. Pisoni, & G. Potts (Eds.), *Cognitive Theory* (Vol. 2). Hillsdale, NJ: Erlbaum.

Schum, D. A. (1986). Probability and the processes of discovery, proof, and choice. *Boston University Law Review, 66,* 825–876.

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review, 17,* 105–151.

Simon, R. J. (1969). Judges' translations of burdens of proof into statements of probability. *Trial Lawyers' Guide, 38,* 103–114.

Simon, R. J. (1970). Beyond a reasonable doubt: An experimental at-

tempt at quantification. *Journal of Applied Behavioral Science, 6,* 203–209.

Sindell v. Abbot Labs, 26 Cal. 3rd. 588, 607–610 (1980).

Smith v. Rapid Transit, Inc., 317 Mass. 469, 58 N.E.2d. 754 (1945).

Thomas, E. A. C., & Hogue, A. (1976). Apparent weight of evidence, decision criteria, and confidence ratings in juror decision making. *Psychological Review, 83,* 442–465.

Thompson, W. C. (1989). Are juries competent to evaluate statistical evidence? *Law and Contemporary Problems, 52,* 9–41.

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior, 11,* 167–187.

Tribe, L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review, 84,* 1329–1393.

Turner v. U.S., 396 U.S. 398, rehearing denied, 397 U.S. 958 (1970).

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgment under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base-rates. In D. Kahneman, A. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Walster, E., Walster, G. W., & Berscheid, E. (1978). *Equity: Theory and research.* Boston: Allyn & Bacon.

Wasserman, D. (1991). The morality of statistical proof and the risk of mistaken liability. *Cardozo Law Review, 13,* 935–936.

Wells, G. L., & Loftus, E. F. (Eds.). (1984). *Eyewitness testimony: Psychological perspectives.* New York: Cambridge University Press.

Wells, G. L., Taylor, B. R., & Turtle, J. W. (1987). The undoing of scenarios. *Journal of Personality and Social Psychology, 53,* 421–430.

Appendix

Bayesian Rationality

Normally, the type of statistical evidence used in these hypothetical situations is not considered to be Bayesian because Bayes's Theorem is a prescription for *combining* evidence, and, in these cases, there is no other evidence to combine with the base rates. These problems are, however, "Bayesian rational" in the sense described by Lempert (1986). Bayesian rationality holds that base rates are taken so seriously that in the absence of additional information the base rate is treated as the posterior probability and, in this sense, these problems are Bayesian. Lempert (1986) has made this same point regarding the paradox of the gate-crasher, and the idea is extended here to the blue bus case by showing that we are implicitly combining the base-rate evidence (e.g., proportion of traffic) with various likelihood ratios whose values happen to be 1.0. Consider, for example, the fact that Ms. Prob's dog is in fact dead as a piece of evidence in the case. Let B-Bus be that a blue bus ran over Ms. Prob's dog, DD be that there is a dead dog, and $\overline{\text{B-Bus}}$ be that a nonblue bus ran over Ms. Prob's dog. Thus,

$$p(\text{B-Bus/DD}) = \frac{p(\text{DD/B-Bus})p(\text{B-Bus})}{p(\text{DD/B-Bus})p(\text{B-Bus}) + p(\text{DD}/\overline{\text{B-Bus}})p(\overline{\text{B-Bus}})}. \quad (1)$$

Simple algebra allows us to reduce Expression 1 to

$$p(\text{B-Bus/DD}) = \frac{p(\text{DD/B-Bus})}{p(\text{DD}/\overline{\text{B-Bus}})} \times p(\text{B-Bus}). \quad (2)$$

Because it is equally likely that the dog would be dead if it was run over by a blue bus or a bus of any other color, the ratio of $p(\text{DD/B-Bus}) \div p(\text{DD}/\overline{\text{B-Bus}})$ must be 1.0 and the value of $p(\text{B-Bus})$, representing the base rate for blue buses, remains .80, for a posterior probability of .80. Of course, if it were the case that a dog is more likely to die if run over by a blue bus than if run over by a nonblue bus or vice versa, then the posterior probability would not be the same as the base rate. In this sense, there is other evidence that is implicitly involved in a Bayesian manner, but its nondiagnostic value does not allow it to impact on the posterior probabilities.

Received November 30, 1990
Revision received July 7, 1991
Accepted November 14, 1991 ■