1-1-1988

Bayes' Theorem in the Trial Process: Instructing Jurors on the Value of Statistical Evidence

David Faigman
UC Hastings College of the Law, faigmand@uchastings.edu

A.J. Baglioni Jr.

Law and Human Behavior, Vol. 12, No. 1, 1988

Bayes' Theorem in the Trial Process

Instructing Jurors on the Value of Statistical Evidence*

David L. Faigman and A. J. Baglioni, Jr.†

The use of statistics and probabilities as legal evidence has recently come under increased scrutiny. Judges' and jurors' ability to understand and use this type of evidence has been of special concern. Finkelstein and Fairley (1970) proposed introducing Bayes' theorem into the courtroom to aid the fact-finder evaluate this type of evidence. The present study addressed individuals' ability to use statistical information as well as their ability to understand and use an expert's Bayesian explanation of that evidence. One hundred and eighty continuing education students were presented with a transcript purportedly taken from an actual trial and were asked to make several subjective probabiliy judgments regarding blood-grouping evidence. The results extend to the trial process previous psychological research suggesting that individuals generally underutilize statistical information, as compared to a Bayesian model. In addition, subjects in this study generally ignored the expert's Bayesian explanation of the statistical evidence.

INTRODUCTION

Statistics and probabilities are receiving increased attention in the law. Since 1960 there has been a dramatic growth of cases using some form of statistical evidence, with the greatest surge coming in the late 1970s (Fienberg & Straf, 1982; Note, 1983). In light of this increased use in the courtroom, legal scholars have begun to debate the merits of various forms of statistical evidence. These commentators have been especially concerned with judges' and jurors' ability to understand and use this evidence. One proposal that has garnered much attention in this debate is

* This study was conducted in fulfillment of the first author's requirement for a Master of Arts degree at the University of Virginia. A version of this paper was presented at the 1983 American Psychological Association convention in Anaheim, California. The authors would like to thank John Monahan, N. Dickon Reppucci, Lois A. Weithorn, and Timothy D. Wilson for comments on earlier drafts of this paper.

† University of Virginia.

0147-7307/88/0300-0001$06.00/0

the proposed explicit use of Bayes' theorem in the trial process. Finkelstein and Fairley (1970) suggested that Bayes' theorem could potentially be used to explain to the trier of fact the proper way to combine certain types of evidence that might otherwise be difficult to understand. Specifically, Bayes' theorem could instruct jurors on how to combine statistical evidence with other, more qualitative, evidence in a trial.[1]

The Finkelstein and Fairley proposal has been vigorously debated; and out of this debate two opposing views have arisen (see generally Weinstein, Mansfield, Abrams, & Berger, 1983). Tribe (1971a), the main proponent of the first view, assailed the use of Bayes' theorem in the trial process, arguing that the trier of

[1] A common expression of Bayes' theorem is as follows:

$$p(A/B) = \frac{p(B/A)\,p(A)}{p(B/A)p(A) + p(B/\text{not } A)\,p(\text{not } A)}$$

As an example of how Bayes' theorem would work in the trial process, consider the following hypothetical case involving a defendant on trial for killing his employer with a ball point pen. It was shown at trial, among other things, that the defendant had fought with his boss over a highly sensitive issue and had stormed out of the office early on the day his boss was killed. The body was found by the cleaning crew at 7:00 p.m. that night. The defendant claimed that he had gone home and stayed there but no one could support this contention. An expert testified that the victim had been stabbed repeatedly with a ball point pen that contained a highly unusual kind of ink. The expert further testified that the defendant's pen contained ink of that same type, and that based on highly reliable data one would expect only 5% of all pens to contain that type of ink. The question is, how should a juror integrate this 5% probability figure into the other evidence already heard? Bayes' theorem addresses this question.

Suppose a hypothetical juror believed, prior to hearing the expert testify, that she was 60% confident (i.e., had a subjective probability of 0.60) that the defendant had stabbed his employer with his pen [i.e., $p(A) = 0.60$]. Therefore, it follows that the probability that he did *not* stab his employer, based on the prior evidence, is 40% [i.e., $p(\text{not } A) = 0.40$]. Further, as the expert testified, the frequency generally of finding the type of ink that was in the murder weapon was only 5%. This is another way of saying that the probability of finding the particular type of ink in the defendant's pen if he did not stab his employer is 5% [i.e., $p(B/\text{not } A) = 0.05$]. The final probability needed to use Bayes' theorem is the probability of finding the ink type in the defendant's pen if he had indeed stabbed his employer. This can be assumed to be 100%, because if the defendant did use his pen to kill his employer then it is certain that its ink type matches the ink type of the murder weapon [i.e., $p(B/A) = 1.0$]. These figures can be substituted in as follows:

$$p(A/B) = \frac{(1.0)(0.60)}{(1.0)(0.60) + (0.05)(0.40)} = 0.967$$

Therefore, Bayes' theorem provides a straightforward device for assessing the probative value of certain evidence that might otherwise be difficult to assess. In this case if the juror had a subjective belief prior to the expert's testimony that there was a 60% chance the defendant had stabbed his employer then she should have a 96.7% subjective belief after hearing the expert.

It should be noted that the merits of using Bayes' theorem to compute the subjective probabilities in this fashion have been debated on philosophical grounds. Unfortunately, space does not permit a discussion of this debate. The reader is referred to several excellent resources that accomplish this task; see in support of Bayes' theorem De Finetti (1972) and Savage (1972); and against, see Horwich (1982) and Shafer (1976). Legal commentators have also addressed these issues; see Brilmayer and Kornhauser (1978), Callan (1982), and Tribe (1971a, 1971b). For a discussion of this debate and its relevance to the psychological issues addressed in the present study, see Faigman (1984).

fact (i.e., a juror or, if a jury is waived, a judge), unsophisticated in the complexity of mathematical techniques, will be overwhelmed by the apparent certainty of those techniques. In *People v. Collins* (1968), the California Supreme Court voiced a similar concern regarding statistical proofs in general when it warned that "[m]athematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search for truth, must not [be allowed to] cast a spell over him" (p. 497). More recently, Saks and Kidd (1981) explicated an alternative view, arguing that rather than be overwhelmed by statistical information, triers of fact are more likely to ignore it. Basing their argument on studies conducted outside a legal context, Saks and Kidd held that people are generally poor processors of quantitative information when qualitative information is available. They suggested that "[t]he more realistic problem is presenting statistical evidence so that people will incorporate it into their decisions at all" (p. 149).

As reflected in the foregoing debate, the Finkelstein and Fairley (1970) proposal rests on two assumptions concerning the capacity of individuals to process information. Foremost is the assumption that triers of fact do not use statistical evidence optimally. Their second assumption is that triers of fact will benefit from a Bayesian explanation of that evidence. Whereas both critics and proponents of Bayesian techniques accept the former assumption, critics argue that a Bayesian explanation will only serve to confuse the trier of fact more than the probabilities it purports to explain.

According to the Federal Rules of Evidence, expert testimony is not admissible unless it will aid the trier of fact to determine some fact in issue. Also, the trial judge must determine whether expert testimony will confuse, prejudice, or waste the time of the trier of fact. The admissibility of an expert's Bayesian calculations, therefore, rests on the degree of assistance of those calculations balanced against the possibility that they will adversely effect the conduct or outcome of the trial. If triers of fact are able to use correctly statistical evidence, testimony intended to explain that evidence would be superfluous and inadmissible. In addition, if the statistical evidence or the Bayesian explanation of that evidence tended to overwhelm or otherwise prejudice the trier of fact, that evidence also would not be admissible. However, if individuals do not attribute proper weight to statistical evidence then testimony explaining the weight this evidence deserves might indeed be admissible.

Use or Misuse of Probabilistic Evidence

In support of their position, Saks and Kidd (1981) relied on a series of studies conducted by Tversky and Kahneman (Kahneman & Tversky, 1972, 1973; Tversky, 1975; Tversky & Kahneman, 1971, 1973, 1974, 1980). Kahneman and Tversky (1973) suggested that people use simplifying strategies or heuristics when assessing multiple bits of information. They found that people tend to ignore quantitative information in favor of more qualitative information when making judgments. Studies in this area have uniformly concluded that people do not intuitively conform to a Bayesian rule for integrating probabilities (Bar-Hillel, 1980; Bar-Hillel & Fischoff, 1981; Borgida & Brekke, 1981; Fischoff & Beyth-Marom, 1983). Although some studies suggest that people process information in a Bayesian fashion but do so conservatively (Edwards, 1968, 1975; Wells & Harvey, 1978), investigations have generally shown that individuals do not intuitively understand rules of statistical inference (Crocker, 1981; Lichtenstein, Slovic, Fischoff, Layman & Coombs, 1978; Nisbet & Ross, 1980). These findings prompted Taylor and Thompson (1982) to comment that apparently people have difficulty recognizing the causal relevance of statistical information when making judgments.

Although the studies cited above demonstrate the problem, the methodologies they adopt do not address the proposal in the legal literature. The proposal to use Bayesian techniques in the trial process incorporates the notion that an explicit probability estimate will be made for the nonstatistical evidence presented at trial. This may be done by the trier of fact or by an expert witness (Ellman & Kaye, 1979). No study has yet tested the introduction of Bayesian techniques into the trial process under either scenario. The present study examines one of these scenarios by comparing subjects' subjective judgments of qualitative information with those same subjects' revised judgments after having heard the quantitative information. In practice, a trier of fact's probabilistic estimation may be made at one of two time points in the trial. One possibility entails the trier's estimation during the trial, when all of the qualitative evidence has been presented but before the quantitative evidence is admitted. Alternatively, and more realistically, at the end of the trial the trier may be asked to consider what his or her estimation of the qualitative evidence had been and then asked to reevaluate it in light of the quantitative evidence.

The purpose of the present study was to assess the two major assumptions underlying the proposed use and criticism of Bayes' theorem in the trial process. First, the study examined individuals' ability to integrate statistical evidence into the other more qualitative evidence of the case. Based on previous research in this area, individuals were expected to underutilize this information. Second, individuals' ability to understand and use a Bayesian presentation was studied. No a priori hypotheses were set forth for this second question.

METHOD

Subjects

Subjects were 180 volunteers (96 females and 84 males) enrolled in continuing adult education courses at several community colleges in Virginia. The mean age of the subjects was 26.6 (median age was 23), with a range of 18 to 59 years.

Procedure

Subjects were randomly assigned to experimental condition and participated in the study in groups ranging from 7 to 35 during their regular evening class time.

It took subjects from 35 to 45 minutes to complete the experiment. The experimenter then conducted an extensive debriefing with each class, including a discussion of their observations and reactions to the study.

The Transcript

The transcript was composed of five parts.

1. *Instructions*. One page of instructions introduced the subjects to the study. The subjects were told that the transcript they would read was from an actual court trial and they were asked to take the role of a juror in the trial. The instructions also notified the subjects that they would find questions within the body of the transcript. It was explained that many of these questions ask what the likelihood that a given statement is true and that they were posed in terms of percentages. Several examples were provided. The subjects were told that there were no right or wrong answers to these questions—all that was requested was their personal estimation. Finally, subjects were told that at the end of the transcript they would be asked to come to a verdict, and to come to a verdict of "guilty" only if the evidence warranted it *beyond a reasonable doubt*.

2. *Pretranscript Questionnaire*. The pretranscript questionaire asked the subjects for personal background information, including sex, age, status as a registered voter (i.e., whether currently registered to vote), juror experience, years of education, years of mathematics, and mathematics courses taken.

3. *Trial Transcript*. The trial transcript was written for use in this study. It contained opening instructions from the judge to the jury as well as opening statements from the prosecuting attorney and the defense counsel. The transcript presented a case in which a male was arrested for breaking and entering a stereo shop and stealing merchandise and cash totaling $3,000.00. During the break-in the defendant allegedly cut his arm on broken glass from the window used to enter the store. The transcript included the direct examination and cross examination of five witnesses: (A) The arresting police officer, who testified to his investigation of the case, and the subsequent arrest of the defendant; (B) An eyewitness, who saw a car similar to the one driven by the defendant outside the stereo shop the night it was broken into. (C) The defendant, who gave vague and inconsistent testimony regarding the stereo receivers found in his apartment and his whereabouts the night of the burglary. He also testified that he had received the cuts and scars on his arm from his current job, construction work; (D) A physician, who testified to taking a blood sample from the defendant and comparing it for blood type with the blood found at the scene of the crime, and he testified that they indeed matched. He also told the jury what proportion of the population had the defendant's blood type and explained to the jury what that figure meant (e.g., in the A blood-type condition: ". . . if you had 100 people that were representative of the population, 40 of them would have type A blood . . ."). (E) A statistician, who testified how Bayes' theorem would evaluate the blood grouping evidence. He presented a chart to the jurors that displayed four prior probabilities (ranging from 1% to 80%) and their accompanying posterior

probabilities. There was no concluding summation by either the defense or prosecution.

4. *Probes*. There were three separate probes, each containing a set of instructions and questions. These probes were placed within the transcript and completed by respondents before going on. (Note that not all subjects completed all three probes—see "Number of Probes," below.) The instructions explained to the subjects that the questions should be answered on the basis of the evidence read up to that point.

On each probe, subjects were asked to state the likelihood that the blood found in the stereo shop was the defendant's blood. In addition, two questions that were specific to the testimony heard just prior to the respective probe were asked. To answer these three questions, subjects had to circle a percentage figure on a scale that reflected their estimate of the likelihood the question was true. The scale had 21 points beginning at "1% or less" increasing in increments of five (rounded off, so the next point was 5%, then 10%, and so on) to a high point of "99% or more." Subjects were also asked, on each probe, to make a determination of guilt based on the evidence already heard.

5. *Posttranscript Questionnaire*. The posttranscript questionnaire contained ten questions. The first question asked subjects to render a verdict of guilty or not guilty. Questions two through six asked subjects to estimate how much weight they gave to each of the five witnesses when deciding on a verdict. The seventh and eighth questions were multiple choice questions designed to gauge the subjects' recollection and understanding of the Bayesian presentation. The ninth question asked the subjects to estimate what they had believed was the likelihood the blood found in the stereo shop was the defendant's blood, before they had read the physician's testimony (the prior probability). The tenth question asked the subjects to give a short explanation for their verdict and to include any observations of the case they had.

Independent Variables

Blood-Grouping Evidence. Blood grouping evidence was selected as the manipulation of probabilistic information for three reasons: It is legally relevant (i.e., it is generally considered admissible; see *State v. Thomas*, 1954; *U.S. v. Kearney*, 1969), it is assumed to be recognized as valid by the lay public, and its impact on the trier of fact has not been previously studied. Subjects were randomly assigned to one of three blood-type conditions, A, O, or AB. Within the transcript a physician testified that for each respective condition either 40%, 20%, or 5% of the population had that blood type.[2]

Number of Probes. The number of probes was varied in an effort to assess

[2] These percentages are not the correct population frequencies. They were used instead of the actual percentages in an effort to increase the range available. The actual percentages for whites in the U.S. are 41.8, 44.4, and 3.8 for A, O, and AB, respectively. For blacks in the U.S., the percentages are 26.5, 49.1, and 4.3, respectively (Blood and Other Body Fluids, 1961). Only one subject's responses had to be eliminated from the analyses because he knew that the percentage quoted by the physician was incorrect.

the effect, if any, quantifying an earlier probability would have on later probability estimates. It was also important to assess the accumulating evidence at several points in the trial. The three most important junctures in the trial were before the physician's statistical evidence, after the physician's statistical evidence, and after the statistician's testimony. Therefore, in one condition, a probe was placed in each of these three places. In a second condition, probes were placed after the statistical evidence and after the statistician's Bayesian testimony. In a third condition, a probe was placed after the statistician's testimony only.

Design. As the above indicates, the experimental design had two distinct phases dependent on the specific questions to be addressed. The over-riding design was a 3 × 3 factorial with blood type and the number of probes as the between-subjects factors. In addition, the responses of those subjects in the two-probe and three-probe conditions were examined as within-subjects factors for changes over the course of the transcript. The primary dependent measures were (1) the subjects' belief that the blood found was the defendant's blood, as determined before the physician's testimony, after the physician's testimony, and after the statistician's testimony, and (2) Bayesian predictions computer calculated using each subject's "prior estimate" (i.e., respondents' subjective probability estimates of the qualitative evidence; Bayesian predictions were calculated from subjects' estimates actually made before the physician's testimony as well as subjects' estimates made at the end of the transcript as to what their prior estimates had been). These measures were analyzed as ratio level data. Also, the subjects' assessment of guilt or innocence was a third dependent measure which was analyzed as a dichotomous categorical variable.

RESULTS

Preliminary analyses indicated that sex, age, education, voter status (i.e., registered versus not registered), or the subjects' mathematics background did not influence responses on any of the dependent measures. Those subjects with a mathematical background did, however, demonstrate a superior understanding of the statistician's Bayesian presentation [G^2 (2) = 8.33, p = .02].

The Probes

Because the subjects had not yet received the blood grouping information, no significant results were expected, and none were found, on the first probe.

As evidenced in Table 1, subjects in the AB condition (that is, the blood type with a 5% population frequency) overall gave significantly more weight to the blood grouping evidence [MANOVA: $F(8,228) = 2.40$, p = .017] than did either of the other two conditions, which did not differ. Individual ANOVAs and *post hoc* contrasts with Bonferoni's inequality imposed indicated that the AB condition had higher estimates than the other two conditions in the belief that the blood

Table 1. Mean Probability[a] Ratings on Probes One, Two, and Three for All Subjects on Whether the Blood Found Was the Defendant's Blood

	Blood type		
Location	A	O	AB
Probe 1 (n = 60)	63.5 (20.2)	61.2 (27.2)	50.2 (27.8)
Probe 2 (n = 120)	62.3 (24.6)	61.8 (29.0)	78.6 (22.1)
Probe 3 (n = 180)	64.4 (25.8)	64.5 (28.8)	77.9 (25.2)

[a] Standard deviations are presented in parentheses.

found in the stereo shop was the defendant's on the second probe [$F(2,119) = 6.1$, $p = .009$] and on the third probe [$F(2,179) = 5.1$, $p = .021$].

The number of probes did affect subjects' judged likelihood estimates that the blood found was the defendant's blood. Subjects who estimated a subjective probability prior to the physician's testimony had higher estimates on the second probe [$F(1,119) = 5.3$, $p = .024$] and on the third probe [$F(2,179) = 5.03$, $p = .008$] than those who did not. Therefore, a comparison of subjects' responses on the first probe they received was conducted, thus creating a one-probe between-subjects design. As Figure 1 illustrates, subjects in the AB condition significantly differed from the other two groups in their probabilistic valuations after the physician's testimony [$F(2,57) = 7.24$, $p = .002$], while the differences were not as great following the statistician's testimony. Moreover, contingency table analysis revealed that subjects in the AB condition were more likely to render a guilty verdict after the physician's testimony than either the A or the O conditions [$G^2(2) = 6.79$, $p = .034$; see Table 2]; whereas no significant differences in verdicts were found for blood-group condition after the statistician's testimony.

Figure 2 illustrates a progressive revision upward of subjects' judgments in the three-probe condition and how those judgments compare to what Bayes' theorem would predict. A within-subjects repeated measures analysis of variance revealed an overall increase in subjective estimates for all blood-type conditions on the likelihood the blood found was the defendant's blood [$F(2,56) = 21.3$, $p < .001$]. Univariate tests revealed that these respondents' subjective estimates significantly increased between the first probe and the second probe [$F(1,57) = 38.3$, $p < .001$] as well as between the second probe and the third probe [$F(1,57) = 4.62$, $p = .036$]. Blood-type specific contrasts conducted on these results revealed that in the AB condition subjects significantly increased their estimates froim probe one to probe two [$F(1,57) = 20.6$, $p < .001$]. Subjects in the 0 condition significantly increased their subjective judgments from probe two to probe three [$F(1,57) = 5.0$, $p = .029$). The upward revisions made by subjects in the A condition were not statistically significant.

To compare subjects' judgments of the likelihood the blood found was the defendant's blood on the second and third probes to Bayesian predictions, a new dependent measure, the Bayesian probability for each subject, was computed using the individual's *prior estimate* and the blood-type condition he or she was in. This new dependent measure was examined via a repeated measures analysis. An ANOVA conducted on the differences between subjects' judgments on the

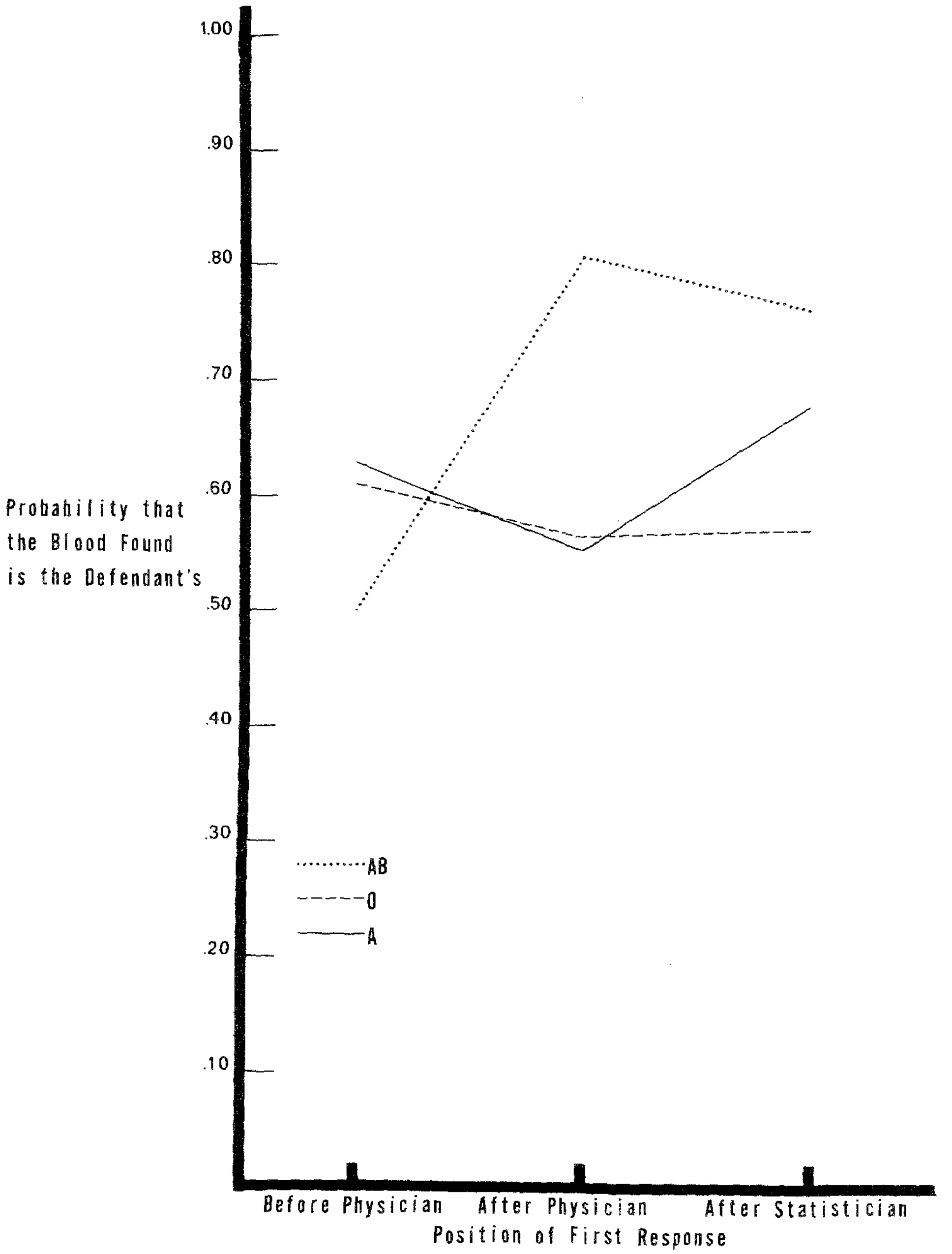

Fig. 1. Comparison of respondents' subjective probability judgments on the first probe they completed (i.e., subjects in the three-probe condition completed their first probe before reading the physician's testimony [$n = 60$]; subjects in the two-probe condition completed their first probe after reading the physician's testimony [$n = 60$]; and subjects in the one-probe condition completed their only probe after reading the statistician's testimony [$n = 60$]).

second probe and their computed Bayesian predictions revealed that the former were significantly lower than the latter [$F(1,57) = 13.7, p < .001$]; while a similar ANOVA conducted on subjects' judgments on the third probe as compared to Bayesian predictions revealed a similar trend [$F(1,57) = 6.6, p = .013$]. Contrasts conducted between subjects' judgments and the Bayesian predictions within each level of blood type revealed that on the second probe subjects' judgments were not significantly different from Bayesian predictions for blood type A but they were for blood types O [$F(1,57) = 5.38, p = .024$] and AB [$F(1,57) = 6.38, p = .010$]. However, on probe three, subjects' judgments were not significantly different for blood types A or O but were for blood type AB [$F(1,57) = 6.38, p = .014$].

Figure 3 depicts a comparison between subjects' probability judgments (for subjects in the two-probe and three-probe conditions) on the second and third probes, as well as their *end-of-transcript estimates* of the prior probability, to what a Bayesian model would predict. Once again, a Bayesian comparison line was calculated, this time using subjects' "prior probability" judgments, as estimated at the end of the transcript, as well as their respective blood grouping condition. For subjects who received three probes, the correlation for judgments made on the first probe and estimates made at the end of the transcript was $r(50) = .537$ ($p < .001$)[3]; and a t-test on these two measures revealed no significant difference.

Repeated measures analysis conducted on the responses of subjects who received multiple probes revealed an overall increase from their prior probabilities estimated at the end of the transcript as compared to the judgments made on the second and third probes [$F(2,116) = 116.0, p < .001$] (see Figure 3). Univariate tests revealed that these subjects' judgments significantly increased from the prior probabilities estimated at the end of the transcript to the second probe [$F(1,117) = 31.05, p < .001$] and from the second probe to the third probe [$F(1,117) = 3.61, p = .06$]. Contrasts conducted on these results revealed that subjects in the AB condition significantly increased their judgments from those

Table 2. Frequency of Guilty versus Not Guilty Verdicts after the Physician's Testimony for Subjects who Received Two Probes

	Verdict	
Blood type	Guilty	Not guilty
A	7	13
O	10	10
AB	15	5

[3] Note that, though significant, this correlation is surprisingly low considering that it should be 1.0. The query at the end of the transcript asked subjects to recall their stated prior subjective probability. This result may lend some support to earlier research that indicated that individuals cannot separate what they "knew" from what they "know" (e.g., Koriate, Lichtenstein, & Fischoff, 1980; Lichtenstein, Fischoff, & Phillips, 1977).

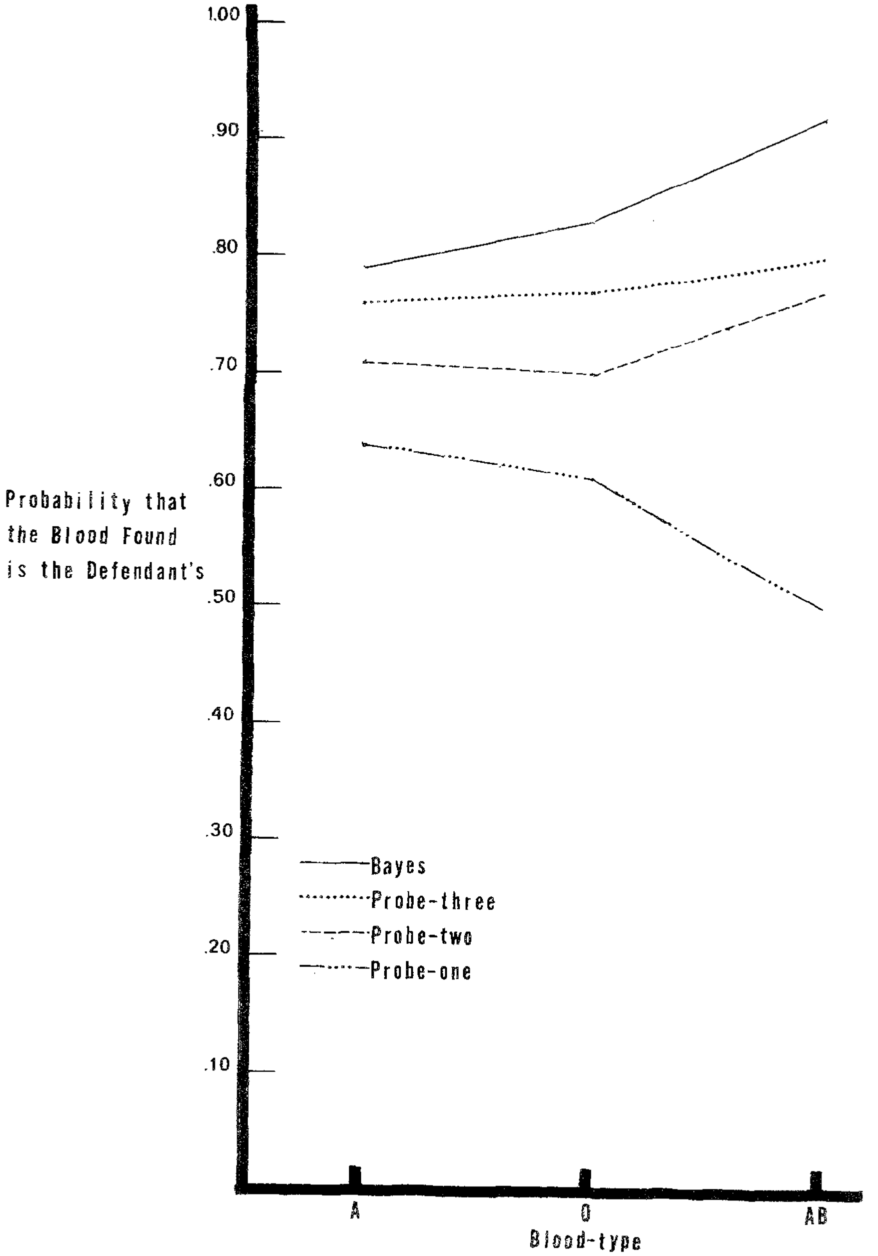

Fig. 2. Comparison of respondents' subjective probability judgments estimated prior to the physician (probe one), after the physician (probe two), and after the statistician (probe three) to a Bayesian comparison line (for subjects in the three-probe condition, $n = 60$).

estimated at the end of the transcript to those made on the second probe [$F(1,117) = 16.97$, $p < .001$] but not from the second to the third probe. Subjects in the O condition did not significantly increase their prior judgments to those made on the second probe, but did significantly revise their judgments from the second probe

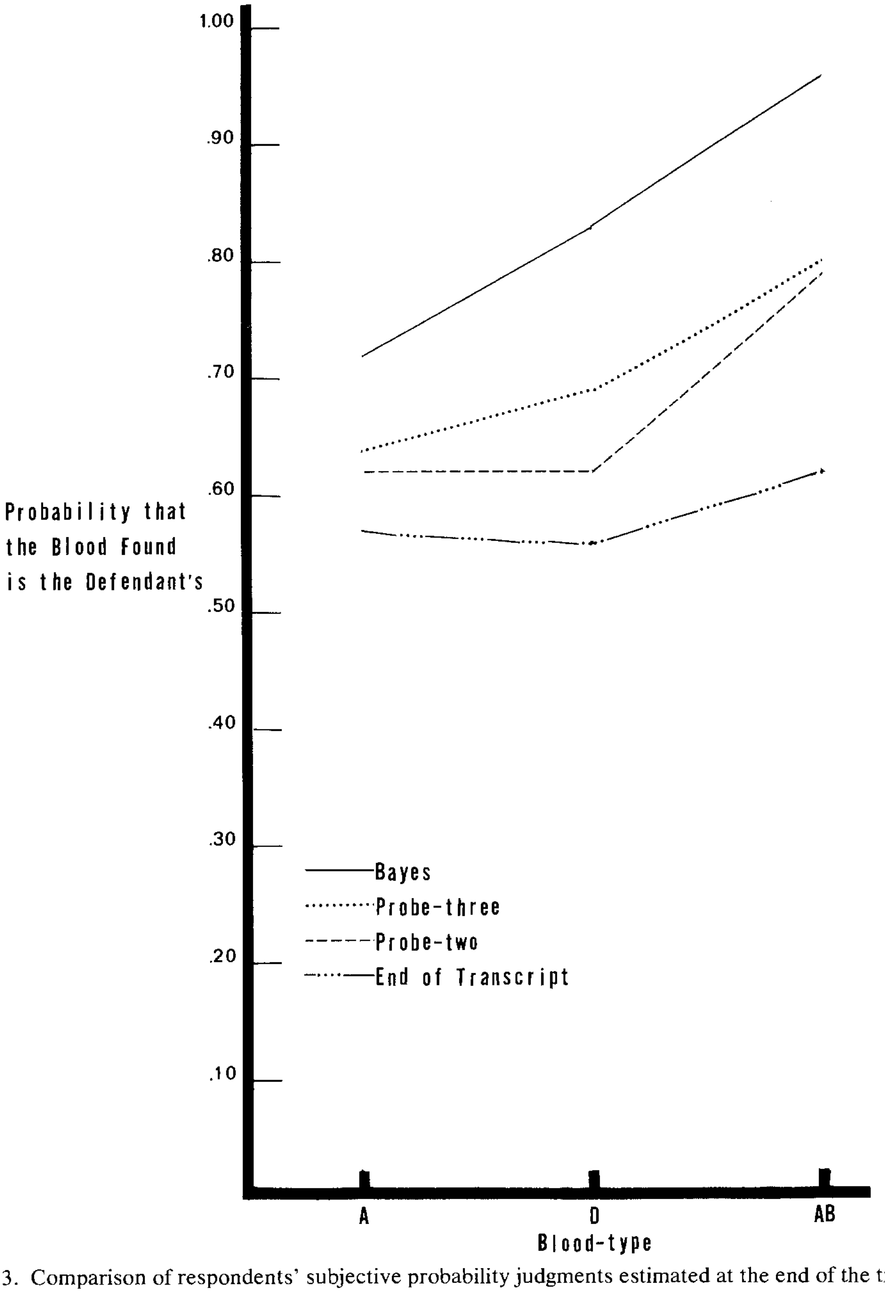

Fig. 3. Comparison of respondents' subjective probability judgments estimated at the end of the transcript (prior probability), after the physician (probe two), and after the statistician (probe three) to a Bayesian comparison line (for subjects in the two-probe and three-probe conditions, $n = 120$).

to the third probe [$F(1,117) = 9.0, p = .003$]. Subjects in the A condition did not significantly revise their judgments.

An ANOVA conducted on the difference scores between subjects' judgments on the second probe and Bayesian predictions (see Figure 3) revealed that the former were significantly lower than the latter [$F(1,117) = 49.9$, $p < .001$]; a similar ANOVA conducted on subjects' judgments on the third probe as compared to Bayesian predictions revealed a similar significant pattern [$F(1,117) = 35.44$, $p < .001$]. Contrasts conducted between subjects' judgments and the Bayesian predictions within each level of blood type revealed that responses on probe two for blood types A, O, and AB were all significantly different from the Bayesian predictions [$F(1,117) = 5.84$, $p = .017$; $F(1,117) = 29.5$, $p < .001$; and $F(1,117) = 19.25$, $p = .001$, respectively]. Similar results were obtained for contrasts between responses on the third probe and Bayesian predictions for blood types A, O, and AB [$F(1,117) = 4.68$, $p = .033$; $F(1,117) = 15.02$, $p < .001$; and $F(1,117) = 18.25$, $p < .001$, respectively].

Posttranscript Questionnaire

Log-linear contingency table analysis revealed no significant relationship between final determinations of guilty versus not guilty verdicts and either blood-type condition or number of probes completed. Contrasts conducted on the weights attributed to each of the witnesses revealed that subjects attributed significantly more weight to the physician than to the police officer [$F(1,178) = 9.63$, $p = .002$], the eyewitness [$F(1,178) = 65.51$, $p < .001$], the defendant [$F(1,178) = 45.25$, $p < .001$], and the statistician [$F(1,178) = 65.67$, $p < .001$]. Also more weight was attributed to the police officer than to the eyewitness [$F(1,178) = 45.34$, $p < .001$], the defendant [$F(1,178) = 25.42$, $p = .002$), and the statistician [$F(1,178) = 22.85$, $p < .001$].

No differences were found between the two multiple choice questions designed to gauge respondents' recollection and understanding of the Bayesian presentation on the frequency of getting one correct rather than the other. In all, 14% of the subjects answered both questions correctly, while 43.6% answered one correctly and 42.5% answered neither question correctly. As noted above, subjects with a mathematics background were more likely to answer these questions correctly than subjects with little or no mathematics background. Importantly, however, no relationship was obtained between subjects' understanding of the Bayesian presentation and either verdicts or use of the statistician's conclusions. Subjects who understood the presentation were no more or less likely to use the information than subjects who did not understand it.

DISCUSSION

The results of this study generally support and extend to the trial process the psychological research available on individuals' use of statistical information. Respondents significantly underutilized the statistical evidence. Indeed, except in the AB blood-group condition (i.e., presented with the 5% figure) and where subjects explicitly stated a prior probability (i.e., the three-probe condition), respondents virtually ignored the statistical evidence. Further, the subjects did not conform to the expectations of either critics (e.g., Tribe, 1971a, 1971b) or proponents (e.g., Finkelstein & Fairley, 1970, 1971) of the courtroom use of Bayes' theorem; they were not overwhelmed by this statistical theorem, nor did they accept the statistician's conclusions. Overall, subjects who estimated a prior probability on the first probe revised their estimates after the statistician testified as did subjects who estimated a prior probability at the end of the transcript. Yet, for both of these groups, the revisions remained significantly below the probabilities about which the statistician had testified (i.e., a Bayesian model). These findings lend support to previous research findings that identified individuals' reluctance to use statistical information when making causal attributions (Saks & Kidd, 1981).

Subjects in the AB condition significantly utilized the blood grouping evidence while the other two conditions failed to recognize its relevance. They not only differed in the valuation of whether the blood found at the scene was the defendant's, but they also were more likely to render a guilty verdict. It is unclear why subjects found the AB blood-type to be particularly probative. One reason may be the extremity of the 5% figure. No studies have systematically investigated individuals' differential use of varying degrees of statistical information. Typically, researchers have used modest ratios such as 80/20 or 70/30. It may be that individuals process probability figures by giving some weight to extreme figures and little or no weight to modest figures, but do not discriminate between the two in any refined manner (Nisbet & Ross, 1980). Support for this interpretation comes from the finding that even where subjects significantly utilized the statistical evidence they nevertheless underutilized it when compared to a Bayesian model.

Ajzen (1977) found that individuals look for causal explanations of events and that statistical information was more fully utilized when it fit a person's causal schemata. Apparently, people heuristically select information that directly pertains to some causal explanation of an event (Borgida & Brekke, 1981; Tversky & Kahneman, 1980). Here, respondents confronted with an extreme probability figure may have been impressed by the significance of the figure and thus adopted it as support for their decision. This raises an interesting question for future research: is there an interaction between the causal relevance and the extreme degree of statistical evidence?

A possibly related set of findings involved the effect of quantifying a prior probability on later likelihood judgments. There are several possible explanations as to why subjects who estimated a probability prior to the physician's testimony subsequently had higher probability judgments than subjects who did not. Schum and Martin (1982) obtained a similar result when they compared individuals' subjective probability judgments for segments of a collection of evidence to individual's judgments for the entire collection of evidence. The most common explanation they identified was the "misaggregation hypothesis," which states that individuals use heuristic strategies to process information and therefore some of the information available is not utilized or fully processed (Kahneman & Tversky, 1973). Another explanation is that people have a "response bias" against using large numbers (Ducharme, 1970) or simply do not want to overestimate the relevance of the information available. Although these explanations may play some part in the findings of the present study, another explanation seems more cogent.

Subjects who received two probes did not differ in their final probability judgments from subjects who received one probe. This finding suggests that simply estimating an earlier probability is not the important factor. Rather it appears that explicitly stating a prior probability before the physician testified sensitized subjects to the blood evidence.

This finding suggests that explicit quantification of the nonstatistical evidence may increase the utilization of the statistical evidence. However, it is probably not the quantification of the nonstatistical evidence that is important, but instead the attention that is drawn to the statistical evidence. It may be that when a manipulation increases the attention given to a probability figure, or when a probability figure is sufficiently extreme to garner attention, it is more likely to be used in a fashion consistent with a Bayesian model.

From a practical standpoint, however, even if courts allow an expert to testify to a Bayesian interpretation of the evidence, they are unlikely to allow jurors to make explicit prior probability valuations before hearing the statistical evidence. This, of course, places a caveat on the finding that explicit quantification of the nonstatistical evidence increased the use of the statistical evidence. Tribe (1971a) contended that if courts do not allow this explicit quantification then there is the possibility the trier of fact will double-count the probabilistic evidence when evaluating the statistician's testimony. No evidence for this assertion was found in this study. In fact, subjects who received only one probe, after the statistician's testimony, actually quoted somewhat lower probability valuations than subjects who responded right after the physician's testimony (see Figure 1). This result may be a function of the subjects' general perception of the statistician as a witness.

Overall, subjects did not differentially weigh any of the witnesses' testimony on the basis of blood type or number of probes. They rated the physician as the witness to whom they gave the most weight when determining a verdict. The police officer was accorded the second most weight. The statistician was only given as much weight as the eyewitness, who admitted to drinking the night of the burglary, and the defendant, whom the subjects saw, on average, as only 36% likely to be telling the truth. However, the statistician's testimony was seen, on average, as accurate (mean = 75.9). Apparently, subjects felt as the following subject succinctly put it: "I personally don't put much weight on statistical deductions as proof of anything."

Although the results of this study have several important implications for the trial process, they also have many limitations. Principally, the methodology employed lacked the complexity and meaning of an actual trial. The subjects did not deliberate their responses in a group. Verdicts did not have consequences for any real person (save perhaps the authors). The experimental materials had to be read and the entire study only lasted approximately 40 minutes per subject. These factors may or may not have differentially influenced the results and hence have limited the generalizability of the study (Bray & Kerr, 1979; Weiten & Diamond, 1979).

Nonetheless, this study has several strengths that buttress the conclusions drawn from it. Foremost, it is based on a foundation of sound empirical research that has obtained consistent results, utilizing various paradigms. Legal decision makers who desire to use social scientific findings to aid in difficult decision making would be well advised to rely on studies that have a sound theoretical foundation (Lind & Walker, 1979).

Legal decision makers face several difficult issues involving the proposed use of Bayesian techniques in the trial process. As previous research has shown, and this study has supported, individuals tend to underutilize statistical information, although extreme probability figures may be utilized more than modest figures. This study has also found that explicit quantification of a prior probability does not hinder, and may in fact aid, an individual's use of statistical information. The results also suggest, contrary to Tribe's (1971a) assertion, that an expert's Bayesian formulation will not overwhelm the average trier of fact. Courts, it seems, should be less concerned with jurors being overwhelmed by the complexity of statistical techniques and more concerned with impressing upon jurors the relevance of those techniques.

REFERENCES

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

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44,* 211–233.

Bar-Hillel, M., & Fischoff, B. (1981). When do base rates affect predictions? *Journal of Personality and Social Psychology, 41,* 671–680.

Blood and other body fluids. (1961). Washington, DC: Federation of American Societies for Experimental Biology.

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

Bray, R. M., & Kerr, N. L. (1979). Use of the simulation method in the study of jury behavior: Some methodological considerations. *Law and Human Behavior, 3,* 107–119.

Brilmayer, L. & Kornhauser, L. (1978). Review: Quantitative methods and legal decisions. *University of Chicago Law Review, 46,* 116–153.

Callan, 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.

Crocker, J. (1981). Judgement of covariation by social perceivers. *Psychological Bulletin, 90,* 272–292.

De Finetti, B. (1972). *Probability, induction and statistics: The art of guessing.* London: John Wiley & Sons.

DuCharme, W. M. (1970). A response bias explanation of conservative inference. *Journal of Experimental Psychology, 85,* 66–74.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmutz, (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.

Edwards, W. (1975). Comment. *Journal of the American Statistical Association, 70,* 291–293.

Ellman, I. M., & Kaye, D. (1979). Probabilities and proof: Can HLA and blood group testing prove paternity? *New York University Law Review, 54,* 1131–1162.

Faigman, D. L. (1984). Setting the odds on justice: Statistics and probabilities in the trial process. Unpublished master's thesis, University of Virginia, Charlottesville.

Fienberg, S. E., & Straf, M. C. (1982). *Statistical assessments as evidence.* Technical Report No. 237, Committee on National Statistics, National Academy of Sciences.

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

Finkelstein, M., & Fairley, W. B. (1971). A comment on "Trial by mathematics," *Harvard Law Review, 84,* 1801–1809.

Fischoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239–260.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Koriat, A., Lichtenstein, S., & Fischoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118.

Lind, E. A., & Walker, L. (1979). Theory testing, theory development, and laboratory research on legal issues. *Law and Human Behavior, 3,* 5–19.

Lichtenstein, S., Slovic, P., Fischoff, B., Layman, M., & Coombs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 551–578.

Lichtenstein, S., Fischoff, B., & Phillips, C. D. (1977). Calibration of probabilities: The state of the art. In H. Jungerman and G. deZeeuw (Eds.), *Decision making and change in human affairs* (pp. 275–324). Amsterdam: D. Reidel.

Nisbet, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, New Jersey: Prentice-Hall.

Note (1983). Admissibility of mathematical evidence in criminal trials. *American Criminal Law Review, 21,* 55–79.

People v. Collins, (1968). 66 Cal. Reptr. 497.

Saks, M. J., & Kidd, R. (1981). Human information processing and adjudication: Trial by heuristics. *Law and Society Review, 15,* 123–160.

Savage, L. J. (1954). *The foundations of statistics.* New York: John Wiley & Sons.

Schum, D. A., & Martin, A. W. (1983). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review, 17,* 105–151.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton, New Jersey: Princeton University Press.

State v. Thomas, (1954) 275 P. 2d 408.

Taylor, S., & Thompson, S. C. (1982). Stalking the elusive "vividness" effect. *Psychological Review, 89,* 155–181.

Tribe, L. H. (1971a). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review, 84,* 1328–1393.

Tribe, L. H. (1971b). A further critique of mathematical proof. *Harvard Law Review, 84,* 1810–1820.

Tversky, A. (1975). Assessing uncertainty. *Journal of the Royal Statistical Society, 36B,* 148–159.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

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

U.S. v. Kearney, (1969). 420 F.2d 170.

Weinstein, J. B., Mansfield, J. H., Abrams, N., & Berger, M.A. (1983). *Cases and materials on evidence* (7th ed.). Mineola, New York: The Foundation Press.

Weiten, W., & Diamond, S. S. (1979). A critical review of the jury simulation paradigm: The case of defendant characteristics. *Law and Human Behavior, 3,* 71–93.

Wells, G. L., & Harvey, J. H. (1978). Do people use consensus information in making causal attributions? *Journal of Personality and Social Psychology, 35,* 279–293.