Do Juries Meet Our Expectations?

Hal R. Arkes\textsuperscript{1,3} and Barbara A. Mellers\textsuperscript{2}

Surveys of public opinion indicate that people have high expectations for juries. When it comes to serious crimes, most people want errors of convicting the innocent (false positives) or acquitting the guilty (false negatives) to fall well below 10\%. Using expected utility theory, Bayes' Theorem, signal detection theory, and empirical evidence from detection studies of medical decision making, eyewitness testimony, and weather forecasting, we argue that the frequency of mistakes probably far exceeds these "tolerable" levels. We are not arguing against the use of juries. Rather, we point out that a closer look at jury decisions reveals a serious gap between what we expect from juries and what probably occurs. When deciding issues of guilt and/or punishing convicted criminals, we as a society should recognize and acknowledge the abundance of error.

KEY WORDS: jury decision making; Bayes' Theorem; signal detection theory; expected utility theory.

Jurors face a difficult task. Their job is to simultaneously punish individuals who commit violent acts and acquit those who are innocent. The job is not easy because the evidence is often ambiguous. Innocent persons do not always appear innocent, and guilty persons do not always appear guilty. Despite the difficulty of the task, our legal system rests on the assumption that juries are generally accurate. We will use three analytic tools to show that jurors are very unlikely to perform at levels of accuracy congruent with our expectations.

Before we begin, we want to acknowledge that not all juries want to be accurate. On occasions, juries make decisions that convey attitudes—regardless of the guilt or innocence of the defendant (Sunstein, 1996). Opinions might be expressed in the form of jury nullification (Kadish & Kadish, 1973), as documented in trials of slavery, prohibition, euthanasia, and prostitution. In this paper, we put these cases aside and focus solely on the issue of accuracy. Our approach is a strictly normative one, in that we assume that jurors wish to maximize utility by minimizing the number of falsely convicted and falsely acquitted defendants. Our claim is that juries make many more errors than most of us say "tolerable."

\textsuperscript{1}Department of Psychology and Center for HOPES, Ohio State University, Columbus, Ohio.
\textsuperscript{2}Haas School of Business, University of California, Berkeley, California.
\textsuperscript{3}To whom correspondence should be addressed at Department of Psychology, Ohio State University, 240N Lazenby Hall, 1827 Neil Avenue Mall, Columbus, Ohio 43210; e-mail: arkes.1@osu.edu.

625
THE DISTRIBUTIONS OF EVIDENCE

During a trial, jurors weigh the testimony, opinions, and facts presented to them and form an overall impression of the culpability of the defendant. The evidence they face can be represented with probability distributions, as shown in Fig. 1. The perceived culpability of the defendant is a subjective continuum with guilt increasing to the right. Perceived culpability is a random variable that may change throughout the trial. We are interested in the value of this random value \( x_i \) when the trial is over and jury deliberation begins.

The probability distributions in Fig. 1 show what might occur over an infinite number of trials. The Y-axis is the chance that the value of the random variable \( x_i \) could arise from a distribution of innocent persons or a distribution of guilty persons. The distributions are normal, and guilty persons are, on average, more culpable than innocent persons. Despite this difference, the distributions overlap. Evidence is often ambiguous.

Jurors make two decisions. In addition to the task of determining the culpability of the defendant, the juror must decide how much evidence he requires to convict. The amount of evidence required for a conviction is represented as a threshold \( T \) in Fig. 1. If the value of the random variable \( x_i \) lies to the left of \( T \), the juror acquits; otherwise, he convicts. The chance of a false acquittal (acquitting a guilty person) is the region of the distribution of guilty persons to the left of \( T \). The chance of a false conviction (convicting an innocent person) is the region of the distribution of innocent persons to the right of \( T \).

Different threshold locations imply different error rates, and as the probability of one error decreases, the probability of the other error increases. Jurors with higher thresholds who require more evidence to convict have lower chances of false convictions, but higher chances of false acquittals. Jurors with lower thresholds who require less evidence to convict have higher chances of false acquittals, but lower chances of false convictions. Unless the distributions can be moved further apart, it is impossible to simultaneously decrease both errors.

![Fig. 1. Distributions of perceived culpability of innocent and guilty defendants. "T" and "T9" represent two possible thresholds a juror might set, above which the juror would vote "guilty" and below which a juror would vote "not guilty."](image)
WHERE “SHOULD” THE THRESHOLD BE?

The optimal location of the threshold depends on the juror’s goals. Given certain goals, the juror needs three theoretical tools. We now discuss each and show how it applies to jury decision making.

**Expected Utility Theory**

One goal is to maximize expected utility. This normative goal implies that the juror will acquit if the expected utility of acquitting exceeds that of convicting; otherwise, he will convict. To achieve this goal, the juror needs expected utility theory (von Neumann & Morgenstern, 1947). To illustrate, consider the four possible outcomes of a jury trial shown in Fig. 2. Columns are decisions, and rows are states of the world. Good decisions are “Correct Rejections” (acquitting the innocent) and “Hits” (convicting the guilty). Bad decisions are “Misses” (acquitting the guilty) and “False Alarms” (convicting the innocent). The psychological costs and benefits of the outcomes, called utilities, are represented as $U_{CR}$ and $U_H$ for correct decisions and $U_M$ and $U_{FA}$ for incorrect decisions. Let us suppose the juror can quantify all four utilities.

To find the threshold, the juror should set the expected utility of “Acquitting” given the evidence equal to the expected utility of “Convicting” given the evidence, as follows:

$$EU(\text{“Acquit”} \mid x_i) = EU(\text{“Convict”} \mid x_i)$$ (1)

---

**Fig. 2.** Four possible outcomes of a jury trial.
This expression implies
\[ p(G \mid x_i)U_M + (1 - p(G \mid x_i))U_{CR} = p(G \mid x_i)U_H + (1 - p(G \mid x_i))U_{FA}, \] (2)
where \( p(G \mid x_i) \) is the juror's posterior belief that the defendant is guilty after listening to, evaluating, and integrating the evidence. The juror can solve for \( p(G \mid x_i) \) in Eq. 2 as
\[ p(G \mid x_i) = \frac{(U_{CR} - U_{FA})}{(U_{CR} + U_H - U_{FA} - U_M)}. \] (3)
This probability is needed to find the optimal threshold.

Sometimes it is convenient to transform a probability \( p \) into an odds ratio, \( p/(1 - p) \). The posterior probability of guilt, \( p(G \mid x_i) \), can be expressed as a posterior odds of guilt when it is divided by
\[ 1 - p(G \mid x_i) = p(I \mid x_i) = \frac{(U_H - U_M)}{(U_{CR} + U_H - U_{FA} - U_M)}. \] (4)
The posterior odds, \( p(G \mid x_i)/p(I \mid x_i) \), then becomes
\[ p(G \mid x_i)/p(I \mid x_i) = \frac{(U_{CR} - U_{FA})}{(U_H - U_M)}. \] (5)

**Bayes' Theorem**

The next tool needed by the juror is Bayes' Theorem.\(^4\) This theorem says that the posterior odds of guilt can also be written as
\[ p(G \mid x_i)/p(I \mid x_i) = \frac{[p(x_i \mid G)/p(x_i \mid I)][p(G)/p(I)]}{}, \] (6)
where \( p(x_i \mid G) \) and \( p(x_i \mid I) \) on the right-hand side of the equation are the probabilities of the evidence, given guilt and innocence, respectively. These probabilities are the heights of the curves under the distributions of guilt and innocence for a given value of \( x_i \) in Fig. 1. The ratio of these probabilities, \( p(x_i \mid G)/p(x_i \mid I) \), is called the likelihood ratio, and the optimal location for the threshold is related to the likelihood ratio.

Let us consider how changes in the evidence would influence the likelihood ratio. Suppose the evidence presented is more diagnostic of innocence than of guilt. At lower levels of \( x_i \), \( p(x_i \mid I) \) is larger than \( p(x_i \mid G) \), and the likelihood ratio would be smaller. Now suppose the evidence is more diagnostic of guilt than of innocence. At higher levels of \( x_i \), \( p(x_i \mid G) \) is larger than \( p(x_i \mid I) \), and the likelihood ratio would be larger.

**Signal Detection Theory**

At this point, the juror needs a third tool called signal detection theory (Green & Swets, 1966). Signal detection theory has two parameters, \( d' \) and \( \beta \). The first parameter, \( d' \), reflects the observer's ability to discriminate signals from noise. In a jury

\(^4\)There is an extensive literature pertaining to the use of Bayes' Theorem to incorporate evidence with prior beliefs (e.g., Feinberg & Schervish, 1986; Finkelstein & Fairley, 1970; Mossman, 2000). Much of it suggests that jurors have a poor understanding of testimony based on Bayes' Theorem (Koehler, 1996; Koehler, Chia, & Lindsey, 1995). Our analysis uses Bayes' Theorem to develop a normative framework. Therefore, the literature on probability judgments is not relevant.
trial, $d'$ is $(\mu_R - \mu_I) / \sigma$, the difference between the means of the distributions for guilty and innocent persons in Fig. 1 divided by the standard deviation of innocent persons. When distributions are normal and variances are equal as in Fig. 1, $d'$ is simply $Z_{p(H)} - Z_{p(FA)}$, the difference between the standard scores associated with probabilities of hits and false alarms.

The second parameter, $\beta$, reflects the observer's tendency to say "Signal" or "Noise." $\beta$ is related to the decision threshold or $T$ in Fig. 1. According to signal detection theory, $\beta$ is defined as the likelihood ratio or $p(x_i | G)/p(x_i | I)$. By rearranging Eq. (6) and solving for the likelihood ratio, $\beta$ can also be expressed as

$$\beta = \left[\frac{p(G \mid x_i)}{p(I \mid x_i)}\right] \left[\frac{p(I)}{p(G)}\right].$$

(7)

This expression tells the juror that $\beta$ depends on the posterior odds of guilt (after the evidence has been presented) and the prior odds of innocence (before the evidence is presented). If the juror replaces the posterior odds of guilt with utilities, as shown in Eq. (5), $\beta$ becomes

$$\beta = \left[\frac{(U_{CR} - U_{FA})}{(U_H - U_M)}\right] \left[\frac{p(I)}{p(G)}\right].$$

(8)

Now the juror knows that the location of the optimal threshold is the value of $x_i$ that corresponds to $\beta$ when $\beta$ is $\left[\frac{(U_{CR} - U_{FA})}{(U_H - U_M)}\right] \left[\frac{p(I)}{p(G)}\right]$. Simply put, the optimal location for the threshold is a simple function of the juror's utilities and prior beliefs about guilt and innocence. $\beta$ can easily be transformed to a value of $x_i$.

In summary, a juror who wants to maximize the expected utility of his decision should listen to the evidence that is sampled from an unknown distribution. After the trial, he has determined a value of $x_i$. Then he sets his threshold, $\beta$, at a location determined by his utilities and prior beliefs, as in Eq. 8. If $x_i$ exceeds the threshold, the juror convicts. Otherwise, he acquits. Of course, jurors do not perform these calculations. Our analysis is purely normative and shows how a juror should decide if his goal is to maximize the expected utility of his decision.

Others have developed models of jury decision making that employ some of the same tools (e.g., Connolly, 1987; Cullison, 1971; Kaplan, 1968; Marshall & Wise, 1975; Mowen & Linder, 1979; Nagel & Neef, 1978; Thomas & Hogue, 1976). However, these approaches differ from ours in several respects. (See Footnote 5.) Most importantly,

5Other methods proposed in the literature (e.g. Connolly, 1987), conceptualize the decision as one based on no trial evidence. The juror assigns utilities to outcomes and then solves for the prior probability of guilt, $p(G)$, that makes the expected utility of convicting equal to the expected utility of acquitting. This probability is the threshold; if a juror's prior belief in guilt exceeds the threshold, he convicts. Otherwise, he acquits. The analysis is done as follows: $EU("Acquit") = EU("Convict")$

or

$$p(G)U_M + (1 - p(G))U_{CR} = p(G)U_H + (1 - p(G))U_{FA}.$$ 

Solving for $p(G)$ yields:

$$p(G) = \frac{(U_{CR} - U_{FA})}{(U_{CR} + U_H - U_{FA} - U_M)},$$

which represents the probability threshold. But this is not what happens in a trial. Jurors base their decisions on evidence $x_i$. For this reason, we believe that our method, which does more fully take the evidence into account, is more complete than these earlier analyses.
we argue that the juror who wants to maximize his expected utility must use all three of the theoretical tools.

We also wish to point out that our model does not require jurors to think in terms of probabilities. Hence our analysis is agnostic with regard to the question as to whether jurors do or do not use probabilistic reasoning in deciding whether to acquit or convict (e.g., Nesson, 1979). Our central point is that expected utility theory, Bayes’ Theorem, and signal detection theory can be used to show that the diagnostic performance of jurors is extremely likely to be far lower than our expectations, irrespective of the jurors’ willingness to base their decision on probabilistic considerations or any other cognitive strategy.

**SOME ILLUSTRATIVE THRESHOLDS**

Imagine a juror believes the two correct decisions (Hits and Correct Rejections) are equally beneficial and the two incorrect decisions (Misses and False Alarms) are equally costly. Let us further suppose the juror believes there is a 50% prior probability the defendant is guilty. The likelihood ratio for this juror becomes

\[ \beta = \frac{p(x_i | G)}{p(x_i | I)} = \frac{(1 - 0)}{(1 - 0)} \frac{0.5}{0.5} = 1. \tag{9} \]

A likelihood ratio of 1 means that the threshold is placed where the heights of the curves in Fig. 1 are identical. This point is labeled \( T \).

A juror with the same utilities who believes there is a 10% prior probability the defendant is guilty has a likelihood ratio of 9 because

\[ \beta = \frac{p(x_i | G)}{p(x_i | I)} = \frac{(1 - 0)}{(1 - 0)} \frac{0.9}{0.1} = 9. \tag{10} \]

This threshold, labeled \( T_0 \) in Fig. 1, is higher than \( T \). A lower prior probability of guilt translates into a higher threshold for conviction.

**FIXING THE PROBABILITY OF ERRORS AND ESTIMATING \( d' \)**

Suppose we work backwards. With the assumptions of normal distributions and equal variances, we can fix the error rates and solve for \( d' \), the average difference between the probability distributions in Fig. 1. Imagine error rates (False Alarms and Misses) of 1%. With this level of accuracy, \( d' \) must be 4.66. The mean of the distribution for guilty persons should be almost five standard deviations higher than the mean of the distribution for innocent persons! Values of \( d' \) this large are virtually never obtained.

What happens if we assume that the probabilities of false convictions and false acquittals are 5%? This level of accuracy implies that \( d' \) is 3.3. The mean of the distribution for guilty persons must be over three standard deviations higher than the mean for innocent persons. Again, evidence this strong is very unlikely. What happens if both error rates are 10%? Even these error rates require a \( d' \) of 2.6. Later in this paper we will show that even medical tests using highly sophisticated technology cannot achieve these levels of \( d' \). Hence we contend that it would very surprising if unaided jurors were able to attain the lofty performance levels consistent with a \( d' \) of approximately 3.0.
What error rates do people say they can tolerate? The two most direct attempts to determine the answer to this question were undertaken by McCauliff (1982) and Simon (1969). McCauliff (1982) asked a sample of 171 federal judges to provide a numerical definition of “reasonable doubt,” the threshold above which defendants would be found guilty in a criminal trial. The median response was 90%. The three most frequent responses were 90, 95, and 100%, with 33, 18, and 12% of the judges choosing each of these responses, respectively. Simon (1969) found similar results, using a sample of state and federal judges. The judges’ interpretations of “reasonable doubt” had a median of 88%, a mean of 89%, and a mode of 100%. When a serious crime such as murder was considered, judges’ values were even higher.

Simon (1969) also asked judges what probability they thought jurors would require before they would find the evidence probative beyond a reasonable doubt. The judges thought that jurors’ thresholds were very similar to their own. The implication of these results is that judges, and probably jurors, are willing to convict if the probability of a false conviction is approximately 5–10%.6

Unfortunately neither McCauliff (1982) nor Simon (1969) asked the judges about the other type of error—namely false acquittals. To get some idea what people view as acceptable probabilities, we asked 133 students at Ohio State University to tell us their maximum acceptable false alarm rate and their maximum acceptable miss rate. Exact wordings of the questions are presented in the Appendix. If the students’ false alarm rates were similar to those of the judges, then perhaps the students’ false acquittal rates might also be reasonable approximations of what the judges would say.

The majority of students were willing to tolerate fewer false convictions than false acquittals. The median response for the largest acceptable false conviction rate was 5%, quite comparable to the responses of the judges. The students’ median response for the largest acceptable false acquittal rate was 8%. To achieve false alarm and miss rates of 5 and 8%, respectively, $d'$ must be at least 3.0. Even with the very best of current technology and investigative methods, courtroom evidence is rarely this diagnostic.

**IS 8% A REASONABLE FALSE ACQUITTL RATE?**

The fact that the students’ deemed the largest acceptable false acquittal rate to be 8% may seem surprising in light of the opinion expressed by many legal commentators that false acquittals are much less troubling than false convictions. Therefore one might suppose that these scholars would be willing to tolerate a false acquittal rate far higher than 8% in order to achieve a low false conviction rate of approximately 5%. For example, one reviewer pointed out that according to Blackstone’s ratio it is better to let 10 guilty persons go free than to convict 1 innocent person. Would not this require that to achieve a false conviction rate of only 5% one

---

6More indirect methods of ascertaining the threshold for “reasonable doubt” have generated lower estimates (e.g., Dane, 1985; Nagel & Neef, 1978). Rather than asking respondents directly about their threshold, these investigators derive the threshold based upon the utilities the respondents assign to the various outcomes.
would have to tolerate a false acquittal rate of 50%, not the puny 8% given by our student sample?

There are three reasons why we think not. First, as DeKay (1996) pointed out, a conviction threshold of 95%, which allows for a 5% false conviction rate, does not imply any particular ratio of false convictions to false acquittals. This ratio is also influenced by the accuracy with which guilty persons are distinguished from innocent ones and the base rate of guilty persons in the population of defendants. Thus Blackstone's ratio of 10 false acquittals to 1 false conviction simply does not presuppose any particular false acquittal rate.

Second, to the best of our knowledge neither Blackstone, John Fortescue, nor any other legal commentator asked ordinary citizens what they thought the maximal false acquittal rate ought to be. We did. It is difficult to know whether our data differs from those that would have been obtained had these earlier scholars polled their fellow countrymen. It may be that these erudite scholars might not have comprised a representative sample of seventeenth or eighteenth century citizens, who probably had more to fear from falsely acquitted criminals than these gentlemen did.

Third, we asked our respondents what were the maximal levels of false convictions and false acquittals a society should accept in order to function. We did not ask for a ratio. It may well be the case that asking for an explicit trade-off between the two types of erroneous verdicts, that is, a ratio, would suggest a different attitude toward the two types of errors than asking for absolute levels of each. This remains an empirical question, as do investigations of the different attitudes toward the two types of errors among different segments of the public and for different crimes.

**WHAT IS d'??**

What is the value of d' in jury trials? It is impossible to know precisely. But it is instructive to examine experimental research that has estimated d' in related tasks. Take, for example, the job of detecting liars. Fortunately, there are numerous studies on this topic, and most of them point to the same conclusion: Most people cannot reliably distinguish between liars and truth tellers (e.g., DePaulo, 1994; Zuckerman, DePaulo, & Rosenthal, 1981; Zuckerman, Koestner, & Alton, 1984). In one study, Kassin and Fong (1999) trained half of the participants to distinguish between persons who had or had not committed a mock crime. The untrained group exhibited a 55.6% accuracy rate, which was close to chance performance (50%). The trained group did even worse at 45.6%.

Even experienced police officers are not proficient at detecting liars (Vrij, 1994). Ekman and O'Sullivan (1991) and Ekman, O'Sullivan, and Frank (1999) found that only a small number of professionals can perform above chance. It seems reasonable to conclude that jurors would fare worse than trained professionals. Indeed, many of the studies using untrained laypersons show that d' is close to 0, indicating no ability whatsoever to distinguish the innocent from the guilty (DePaulo, Stone, & Lassiter, 1985).

What about polygraphs? Szucko and Kleinmuntz (1981) conducted a laboratory study in which 15 participants carried out a mock crime, and 15 did not. Polygraph
records for the 30 participants were examined by six interpreters who judged the likelihood that each individual, who denied his culpability, was lying. $d_s$ ranged from 0.6 to 0.9 across interpreters. In six other laboratory studies, $d_s$ were impressive and ranged from 1.5 to 2.3 (see Swets, 1996). Raskin and Honts (2000) summarized the results of the nine highest quality laboratory studies and the four highest quality field studies using the “comparison question” polygraph test methodology. The $d$’s were remarkable—2.68 and 2.72, respectively. However each of the 13 studies Raskin and Honts (2000) summarized included an “inconclusive” category into which a large number of the polygraphers’ decisions fell—up to 46% of the decisions in the case of one study. Raskin and Honts (2000) excluded these inconclusive decisions when calculating the percentage of correct and incorrect decisions. The elimination of such inconclusive decisions results in the inflation of $d'$. Hence, even with this very generous mode of scorekeeping, $d'$ levels did not achieve the 3.0 required by a false conviction rate of 5% and a false acquittal rate of 8%.

Evidence brought to bear in a jury trial may involve eyewitness testimony. Again, experimental results suggest that discrimination between old and new faces is often poor (Loftus, 1996). Shapiro and Penrod (1986) provided a meta-analysis of eyewitness facial identification with information on $d'$ for 120 studies. The single largest value of $d'$ was 0.8.

Medical tests provide a benchmark against which we can compare estimates of $d'$ in legal settings. In medicine, $d_s$ are typically higher, but they still do not exceed 3.0. CT scans (Computed Tomography scans) and RN scans (Radionuclide scans) are commonly used to detect brain lesions. $D_s$ for six radiologists ranged from 2.4 to 2.9 with CT scans and 1.4 to 1.7 with RN scans (from Swets, 1996).

$D_s$ have also been calculated for tests of cervical cancer. Such tests are performed by experts who distinguish normal cells from abnormal cells using photomicrographs. Using 10 observers and 6,000 slides, Bacus et al. (1984) found that $d'$ was 1.6. Computer-based evaluations of the same slides resulted in a $d'$ of 1.8 (Bacus, 1982, as cited in Swets, 1996, p. 45).

Mammograms are a test used by millions of women annually. The purpose of such tests is to detect breast cancer in its early stages. Even though the test is widely recommended by health organizations, the average $d'$ of the test is only 1.3 (Swets, Dawes, & Monahan, 2000). Likewise, men often rely on the PSA (Prostate Specific Antigen) test to detect prostate cancer in its early stages. Etzioni, Pepe, Longton, Hu, and Goodman (1999) show that the PSA test has a $d'$ of approximately 2.

Weather forecasting provides another benchmark for comparison. Again, $d'$s are higher than in legal domains, but they do not exceed 3.0. Mason (1982) found that $d'$ for forecasts of rain in Chicago was 1.5, $d'$ for forecasts of minimum temperature near Albuquerque, New Mexico, was 1.7, and $d'$ for forecasts of tornados in areas delineated by the Severe Storm Forecast Center in Kansas City was 1. Values of $d'$ for fog-risk forecasts at the airport in Canberra, Australia, issued 24, 18, and 12 hrs earlier were 0.8, 1.0, and 1.2, respectively.

A final benchmark is provided by discrimination required to accurately assign individuals to jobs, as in the military. The Armed Forces Qualification Test is used by the Navy for personnel selection. This test predicts the success of individuals in jobs such as quartermaster, signal-man, electrician’s mate, and mess management. With
success defined as performance at or above the 50th percentile and failure below, the
test yields disappointingly low \( d' \)’s that range from 0.6 to 0.8 (Swets, 1996, p. 50).

In sum, experimental estimates of discrimination in legal settings suggest that
\( d' \) is typically less than 1 and close to 0. An exception is polygraphs, with \( d' \)s ranging
from 1.5 to 2.3. In comparison, medical tests have values of \( d' \)’ ranging from 1.9 to 2.9,
and weather forecasters have values of \( d' \) ranging from 0.8 to 1.7. These benchmarks
show that, even with enormous technological sophistication, it is not easy to achieve
levels of \( d' \) that exceed 2. A \( d' \) of 3 needed to achieve a false conviction rate of 5%
and a false acquittal rate of 8% is extremely rare.

We acknowledge that tasks such as weather forecasting and cancer screening
differ in many important ways from the duty confronting a juror. However the former
tasks are devoid of interpersonal sparring, include minimal morality considerations,
and usually involve mere detection. Thus a case can be made that they are easier
than jury decision making, and therefore denote an optimistic upper limit for the
performance of jurors who must face a more daunting challenge.

**IS THERE SOME WAY AROUND THIS PROBLEM?**

A concerned reader might believe—indeed, hope—that there is something
wrong with our analysis. Jury decisions cannot be that bad. We discuss two potential
solutions and argue that neither solves the problem.

**Adjusting Prior Beliefs**

Perhaps with the “right” prior probability of guilt, a juror could achieve accept-
able error rates without stupendously diagnostic evidence (\( d' > 3 \)). If \( p(G) \), one’s
prior belief of guilt, was low, would the rate of incorrect convictions be small with
only modest evidence? Table 1 answers this question. Error rates are shown for differ-
ent levels of \( d' \) and different prior beliefs, \( p(G) \). For purposes of simplicity, we assume
that the utilities for correct and incorrect decisions are 1.0 and 0.0, respectively.

The first row presents misses and false alarms when \( d' \) is 0.5; evidence is weakly
diagnostic. A juror with a 33% prior belief in guilt would acquit guilty persons 87%
of the time. With this error rate, why conduct a trial at all? A juror with a 50% prior
belief of guilt has identical misses and false alarms that are unacceptably high at 40%.
Finally, a juror with a 67% prior belief of guilt convicts innocent defendants 87% of
the time. With a \( d' \) of 0.5, no prior beliefs produce acceptable error rates.

The second and third rows of the table show errors when \( d' \) is 1.0 and 2.0,
respectively. In both rows, we see the same problem, although to a lesser extent.
With a 33% prior belief of guilt, false alarms are low, but misses range from 26 to
58%. With a 67% prior probability of guilt, misses are under 13%, but false alarms
range from 26% to 58%—inexcusably high. Again, there is nowhere to place one’s
threshold that makes both error rates acceptable.

It is not until \( d' \) is at least 3.0 that errors become reasonable. When the prior
probability of guilt is 50%, mistakes of both types are 7%. When \( d' \) is 4.0, the mistakes
of both types are 2%. Unfortunately, these levels of \( d' \) are virtually never obtained.
We conclude that high or low prior probabilities of guilt do not alleviate the problem: in order to achieve acceptable levels of errors, unrealistic levels of $d'$ are required.

**Jury Decisions Are Not Made by Individuals**

Although it may be true that individual jurors might have disappointingly low levels of $d'$, perhaps groups are better. Do groups make more accurate decisions than individuals? Kerr, MacCoun, and Kramer (1996) compared the performance of individuals to that of groups in a variety of tasks. Some studies examined the extent to which individuals versus groups used extraevidentiary information, such as the attractiveness of the defendant (Bray et al., 1978; Izzett & Leginski, 1974; MacCoun, 1990). Kerr et al. (1996) reported that individuals usually performed better than groups.

Kerr et al. also reviewed studies of the “joinder bias” (Davis, Tindale, Nagao, Hinsz, & Robertson, 1984; Tanford & Penrod, 1984). Some offenses can be “joined” in a single trial. This action is based on the assumption that jury verdicts would be the same if offenses were tried together or separately. The “joinder” bias occurs when joint offenses lead to more convictions than separate offenses. Again, Kerr et al. (1996) found that joinder biases were less pronounced with individuals than groups.

Two more studies that appeared after Kerr et al.’s review are relevant. Sommer, Horowitz, and Bourgeois (2001) found that both individuals and groups violated the rules of determining negligence in order to make damage awards they felt were appropriate. Contrary to the prior findings, London and Nunez (2000) found evidence for the superiority of groups. Jurors who deliberated in groups were less sensitive to inadmissible evidence than jurors who rendered verdicts by themselves. In sum, the majority of the evidence does not support the conclusion that groups have larger $d$'s than individuals.

If groups are no better than individuals, why use groups? The Condorcet Jury Theorem states that “... under certain conditions a majority of a group, with limited information about a pair of alternatives, is more likely to choose the ‘better’ alternative than any one member of the group (Ladha, 1992, p. 617). This might lead one to expect that groups will manifest higher levels of $d'$ than individuals. However, one of the premises upon which the Condorcet Jury Theorem rests is that individuals vote independently. This assumption is obviously violated in most juries. Jurors are often heavily influenced by group deliberation and discussion. This process is most unlikely to satisfy the requirement for independence. Indeed, in his analysis of the Condorcet Jury Theorem, Ladha (1992) showed that the probability that the group will select the correct alternative is inversely related to the average correlation between pairs of group members.

We can examine the effects of juror dependence by making assumptions about $d'$ and $p(G)$ and examining error rates for groups with varying levels of influence. Let us begin with the unlikely case in which the jurors have no influence upon one another. If the defendant is guilty, jurors make 12 independent observations from the distribution of guilty persons in Fig. 1. Fig. 1 pertains to a sample of 1, but a sample of 12 independent observations would result in probability distributions with smaller—“tighter”—standard deviations.
Table 1. Percentages of Incorrect Acquittals (Misses) and Incorrect Convictions (FAs) Given the Average Strength of the Evidence \((d')\) and the Prior Probability of Guilt \([p(G)]\)

<table>
<thead>
<tr>
<th>Average strength of evidence</th>
<th>Prior probability of guilt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>33%</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Misses</td>
<td>87</td>
</tr>
<tr>
<td>FAs</td>
<td>5</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Misses</td>
<td>58</td>
</tr>
<tr>
<td>FAs</td>
<td>12</td>
</tr>
<tr>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Misses</td>
<td>26</td>
</tr>
<tr>
<td>FAs</td>
<td>9</td>
</tr>
</tbody>
</table>

Tables 1 and 2 represent the extremes of juror dependence. Table 1 represents complete dependence among jurors (equivalent to 1 juror) and Table 2 represents complete independence among jurors. With independence, errors rates less than 20% occur with a \(d'\) of 2. Nonetheless, errors are still above 10%. More realistic assumptions about juror independence would cause error rates to fall somewhere between these two extremes. Our point is simply that, even with the exceedingly unlikely assumption of complete independence of jurors, error rates still exceed 10%.

CONCLUSIONS

Only a small proportion of the events that could become legal cases ever turn into legal cases, and even fewer end up in jury trials. Most are dismissed or plea-bargained (Greenberg & Ruback, 1982). Chen (1991) examined a random sample of approximately 80,000 felony arrests made during 1977 and 1978 under the jurisdiction of the Los Angeles Superior Court. Only 4.2% of the sampled cases went to trial in this court. The majority of the others were either dismissed, because the evidence was insufficient, plea bargained, because the defendant realized the evidence was

Table 2. Percentages of Incorrect Acquittals (Misses) and Incorrect Convictions (FAs) Given the Average Strength of the Evidence \((d')\) and the Prior Probability of Guilt \([p(G)]\) Assuming the 12 Jurors are Completely Independent

<table>
<thead>
<tr>
<th>Average strength of evidence</th>
<th>Prior probability of guilt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>33%</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Misses</td>
<td>45</td>
</tr>
<tr>
<td>FAs</td>
<td>36</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Misses</td>
<td>33</td>
</tr>
<tr>
<td>FAs</td>
<td>29</td>
</tr>
<tr>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Misses</td>
<td>17</td>
</tr>
<tr>
<td>FAs</td>
<td>15</td>
</tr>
</tbody>
</table>
quite sufficient, or referred to a lower court after the offense was downgraded to a misdemeanor. Those few felony cases that made it to juries were probably the most difficult ones, because the evidence was neither flimsy enough to warrant dismissal or abundant enough to compel a guilty plea.

It is ironic that the recent use of forensic science evidence with its relatively high levels of diagnosticity may not be the complete solution to the problem we have discussed. Cases in which there is highly diagnostic forensic evidence are likely to result in a guilty plea and thus will never be heard by a jury. Such evidence may reduce the total number of cases heard by juries but may not reduce the difficulty of those cases which go to a jury, because such cases will likely be devoid of the highly diagnostic forensic evidence. Similarly, if a large number of relatively independent pieces of evidence are presented at trial, even if each has a modest \( d' \), their accumulated diagnosticity may be much higher than the unimpressive levels we have documented for individual modes of evidence, such as eyewitness testimony, polygraph testing, etc. Indeed, if a cascade of such independent tests all point toward guilt, then the aggregation will have a \( d' \) that will never be appreciated by any jury, because the defendant will very likely plead guilty to the original or to a lesser charge.

Despite the relatively small number of jury trials, the impact of jury decisions is felt far and wide. Assumptions about what juries would do if faced with the decision are used to justify a host of actions, including insurance settlements, divorce settlements, and industrial fines. Juries would presumably make accurate decisions if confronted with these easier cases, but are they accurate with the difficult cases they get? Our analysis suggests that error rates are sizeable. Even though jury decisions are highly regarded methods for determining justice, juries make mistakes—far more than most of us think.

Technological innovations have exposed some juries’ errors; DNA testing, for example, has revealed mistakes that have sent defendants to death row. Technological advances can increase \( d' \), and such an increase is the only way to simultaneously reduce false convictions and false acquittals. Scientific advances may eventually improve the accuracy of jury decision making. In the meantime, we should recognize that juries make sizeable errors, even when they are doing their level best.

**APPENDIX**

Below is the exact phrasing of the questions we posed to undergraduates.

Consider a serious crime, such as murder or rape. Any legal system that convicts people for such crimes will occasionally make mistakes for the simple reason that the evidence is usually ambiguous. If we make it easy to convict people, we will convict more guilty people, but we are also more likely to convict some innocent people, too. If we make it easy to acquit people, we will let more innocent people go free, but we also let more guilty people go free, too.

It would be great if we could live in a society that *always* acquits the innocent. But that can’t happen. If we did that, we would have to acquit everyone charged, and guilty people would go free. As long as there are convictions, some innocent people will be convicted. We would like the percentage of incorrect conviction rates to be
zero, but that is impossible. What do you think is the largest percentage of incorrect convictions (convicting the innocent) that a law-abiding society should accept in order to function?

It would also be great if we could live in a society that always convicts the guilty. But that can’t happen either. If we did that, we would have to convict everyone, and innocent people would be convicted. As long as there are acquittals, some percentage of guilty people will be acquitted, too. We would like the percentage of incorrect acquittals to be zero, but that is impossible. What do you think is the largest percentage of incorrect acquittals (letting the guilty go unpunished) that a law-abiding society should accept in order to function?

ACKNOWLEDGMENT

We thank Robert MacCoun, Greg Mitchell, Robert Sorkin, and Trisha Van Zandt for thoughtful comments on earlier drafts.

REFERENCES


