Institute for Policy Research
Northwestern University
Working Paper Series

WP-06-05

Estimating the Accuracy of Jury Verdicts

Bruce D. Spencer
Faculty Fellow, Institute for Policy Research
Professor of Statistics
Northwestern University

Version date: April 17, 2006;
rev. May 4, 2007
Forthcoming in Journal of Empirical Legal Studies

Abstract

Average accuracy of jury verdicts for a set of cases can be studied empirically and systematically even when the correct verdict cannot be known. The key is to obtain a second rating of the verdict, for example the judge's, as in the recent study of criminal cases in the U.S. by the National Center for State Courts (NCSC). That study, like the famous Kalven-Zeisel study, showed only modest judge-jury agreement. Simple estimates of jury accuracy can be developed from the judge-jury agreement rate; the judge's verdict is *not* taken as the gold standard. Although the estimates of accuracy are subject to error, under plausible conditions they tend to overestimate the average accuracy of jury verdicts. The jury verdict was estimated to be accurate in no more than 87% of the NCSC cases (which, however, should *not* be regarded as a representative sample with respect to jury accuracy). More refined estimates, including false conviction and false acquittal rates, are developed with models using stronger assumptions. For example, the conditional probability that the jury incorrectly convicts given that the defendant truly was not guilty (a "type I error") was estimated at 0.25, with an estimated standard error (s.e.) of 0.07, the conditional probability that a jury incorrectly acquits given that the defendant truly was guilty ("type II error") was estimated at 0.14 (s.e. 0.03), and the difference was estimated at 0.12 (s.e. 0.08). The estimated *number* of defendants in the NCSC cases who truly are not guilty but are convicted does seem to be smaller than the *number* who truly are guilty but are acquitted. The conditional probability of a wrongful conviction, given that the defendant was convicted, is estimated at 0.10 (s.e. 0.03).

I. INTRODUCTION

Although it has long been known that juries sometimes make incorrect decisions (Borchard 1932), researchers have devoted relatively little attention to quantifying the error rate for jury verdicts. The reasons for jury error are many, and can include misleading or incomplete evidence or testimony and failures in reasoning relating to complexity of evidence or the law.[1] Use of DNA analysis after the trial has shown that incorrect decisions were made by juries even in cases where the death penalty was assigned. Yet, such direct assessments of accuracy are not possible on a wide scale because only atypically is the correct verdict known, and it is difficult to generalize from those cases to the more typical cases where the correct verdict is not knowable. Fortunately, statistical methods can be used to estimate the average accuracy of jury verdicts if a second assessment of the cases is available. In the field of sample surveys, it is commonplace to estimate sampling accuracy even though the true value is not observed – the key is to use replication. A similar approach can be taken to estimate the accuracy of jury verdicts when "replications" such as a judge's decision on the cases are available. Although judges' verdicts are not routinely known, they have been recorded and studied in the famous work by Kalven and Zeisel (1966), *The American Jury,* and much more recently in the study of hung juries by the National Center for State Courts (NCSC), described in Hannaford-Agor et al. (2003).

Jury accuracy refers to the average probability for a set of cases that the jury verdict is, in some sense, correct. "Correct" can be interpreted in a variety of ways. A "procedural viewpoint" considers a "correct" decision to be one which applies the legal standards correctly: if proof is not demonstrated to the standards prescribed by the law, the defendant should be acquitted.[2] Thus, if a person tried for a crime truly committed the crime, but the evidence was lacking, the procedurally correct decision would be acquittal. This interpretation was adopted by S. D. Poisson (1837) in his early analysis of jury error; reviewing his work, Gelfand and Solomon (1973, 272) characterize the jury's responsibility

[1] In *Speiser v. Randall,* 357 U.S. 513, 525, the Supreme Court observed: "There is always in litigation a margin of error, representing error in factfinding . . ." The incompetence and/or inattention of judges and juries (drunk, asleep, poorly educated, bored) or the bias (racial prejudice, religious beliefs, social class bias, political views) are obviously also factors.

[2] 'The meaning of "correct" application of legal standards is not perfectly obvious. Not only is the law rife with unsettled legal issues, but even bedrock, established principles such as the "beyond a reasonable doubt" standard are subjective in application.

as deciding whether the defendant is "convictable" rather than guilty. An alternative, "omniscient viewpoint", holds that the correct decision is the one that would be reached by an impartial and rational observer with perfect information (including complete and correct evidence) and complete understanding of the law. If the defendant committed the crime, the correct decision is guilt, regardless of the strength of evidence. Although this definition of correct decision could be ambiguous in certain settings involving legal complexities, where even experts would disagree, in many applications it could be precise enough. This perspective has the advantage of being defined independently of the evidence and courtroom presentations, and conforms to popular notions of justice.

The rate of agreement between jury's verdict and judge's verdict provides an important *indicator* of jury accuracy. As discussed in Section II, the agreement rates for criminal cases excluding hung jury cases are similar for the Kalven-Zeisel and NCSC studies, at just under 80%. Such an agreement rate is not cause for complacency, given that agreement by chance alone would exceed 60% for each study.

In this paper we show that an estimate of jury accuracy can be derived from the observed agreement rate even when the correct verdict is unknown. In Section III, an estimator of jury accuracy is developed that has three components of error, survey error from estimating the agreement rate, specification error arising because differential accuracy between judge and jury is not observed and the dependence between judge and jury verdicts is not known, and identification error arising because we cannot distinguish correct agreement from incorrect agreement. The specification error will be one sided, leading to overestimates of jury accuracy, provided that two conditions hold: (i) errors in the judge's and jury's verdicts for a case are either statistically independent or positively dependent, and (ii) the judges' verdicts are no less accurate on average than the juries', even though for individual cases the judge's verdict may be incorrect when the jury's verdict is correct. The identification error is similarly one-sided, always. From the observed agreement rates, the probability of a correct verdict by the jury is estimated at 87% for the NCSC cases and 89% for the Kalven-Zeisel cases. Those accuracy rates correspond to error rates of 1 in 8 and 1 in 9, respectively. The accuracy rates apply to both the "procedural" and the "omniscient" interpretations of correct verdict noted earlier. Caveat: the NCSC cases were not chosen with equal probabilities as a random sample, and the estimates of accuracy should not be generalized to the full caseload in the four

jurisdictions let alone to other jurisdictions. The "optimistic" property of the estimator of accuracy holds under both interpretations of accuracy (process or outcome) so long as the conditions on dependence and differential accuracy of judge and jury apply.

Two types of errors can be made in a criminal trial: an innocent person can be convicted or a guilty person can be found not guilty. In the language of statistical hypothesis testing, with innocence as the "null hypothesis", the former is a type I error and the latter is a type II error. That is, the jury commits a type I error if it convicts a person who truly is not guilty, and it commits a type II error if it acquits a person who truly is guilty. The U.S. Constitution in effect requires that type I error rates be kept low, and type I errors are viewed as far worse than type II errors.[3] Estimates of type I and type II error rates are developed with statistical models that explicitly allow for latent (unobserved) states for the correct verdict (guilty, not guilty). Such an analysis was carried out by Gastwirth and Sinclair (1998, 2004) for burglary and auto-theft cases in Kalven-Zeisel study, using the Hui-Walter (1980) methodology and invoking the assumption that the probabilities of type I and type II errors did not vary by type of crime. In estimating probabilities of type I and type II errors in the NCSC cases, we exploit the availability of assessments of the strength of evidence (Section IV) in the NCSC cases and avoid the assumption of constancy of error rates by type of crime. The analysis suggests, subject to limits of sample size and possible modeling error, that juries in the NCSC cases may have *higher* type I error rates than type II error rates but, since the correct verdict is more often "guilty" than "not guilty", the juries

[3] The leading case on the standard of proof in criminal cases, *In re Winship,* 397 U.S. 358 (1970), holding that the "reasonable doubt" standard is constitutionally-required, makes some observations on probabilities and the costs of error. The majority opinion by Justice Brennan, referring to reasonable doubt, says:

> It is a prime instrument for reducing the risk of convictions resting on factual error. The standard provides concrete substance for the presumption of innocence-that bedrock 'axiomatic and elementary' principle whose 'enforcement lies at the foundation of the administration of our criminal law.' . . .

Justice Harlan's concurring opinion observes:

> In a civil suit between two private parties for money damages . . . we view it as no more serious in general for there to be an erroneous verdict in the defendant's favor than for there to be an erroneous verdict in the plaintiff's favor . . . In a criminal case, on the other hand, we do not view the social disutility of convicting an innocent man as equivalent to the disutility of acquitting someone who is guilty. . . In this context, I view the requirement of proof beyond a reasonable doubt in a criminal case as bottomed on a fundamental value determination of our society that it is far worse to convict an innocent man than to let a guilty man go free.

commit fewer type I errors than type II errors. Analysis also suggests that the judge is more accurate overall by a small amount, but the juries have smaller type I error rates than the judges. (Section V)

The importance of the accuracy of jury verdicts has long been known, but the possibility of estimating their accuracy on a large scale does not appear to have been widely appreciated. More empirical studies could be conducted to estimate and compare accuracy rates over time, across jurisdictions, for different kinds of cases, for consistency between demographic characteristics of defendant and jury, etc. Recommendations for future work are presented (Section VI).

II. MEASURING JUDGE-JURY AGREEMENT

Empirical studies of rates of agreement on verdict between judge and jury include the classic *The American Jury* by Kalven and Zeisel (1966), covering more than 3500 trials in 47 states and Washington D.C. in 1954-1955 and 1958 and a recent study by Eisenberg et al. (2005) comparing the Kalven-Zeisel data with the NCSC data. The NCSC data came from a convenience sample of four metropolitan areas chosen for various reasons: Los Angeles and Washington D.C. chosen because they had high rates of hung juries, Maricopa County (which includes Phoenix) in Arizona because it used an innovative procedure to allow judges to avoid hung juries, and the Bronx because it had a high volume of cases. The NCSC study produced judge and jury data on 290 non-hung-jury trials for noncapital criminal cases in 2000-2001, selected with unequal rates from the four jurisdictions. The unequal sampling rates imply that the results for the NCSC sample cases should be weighted if they are to generalize to the full caseload in the four jurisdictions. No such weighting is employed in the present analysis, and the statistical inferences do not extend outside the cases in the NCSC study.[4] The studies are important in that they surveyed the judges concerning their beliefs about the cases. Judges were asked, among other things, “If you had decided this case in a bench trial, would you have

[4] Sampling weights were not available. Sampling weights can be developed if sufficient information about the sampling rates and participation rates is available, but such information needs to be developed as the sample design is implemented and the survey is fielded. Future judge-jury agreement studies should be designed to provide sampling weights.

rendered a verdict for the prosecution or for the defense?" For detailed discussion of the data see Eisenberg et al. (2005) and Hannaford-Agor et al. (2003).[5]

The overall rate of agreement between judge and jury, say $\hat{\rho}$, is 0.80 for criminal cases studied by Kalven and Zeisel and is 0.77 for the NCSC cases. Those agreement rates are quite modest compared to what one would get by chance. The rate of agreement one would get by chance is 0.62 (the same value for each dataset), given the observed proportions of cases classified by the judge as guilty or not guilty and the corresponding proportions by jury.[6] A chance-corrected measure of agreement is Cohen's "kappa" statistic, κ, which ranges from -1 (complete disagreement) to 0 (agreement at chance level) to +1 (perfect agreement). The values of κ for the Kalven-Zeisel data and NCSC data are 0.47 and 0.38, which Fleiss (1981, 218) interprets as showing only fair to poor agreement beyond chance.

[5] The NCSC study asked questions about a variety of counts, and the overall classification of the outcome as "guilty" or "not guilty" involved a number of steps. The classifications used here are based on the coding for the Eisenberg et al (2005) analysis. The instructions were kindly provided by Professor T. Eisenberg.

[6] To calculate the rate of chance agreement, say $\hat{\rho}_{\text{chance}}$, treat the guilty and not guilty rates for judge and for jury as fixed and consider classifications by judge and jury to be independent. E.g., the Kalven-Zeisel data in Table 1 give $\hat{\rho}_{\text{chance}} = 0.321 \times 0.165 + 0.679 \times 0.835 = 0.62$.

Table 1: Distributions of Judge-Jury Agreement, Excluding Hung Jury Cases

	Jury: Not Guilty	*Jury: Guilty*
A. Kalven and Zeisel Data		
Judge: Not Guilty	14.2%	2.3%
Judge: Guilty	17.9%	65.6%
B. NCSC Data		
Judge: Not Guilty	12.8%	5.5%
Judge: Guilty	17.6%	64.1%

Note: Panel A derives from the Kalven and Zeisel data, excluding hung jury cases (5.5% of cases). Numbers computed from Eisenberg et al. (2005, Table 2). Panel B is based on the NCSC data on 290 trials without hung juries. Proportions were computed from counts in Eisenberg et al. (2005, Table 3).

III. ESTIMATING ACCURACY FROM AGREEMENT

A. The Relationship between Agreement and Overall Accuracy for the Jury

The choices of verdicts by judge and jury can be represented by a simple yet comprehensive model. First we define notation. Consider a study having N cases with a different jury for each case. For case i the judge's probability of choosing a correct verdict is p_i^A, the jury's probability of choosing the correct verdict is p_i^B. Depending on one's perspective, the values of p_i^A and p_i^B could all be 0 or 1, but they do not need to be. The probability that the judge and jury choose the same verdict (correctly or not) is ρ_i.

Kruskal (1988) cautioned against casually assuming independence holds. If the judge and the jury for a case were to choose their verdicts independently, the probability of agreement would be $p_i^A p_i^B + \left(1-p_i^A\right)\left(1-p_i^B\right)$. The choices could be dependent, however, because both judge and jury are presented with the same courtroom evidence and to

some extent are subject to common community pressures and biases. The difference between the actual probability of agreement and the probability that would hold if the choices were independent is denoted by ω_i, that is, $\omega_i = \rho_i - p_i^A p_i^B - \left(1 - p_i^A\right)\left(1 - p_i^B\right)$.[7] Denote the averages for the N cases in the study by $p^A = \sum_i p_i^A / N$, $p^B = \sum_i p_i^B / N$, $\rho = \sum_i \rho_i / N$, $\omega = \sum_i \omega_i / N$, and denote the differential accuracy between judge and jury by $\delta = p^A - p^B$. If case i is unusually difficult (or easy), one might find that both p_i^A and p_i^B are lower (or higher) than average. To clarify this concept, define the covariance between judge's and jury's probability of being correct as $\sigma_{AB} = \sum_i \left(p_i^A - p^A\right)\left(p_i^B - p^B\right) / N$. In contrast to ω, which reflects within-case dependencies for a verdict conditional on the probabilities of being correct, σ_{AB} reflects linear dependence of the judges' and juries' probabilities of being correct. Define the parameter γ to reflect the more general dependence, $\gamma = \omega + 2\sigma_{AB}$.

The relation between the agreement rate and jury accuracy is shown explicitly by the mathematical identity

$$\rho = 2(p^B)^2 + 2(\delta - 1)p^B + 1 + \gamma - \delta. \qquad (1)$$

This equation allows us to interpret γ as the excess agreement that is expected when independence is not present. Equation (1) must be satisfied by any set of p^B, ρ, γ, and δ that occur in practice. Not all values can occur; it is necessary that agreement be bounded below, $\rho \geq \gamma + (1 - \delta^2)/2$.

We can solve (1) for the jury accuracy rate p^B but in doing so we confront an "identification problem" (Manski 1995). The problem arises because agreement occurs when the judge and jury are both correct or are both incorrect, and unless we use outside knowledge we cannot know whether agreement occurs mostly because judge and jury

[7] The parameter ω_i is Lehmann's (1966) measure of positive quadrant dependence for a two-by-two table that shows correct or incorrect decision by judge and jury. If the probabilities p_i^A, p_i^B, and ρ_i take only the values 0 or 1, then $\omega_i = 0$.

both tend to be correct or both tend to be incorrect. That is, on the basis of agreement alone we cannot distinguish jury accuracy, p^B, from inaccuracy, $1-p^B$.

We can represent the identification error mathematically. Equation (1) is quadratic in p^B, with roots, $p^B_{low} = .5\left(1-\delta-\sqrt{2(\rho-\gamma)-1+\delta^2}\right)$ and $p^B_{high} = .5\left(1-\delta+\sqrt{2(\rho-\gamma)-1+\delta^2}\right)$. At least one of the roots must be between 0 and 1 because p^B must equal one of the roots and $0 \le p^B \le 1$. The root p^B_{high} is the same distance above $.5(1-\delta)$ as the root p^B_{low} is below. To resolve the dilemma of which root equals p^B, we will choose the larger root constrained to not exceed 1, namely $\min\left\{p^B_{high}, 1\right\}$, knowing that jury accuracy may be overstated. The overstatement is equal to the *identification error*, defined here as $\min\left\{p^B_{high}, 1\right\} - p^B$. The identification error is non-negative. If p^B is actually equal to p^B_{low}, then p^B does not exceed either p^B_{high} or 1 and so the identification error is positive. The only other alternative is that p^B is equal to the larger root, and in that case the identification error is 0.

B. An Estimator of Jury Accuracy

To obtain an estimator of jury accuracy, we substitute the observed agreement rate $\hat{\rho}$ for ρ in the formula for p^B_{high} and constrain the result so that it conforms to constraints on p^B. (Thus, the estimator may not exceed 1 nor may it fall below the value that would result ρ were equal to its lower bound of $\gamma + (1-\delta^2)/2.$) The estimator may be written as an explicit function of $\hat{\rho}$, γ, and δ,

$$f(\hat{\rho},\gamma,\delta) = \begin{cases} 1 & \text{if } 1+\delta < \hat{\rho}-\gamma \\ 0.5(1-\delta+\sqrt{2(\hat{\rho}-\gamma)-1+\delta^2}) & \text{if } (1-\delta^2)/2 \le \hat{\rho}-\gamma \le 1+\delta \\ 0.5(1-\delta) & \text{if } \hat{\rho}-\gamma < (1-\delta^2)/2. \end{cases} \qquad (2)$$

If we do not know the values of δ and γ, we set them to 0 in (2). Our estimator of p^B for the NCSC data and Kalven-Zeisel data is thus $\hat{p}^B = f(\hat{\rho},0,0)$.

The estimator $\hat{p}^B$ is subject to three kinds of error, identification error, survey error and specification error. Identification error can be shown to equal $f(\rho,\gamma,\delta) - p^B$. *Survey error* includes sampling error, nonresponse error, response error, data processing error, and all other errors that may cause $\hat{\rho}$ to differ from ρ. The survey error is equal to $f(\hat{\rho},\gamma,\delta) - f(\rho,\gamma,\delta)$. The *specification error* is equal to $f(\hat{\rho},0,0) - f(\hat{\rho},\gamma,\delta)$ and reflects the incorrect choices of δ and γ. The total error in the estimate is $f(\hat{\rho},0,0) - p^B$, which equals the sum of identification error, survey error, and specification error.

The following considerations make it plausible that specification error is non-negative and, as a result, that the estimator $\hat{p}^B$ tends to overstate accuracy. The consequences of using the wrong value of γ are easily seen, because mathematically $f(\hat{\rho},\gamma,\delta)$ decreases or remains the same as γ increases. For substantive insight, consider the two components of $\gamma = \omega + 2\sigma_{AB} \geq 0$. The common environment and overlap in exposure to evidence and questioning make it plausible that $\omega \geq 0$. Although ω_i could be negative for cases where juries perceive the judges' preferences for a particular verdict and react in the opposite direction, that is probably the exception rather than the rule. It is also plausible that, on average, cases that are more difficult for the judge are also more difficult for the jury, and thus $\sigma_{AB} \geq 0$. Together, these casual observations suggest that $\gamma \geq 0$ for both the "procedural" and the "omniscient" interpretations of correct verdict (Section I). The consequences of using the wrong value of δ are also easily seen, because $f(\hat{\rho},\gamma,\delta)$ decreases or remains the same as δ increases. It is plausible that even though some judges' verdicts less accurate than some juries', the judges are more accurate on average, both for statistical and non-statistical reasons. Statistical analyses discussed in Section V.B yielded estimates of δ of 0.02 and 0.05 for the NCSC cases, with standard errors respectively estimated at 0.06 and 0.05. Similarly, Gastwirth and Sinclair (1998, 63) estimated δ to be 0.17 for burglary and 0.15 for auto-theft cases in the Kalven-Zeisel study. Judges may have more information about the defendant and see more evidence than juries, further supporting the plausibility of $\delta \geq 0$ for the "omniscient" interpretation of correct verdict. Judges' greater experience and knowledge of the law also support the plausibility of $\delta \geq 0$ for the "procedural" interpretation of correct verdict. Thus, it is

plausible that both $\gamma \geq 0$ and $\delta \geq 0$ and hence specification error is non-negative.[8] We also know that identification error is non-negative. It is plausible, then, that together the identification error and specification error contribute an upward bias (if any) in $\hat{p}^B$ and, if the observed agreement rate $\hat{\rho}$ has expected value approximately equal to the actual rate ρ, the estimator $\hat{p}^B$ will if anything tend to overstate accuracy (under either the "procedural" or "omniscient" interpretation of correct verdict).

C. Empirical Estimates of Jury Accuracy

Estimates of jury accuracy based on $\hat{p}^B$ are 0.87 for the NCSC cases and 0.89 for the Kalven-Zeisel cases. These are optimistic estimates if, as discussed above, the judge is at least as accurate as the jury $(\delta \geq 0)$ and the dependence is non-negative $(\gamma \geq 0)$. Table 2 shows those estimates along with estimates based on alternative choices of δ and γ in (2). For example, if $\delta = 0.10$ and $\gamma = 0$, the estimates of accuracy drop by about 0.05 for both the NCSC and the Kalven-Zeisel cases, to 0.82 and 0.84, respectively. The probability of correct decision by judge is obtained by adding δ to probability for jury.

[8] If $\delta < 0$ but $\gamma \geq 0$ for a set of cases, $\hat{p}^B$ provides an optimistic estimate of the *judges'* accuracy.

Table 2: Alternative Estimates of the Probability of a Correct Jury Decision, Given that the Jury Reached a Decision.

δ $(\gamma = 0)$	γ $(\delta = 0)$	$\hat{\rho} = 0.769$ (NCSC)	$\hat{\rho} = 0.798$ (Kalven-Zeisel)
-0.25	-0.250	1.00	1.00
-0.20	-0.200	0.98	1.00
-0.10	-0.090	0.92	0.94
0.00	0.000	0.87	0.89
0.10	0.065	0.82	0.84
0.20	0.115	0.78	0.80
0.25	0.137	0.76	0.78

Note: Estimates of probabilities are based on (2).

It is remarkable and perhaps important to note that the two datasets lead to such similar $\hat{p}^B$ estimates despite the wide separation in time, the growth in the use of plea bargaining to reduce caseload, different demographic mixes of defendants, changes in available prison space. Whether this similarity indicates a coincidence or a stability or robustness of the judicial system is unclear, however, as the geographic areas differed between the two studies and the sample selection processes differed; see Eisenberg et al. (2005, 173) for further discussion.

D. Considerations of Sampling Error and Generalizability

It is not easy to generalize empirical findings of accuracy for the NCSC and Kalven-Zeisel studies to other sets of cases, past or current. The sample sizes in the NCSC study were not proportional to caseloads in the jurisdictions and analyses from the 290 cases should not be generalized to all cases in the four areas unless either weighting or statistical controlling for area is used. In the current analysis, we do not seek to generalize to the full caseload for the four jurisdictions in the NCSC study. Furthermore, the four jurisdictions were not chosen to be representative of all jurisdictions, and generalization beyond the four jurisdictions in the study is not justified. In order to generalize from the NCSC study,

one would need additional studies showing similar results, so that one could feel some confidence that the accuracy rates were stable.

Although the data for the studies cannot be viewed as coming from a random sample of n cases, we can still get some insight into sampling variability by calculating the standard error as if simple random sampling had been used. A convenient method for estimating the standard error for a statistic calculated from a simple random sample is the jackknife method, which omits one case at a time, recalculates the statistic of interest, and estimates the standard error as the square root of $1-1/n$ times the sum of squared deviations of the resulting values about their average (Efron and Tibshirani 1993, 136). The estimated standard error for $\hat{p}^B$ is less than 0.01 for the Kalven-Zeisel data and less than 0.02 for the NCSC data. These standard errors may be viewed as providing optimistic bounds on the variation one might get from replicating the studies under similar conditions. If one had random samples of judges, one would calculate the jackknife estimate of standard error by omitting the data from one judge at a time; such estimates of standard error tend to be larger than those computed under the assumption of simple random sampling of cases. In any event, the standard errors are only suggestive for the estimates from the Kalven-Zeisel and NCSC studies, since random sampling was not used to develop the studies. Nor do the standard errors reflect the effect of nonresponse or other nonsampling errors. Rather, the standard errors can be used as measures of the sensitivity of the estimates to changes in the data.

IV. STRENGTH OF EVIDENCE

A decision by the judge or jury to acquit could indicate belief that the defendant did not commit the crime or, alternatively, it could indicate that proof was not demonstrated beyond a reasonable doubt. A possible way to distinguish between the two is to classify cases according to the strength of evidence. Question 12 of Sample II (Kalven and Zeisel 1966, 532) asked the judge.

From the factual evidence in the case was the defendant's guilt or innocence
1 ☐ very clear?
2 ☐ a close question whether or not he was guilty beyond a reasonable doubt?

Looking at a subset of Kalven-Zeisel data, Gastwirth and Sinclair (2004, 171-172) found that judge and jury agreed in 614 of 675 cases where the judge rated the evidence "clear" but agreed in only 307 of 516 cases where it was "close", so the odds of agreement were almost 7 times greater for the clear cases than the close cases.

The NCSC study asked jurors and judges to mark an answer on a 7 point scale to the following question.

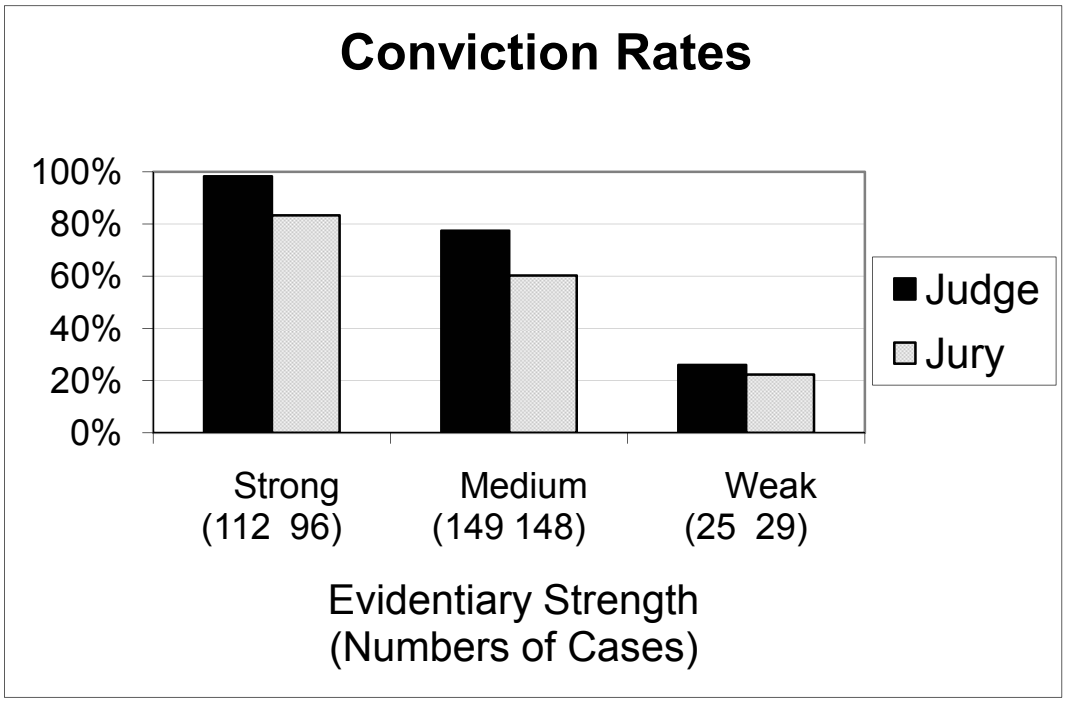
All things considered, how close was this trial?

	1 2 3 4 5 6 7	
Evidence strongly favored prosecution	O O O O O O O	Evidence strongly favored defense

We follow Eisenberg et al (2005, 186) in interpreting the response as an assessment of the net strength of evidence for conviction, where a response of 1-2 indicates strong, 3-5 indicates medium, and 6-7 indicates weak evidence for conviction. The interpretation is reasonable but not perfect – a respondent who perceived the both sides' evidence as weak might be inclined to mark a middle box (3 or 4), which we would misinterpret as medium evidence. In this sense, the Kalven-Zeisel question seems clearer. Further, the strength of evidence is reported for the trial as a whole and not necessarily for the count relating to the verdict being analyzed. Jurors on the same case varied in their ratings. To obtain a single jury rating of evidence for a case, we used the averages of the jurors 7-point scale ratings as calculated by Eisenberg et al (2005). The judge's and jury's assessments of the evidence as strong, medium, or weak matched only 56% of the time, and agreement was poor as indicated by $\kappa = 0.22$ (Underlying data are in Eisenberg et al (2005, Table 3) and in Table 3, below). It is plausible that the evaluations of evidence are comparative and that judges and juries compare to different standards. It can be argued that evidence of guilt in most criminal cases is strong because prosecutors want to keep their conviction rates high and will avoid going to trial with weak evidence. Judges see several (often many) criminal cases. Therefore, a case rated weak may nonetheless have evidence of guilt that is legally sufficient. Figure 1 shows that judges convicted in 98% of the 112 cases they rated as having strong evidence for conviction, they convicted in more than three quarters (78%) of the 149 cases they rated as having medium evidentiary strength, and they convicted in one quarter (26%) of the 25 cases where they rated the evidence for conviction as weak. The basis for jurors' comparative judgments about

evidence is less clear (television?), and juries convicted in 88% of the 96 cases they rated as having strong evidence for conviction, they convicted in 60% of the 148 cases they rated as having medium evidentiary strength, and they convicted in 22% of the 29 cases where they rated the evidence for conviction as weak. The data on strength of evidence is critically important for fitting statistical models with latent classes, as discussed in the next section.

Figure 1. Conviction Rates in NCSC Cases by Judges' and Juries' Assessments of Strength of Evidence for Conviction.



V. PROBABILITY MODELS WITH LATENT CLASSES

A. Log-linear Models

The correct or *true* (but unobserved) state of a case will be denoted by U, the decision or classification by the judge will be denoted by A, and that by the jury by B. Each of U, A, and B can take the values $0,1$, corresponding to "not guilty" and "guilty", respectively. The judge's assessment of evidentiary strength is denoted by C and the jury's by D; each of C and D can take values 1 (weak), 2 (medium), and 3 (strong). Table 3 provides the empirical frequency distribution of cases by attribute patterns including evidentiary strength.

Table 3: Distribution of NCSC Cases by Observed Attribute Patterns

Count	*A*	*B*	*C*	*D*	*Count*	*A*	*B*	*C*	*D*
7	0	0	1	1	0	1	0	1	1
6	0	0	1	2	0	1	0	1	2
0	0	0	1	3	0	1	0	1	3
8	0	0	2	1	6	1	0	2	1
15	0	0	2	2	26	1	0	2	2
0	0	0	2	3	1	1	0	2	3
0	0	0	3	1	2	1	0	3	1
0	0	0	3	2	11	1	0	3	2
0	0	0	3	3	1	1	0	3	3
1	0	1	1	1	0	1	1	1	1
2	0	1	1	2	3	1	1	1	2
2	0	1	1	3	2	1	1	1	3
1	0	1	2	1	1	1	1	2	1
7	0	1	2	2	41	1	1	2	2
1	0	1	2	3	36	1	1	2	3
0	0	1	3	1	2	1	1	3	1
1	0	1	3	2	36	1	1	3	2
0	0	1	3	3	52	1	1	3	3

A is judge's classification as guilty (1) or not guilty (0) and *B* is jury's. *C* is judge's assessment of strength of evidence (1 = weak, 2 = medium, 3 = strong) and *D* is jury's assessment.

We will consider various probabilities. Generally, the notation p_x^X will denote the probability that $X = x$, p_{xy}^{XY} will denote the probability that both $X = x$ and $Y = y$, and

$p_{x\,\bullet y}^{X\bullet Y}$ will denote the conditional probability that $X = x$ given that $Y = y$. Thus, the probability that $U = u$ is denoted by p_u^U, for $U = 0, 1$, and we have $1 = p_0^U + p_1^U$. Note that p_u^U depends on the mix of cases. The conditional probability that $A = a$ given $U = u$ is denoted by $p_{a\,\bullet u}^{A\bullet U}$, the conditional probability that $B = b$ given $U = u$ is denoted by $p_{b\,\bullet u}^{B\bullet U}$, and the conditional probability that $A = a$ and $B = b$ given that $U = u$ is denoted by $p_{a\,b\bullet u}^{AB\bullet U}$. The probability that the judge's decision is correct for a case with correct status u is $p_{u\,\bullet u}^{A\bullet U}$, and the corresponding probability that the jury's decision is correct is $p_{u\,\bullet u}^{B\bullet U}$. Thus, the probabilities of type I errors are $p_{1\,\bullet 0}^{A\bullet U}$ and $p_{1\,\bullet 0}^{B\bullet U}$, and the probabilities of type II errors are $p_{0\,\bullet 1}^{A\bullet U}$ and $p_{0\,\bullet 1}^{B\bullet U}$. These probabilities are related to the overall probabilities of correct decision defined in Section III.A, $p^A = p_{0\,\bullet 0}^{A\bullet U} p_0^U + p_{1\,\bullet 1}^{A\bullet U} p_1^U$ and $p^B = p_{0\,\bullet 0}^{B\bullet U} p_0^U + p_{1\,\bullet 1}^{B\bullet U} p_1^U$.

Table 3 showed attribute patterns involving the four observable attributes, A, B, C, D; we want to consider the unobserved attribute U as well. The probability that case i has the set of attributes $(U_i, A_i, B_i, C_i, D_i)$ equal to (u, a, b, c, d) is denoted by $p_{u\,a\,b\,c\,d\,i}^{U\,A\,B\,C\,D}$. There are 72 possible patterns of attributes, allowing for true state U. We approximate the probability that a case has one of the 72 possible patterns of attributes by a hierarchical log-linear model of the form

$$\begin{aligned}\log p_{u\,a\,b\,c\,d\,i}^{U\,A\,B\,C\,D} = {} & \alpha + \lambda_u^U + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_d^D + \lambda_{ua}^{UA} + \lambda_{ub}^{UB} \\ & + \lambda_{uc}^{UC} + \lambda_{ud}^{UD} + \lambda_{ac}^{AC} + \lambda_{bd}^{BD} + \lambda_{cd}^{CD} + \lambda_{uac}^{UAC} + \lambda_{ubd}^{UBD}.\end{aligned} \tag{3a}$$

Use of log-linear models to model agreement is not new (Agresti 1992; 2002); the idea is that the logarithm of the expected proportion of cases with a given an attribute pattern can be represented by a linear regression model. We impose the conventional constraints (Haberman 1979; 1988, p.196) $\sum_u \lambda_u^U = \sum_a \lambda_a^A = \cdots = \sum_u \lambda_{ua}^{UA} =$

$\sum_a \lambda_{ua}^{UA} = \cdots = \sum_a \lambda_{ubd}^{UBD} = \sum_b \lambda_{ubd}^{UBD} = \sum_d \lambda_{ubd}^{UBD} = 0$. The constraints do not affect the estimates of the probabilities. The interaction parameters λ_{ua}^{UA} and λ_{ub}^{UB} explicitly allow for dependence of the judge's classification and the jury's classification on the true state. Similarly, the parameter λ_{ac}^{AC} allows for dependence between the judge's classification and

judge's rating of the strength of evidence and λ_{bd}^{BD} allows for analogous dependence on the jury side; the parameters λ_{uac}^{UAC} and λ_{ubd}^{UBD} allow that dependence to vary with the true state U. The parameter λ_{cd}^{CD} allows for dependence between the judge's rating and the jury's rating of the strength of evidence. The model assumes that the judge's verdict (A) is conditionally independent of the jury's verdict (B) given the assessments of strength of evidence (C, D) and the true state U. Note that this is an oversimplification, in that it ignores dependence induced by strong but misleading evidence, for example. Assignments of attributes to cases are assumed to occur independently across cases. The model was fitted to the data in Table 3 using the DNEWTON program of Haberman (1988); standard errors were obtained by jackknifing one case at a time, as discussed in Section III.D.

The limited sample size of the NCSC data, especially in the context of latent class models such as (3a), can lead to large standard errors for estimates of probabilities of interest. For the data at hand, we can attempt to obtain smaller standard errors by looking at sub-models of (3a) – i.e., models obtained by restricting additional parameters in (3a) to be zero – that fit the data in Table 3 about as well as (3a) despite having fewer parameters. A simplified version of model (3a) omits three-way interactions among the true state, judge's classification, and judge's assessment of strength of evidence and among the true state, jury's classification, and jury's assessment of strength of evidence. That is, the model is given by (3a) but with the restrictions that

$$\lambda_{uac}^{UAC} = \lambda_{ubd}^{UBD} = 0. \qquad (3b)$$

Under this model, the standard errors of the estimates decrease or remain the same, while the overall fit to Table 3 is almost as close as for model (3a). In light of model checking discussed below, the model (3b) may be viewed as providing estimates that are not any more biased than those under model (3a). Although this statement may seem weak, the fact is that no model will be perfectly correct, and hence estimates from any model may be biased to some degree. Three other sub-models of (3a) were considered, labeled (3c)-(3e), along with an additional model, labeled (4), that included an additional latent variable for strength of evidence; none of those models was found to be satisfactory for the NCSC data, as discussed in Section V.D, below.

B. Estimated Probabilities of Correct Verdict

Table 4 summarizes the results of fitting the various models. The estimates of probabilities should be considered only for models (3a) and (3b); results for the other models are presented for completeness, because the models might be relevant for future datasets. The estimated probability of a correct decision by the jury ($\hat{p}^B$) ranges from 0.83 (0.03) to 0.85 (0.04); estimated standard errors are in parentheses. The estimated probability of correct decision by the judge ($\hat{p}^A$) is slightly higher, ranging from 0.87 to 0.88 with estimated standard errors of 0.04. The estimated difference in accuracy rates, $\hat{p}^B - \hat{p}^A$, is alternatively estimated at 0.02 (0.06) and 0.05 (0.05) under models (3a) and (3b), respectively. Overall, the data suggest, that p^B for the NCSC study is around 84% and that the judges are a bit more accurate on average than the juries. Given the limitations of the data, standard errors, and the uncertainty of the model specifications, the estimates cannot be viewed as conclusive. Recall from Section III.D that the standard errors are based on an assumption of simple random sampling, which might yield standard errors that are too small.

The jury appears to be more accurate than the judge when the true verdict was not guilty ($\hat{p}^{B \bullet U}_{0 \bullet 0} > \hat{p}^{A \bullet U}_{0 \bullet 0}$), although the standard errors are relatively large; the estimates of $\hat{p}^{B \bullet U}_{0 \bullet 0} - \hat{p}^{A \bullet U}_{0 \bullet 0}$ under (3a) and (3b) are 0.19 (0.12) and 0.12 (0.10), respectively. The estimates tentatively suggest that the type I error rate is lower for juries than judges. Conversely the judge appears to be more accurate when the true verdict was guilty ($\hat{p}^{A \bullet U}_{1 \bullet 1} > \hat{p}^{B \bullet U}_{1 \bullet 1}$); the estimates of $\hat{p}^{A \bullet U}_{1 \bullet 1} - \hat{p}^{B \bullet U}_{1 \bullet 1}$ are, respectively, 0.09 (0.05) and 0.12 (0.04).[9] These findings are not unexpected in light of the data in Table 1 showing juries more likely to acquit and judges more likely to convict.

From the preceding estimates, it appears that type I error rates could be *higher* than type II error rates. For example, the jury's type I error rate is $\hat{p}^{B \bullet U}_{1 \bullet 0} = 1 - \hat{p}^{B \bullet U}_{0 \bullet 0}$ and the jury's type II

[9] These standard error estimates are smaller than those for $\hat{p}^{B \bullet U}_{0 \bullet 0} - \hat{p}^{A \bullet U}_{0 \bullet 0}$ because the latter is based on a smaller effective sample size, since $\hat{p}^U_0 \leq 0.28$.

error rate is $\hat{p}_{0\bullet 1}^{B\bullet U} = 1 - \hat{p}_{1\bullet 1}^{B\bullet U}$, so the difference, type I error rate minus type II error rate, is estimated under the two models as 0.08 (0.09) and 0.12 (0.08), respectively. If real, this finding runs counter to beliefs that juries should have higher type II error rates than type I. However, we estimate that juries commit fewer type I errors type II errors. The distinction occurs because the estimates suggest that almost 3/4 of the defendants actually are guilty – models (3a) and (3b) estimate the proportion of cases that truly are not guilty as 0.27 (0.05) and 0.28 (0.04), respectively.

Table 4. Estimates of Probabilities Under Various Log-linear Models. (Estimated standard errors in parentheses)

	model					
statistic	(3a)	(3b)	(3c)	(3d)	(3e)	(4)
$\hat{p}_0^U$	0.27 *(0.05)*	0.28 *(0.04)*	0.15 *(0.03)*	0.31 *(0.04)*	0.94 *(0.17)*	0.31 *(0.12)*
$\hat{p}_1^U$	0.73 *(0.05)*	0.72 *(0.04)*	0.85 *(0.03)*	0.69 *(0.04)*	0.06 *(0.17)*	0.69 *(0.12)*
$\hat{p}^A$	0.87 *(0.04)*	0.88 *(0.04)*	0.91 *(0.01)*	0.86 *(0.03)*	0.25 *(0.16)*	0.85 *(0.09)*
$\hat{p}^B$	0.85 *(0.04)*	0.83 *(0.03)*	0.82 *(0.03)*	0.85 *(0.03)*	0.34 *(0.09)*	0.83 *(0.02)*
$\hat{p}^A - \hat{p}^B$	0.02 *(0.06)*	0.05 *(0.05)*	0.09 *(0.03)*	0.01 *(0.04)*	-0.09 *(0.08)*	0.02 *(0.01)*
$\hat{p}_{0\bullet 0}^{A\bullet U}$	0.61 *(0.09)*	0.63 *(0.09)*	0.85 *(0.05)*	0.57 *(0.07)*	0.20 *(0.04)*	0.57 *(0.19)*
$\hat{p}_{1\bullet 1}^{A\bullet U}$	0.97 *(0.03)*	0.98 *(0.02)*	0.93 *(0.02)*	0.99 *(0.01)*	0.99 *(0.01)*	0.98 *(0.02)*
$\hat{p}_{0\bullet 0}^{B\bullet U}$	0.80 *(0.09)*	0.75 *(0.07)*	0.94 *(0.01)*	0.75 *(0.06)*	0.31 *(0.03)*	0.73 *(0.12)*
$\hat{p}_{1\bullet 0}^{B\bullet U}$	0.20 *(0.09)*	0.25 *(0.07)*	0.06 *(0.01)*	0.25 *(0.06)*	0.69 *(0.03)*	0.27 *(0.12)*
$\hat{p}_{1\bullet 1}^{B\bullet U}$	0.87 *(0.04)*	0.86 *(0.03)*	0.80 *(0.03)*	0.89 *(0.03)*	0.78 *(0.03)*	0.88 *(0.06)*
$\hat{p}_{0\bullet 1}^{B\bullet U}$	0.13 *(0.04)*	0.14 *(0.03)*	0.20 *(0.03)*	0.11 *(0.03)*	0.22 *(0.03)*	0.12 *(0.06)*
$\hat{p}_{1\bullet 0}^{B\bullet U} - \hat{p}_{0\bullet 1}^{B\bullet U}$	0.08 *(0.09)*	0.12 *(0.08)*	-0.14 *(0.05)*	0.15 *(0.06)*	0.47 *(0.05)*	0.15 *(0.02)*
$\hat{p}_{0\bullet 1}^{U\bullet A}$	0.13 *(0.05)*	0.13 *(0.05)*	0.03 *(0.01)*	0.17 *(0.04)*	0.92 *(0.20)*	0.16 *(0.13)*
$\hat{p}_{0\bullet 1}^{U\bullet B}$	0.08 *(0.04)*	0.10 *(0.03)*	0.01 *(0.005)*	0.11 *(0.03)*	0.93 *(0.19)*	0.12 *(0.09)*
$\hat{p}_{1\bullet 0}^{U\bullet A}$	0.13 *(0.12)*	0.07 *(0.07)*	0.34 *(0.10)*	0.05 *(0.04)*	0.005 *(0.01)*	0.08 *(0.08)*
$\hat{p}_{1\bullet 0}^{U\bullet B}$	0.30 *(0.09)*	0.32 *(0.09)*	0.55 *(0.09)*	0.24 *(0.07)*	0.05 *(0.12)*	0.27 *(0.02)*
L^2 [a]	4.87	5.27	21.99	53.27	52.56	22.00
d.f. [a]	13	16	20	20	21	20
AIC_{ent} [a]	2.5814	2.5711	2.5872	2.6499	2.6399	2.6167
BIC_{ent} [a]	2.7276	2.6974	2.6869	2.7446	2.7329	2.7695
ent [a]	2.5002	2.5010	2.5318	2.5895	2.5882	2.5318
GH_{ent} [a]	2.5601	2.5677	2.5867	2.6429	2.6399	2.5749

See text for explanation of notation and models.

The proportion of convicted defendants who actually were not guilty, i.e., the conditional probability of a wrongful conviction, may be estimated by $\hat{p}_{0\bullet 1}^{U\bullet B}$. The estimates under the two models are 0.08 (0.04) and 0.10 (0.03).

C. Additional Limitations

The interpretation of the latent variable U requires care: does it refer to correct status under the omniscient viewpoint, or under the procedural viewpoint, or something else? Although we want U to refer to the omniscient view of the true state, the data analysis uses statistical patterns of agreement observed in real cases to estimate the parameters in the models. Cases where judge and jury agree on acquittal because the evidence is simply weak may be classified as having higher odds that $U = 0$ for that reason rather than because they agree on acquittal from the omniscient viewpoint described in Section I. The proportion of such cases is not known, due to the question wording concerning evidentiary strength (see Section IV), but if those cases are solely within the group classified as having weak evidentiary strength, their proportions are not large (7% according to the judge's assessment and 9% according to the jury's). A labor-intensive way to investigate the validity of U would be to review (a sample of) the cases one by one and compare their predicted status of U, namely $\hat{p}_0^U, \hat{p}_1^U$, with the record (transcripts, interviews with judge and jury, etc.).

The standard error of an estimator depends on the sampling design used for the study. The present analysis does not formally generalize its estimates of accuracy beyond the cases in the NCSC, and thus there is no sampling of judges or juries involved. If the cases were taken as a random sample from a larger population of cases, the calculation of standard error would need to account for clustering of cases by judges; such a calculation is easily carried out with the jackknife or other methods.

The effects of dependence between judge and jury error on the same case are still pertinent to the empirical estimates. Vacek (1985) has shown that positive dependence leads to overstatement of the accuracy rates in estimates based on the Hui-Walter model, analogous to what was demonstrated in Section III, and it is conjectured that the effect is in the same direction for the log-linear models such as (3a)-(3b). Although models for

dependence have been proposed for the Hui-Walter model, some require an additional "rater" in addition to the judge and the jury, and so they are not applicable to the NCSC data (Qu, Tan, and Kutner 1996; Yang and Becker 1997). The estimation of dependence using the adaptation of the Hui-Walter model by Sinclair and Gastwirth (1996, 966) would involve yet other simplifying assumptions. It would be desirable to test the models directly, but to do that would require data where the true state was known with some degree of certainty.

D. Alternative Models

The remainder of this section is technical and discusses alternative models that were tried but found unsatisfactory. The reader should feel free to skip this section with no loss of continuity. To compare alternative models, 5 sets of statistics will be considered in addition to estimates of standard errors. The likelihood-ratio chi-square statistic (L^2 in Table 4) and the degrees of freedom (d.f.) is useful for comparing sub-models with the more general model, but the distributional properties depend on the more general model being correct. An alternative measure, that does not depend on an assumption that the model is correct, is an estimate of expected entropy, say ent, which may be written as $\text{ent} = \sum_j \left(n_j / n\right) \ln\left(\hat{n}_j / \hat{n}\right)$ where n_i is the observed count in cell j of Table 3 ($j = 1, \ldots, 36$), $\hat{n}_i$ is the estimated count based on the fitted model, and the total count is $n = \hat{n} = 271$. Gilula and Haberman (1994, 650-651; 1995, 1138) provide a bias-corrected version of ent, which we will call GH_{ent}. Alternative measures of model performance are the AIC criterion of Akaike and the BIC criterion of Schwarz; these will be scaled to be comparable to ent, and in fact we will use $\text{AIC}_{\text{ent}} = \text{ent} + \left(36 - \text{d.f.}\right)/n$ and $\text{BIC}_{\text{ent}} = \text{ent} + \left(36 - \text{d.f.}\right) \ln\left(n\right) / \left(2n\right)$. We will refer to these various measures as penalty measures; see Gilula and Haberman (2001) for discussion.

The fit of model (3a) as assessed by the likelihood ratio chi-square statistic (L^2 in Table 4) is 4.87 with 13 degrees of freedom, whereas (3b) has L^2 = 5.27 with 16 degrees of freedom; the difference in L^2 is quite small, only 0.40 with 3 degrees of freedom. The measures AIC_{ent} and BIC_{ent} are smaller for (3b) than for (3a), and ent and GH_{ent} are only slightly larger. Those measures are subject to appreciable sampling error, however;

e.g., the difference between GH_{ent} for the two models is only 0.008 whereas the estimated standard error of the difference is 0.02. There is no real evidence that (3b) is less appropriate a model than (3a). The tradeoff of possible model bias against reduced standard errors looks favorable for model (3b).

Additional submodels of (3a) will now be described, and then a model with an additional latent variable representing strength of evidence. None of the submodels were as successful as the submodel (3b), however, in that the measure of fit deteriorated and the estimates of probabilities changed markedly relative to the estimated standard errors. One further simplification from (3a) is assume away any two-way interactions between the true state and the assessments of the strength of evidence. This model is (3b) with the additional restriction that

$$\lambda_{uc}^{UC} = \lambda_{ud}^{UD} = 0 . \qquad (3c)$$

The lack of fit increased from L^2 = 5.27 for (3b) with 16 degrees of freedom to L^2 = 21.99 with 20 degrees of freedom. The change is statistically significant (*p*-value of 0.002) when assessed with a chi-square distribution. Although estimated standard errors decreased (except for $\hat{p}_{1\bullet 0}^{U\bullet A}$), the estimate $\hat{p}_{1\bullet 0}^{B\bullet U}$ of the probability of a type I error decreased by 0.20 from model (3b) to (3c), a change that at 2.7 times its estimated standard error (0.07) is unlikely if both models (3b) and (3c) are correct. (Note that if model (3c) is correct, so are models (3b) and (3a). Standard errors for differences across models were estimated with the jackknife and are not shown in Table 4.) Other changes were large in magnitude but not as large in terms of standard errors, and could reflect chance variation in light of the multiple comparisons being made. The estimated proportion of true verdicts decreased by 0.12 from (3a), from 0.27 in (3a) to 0.15 in (3c), a difference that was large in magnitude compared to its estimated standard error (0.05). The estimate $\hat{p}_{1\bullet 0}^{U\bullet A}$ increased enormously, from 0.13 in (3a) to 0.34; the estimated standard error for the change is 0.12. The changes in the various penalty measures were increases over their values from (3b) except for decrease of 0.01 in BIC_{ent}, which is not appreciable in light of the sampling variability. This model does not seem as trustworthy as (3b).

An alternative simplification from (3b) is to assume away any two-way interactions between classification of the case and assessment of strength of evidence, whether by the judge or jury. This model is (3b) with the additional restriction that

$$\lambda_{ac}^{AC} = \lambda_{bd}^{BD} = 0. \qquad (3d)$$

Figure 1 casts doubt on model (3d), however. The lack of fit increased from L^2 = 5.27 for (3b) with 16 degrees of freedom to L^2 = 53.27 with 20 degrees of freedom. The penalty measures were all much larger than for (3a).

One final simplification from (3b) is a cynical one: disallow direct interaction between the true state and either the judge's classification or the jury's classification. Under this model, the judge's (or jury's) classification of the case is conditionally independent of the true state, given the assessment of strength of evidence. This model, if correct, would indicate profound problems with the legal system. Formally, this model is (3b) with the additional restriction that

$$\lambda_{ua}^{UA} = \lambda_{ub}^{UB} = 0. \qquad (3e)$$

As shown in Table 4, this model estimates that 94% of the cases truly are not guilty, which is markedly (and significantly) different than the estimate of 27% under the more general model (3a). The difference in the estimates $\hat{p}_0^U$ is 0.67, which is much larger than its estimated standard error (0.16). The implications of having such a high proportion of truly not guilty cases are low estimates of judge's and jury's probabilities of correct decisions, 0.25 and 0.34, respectively. Further, the penalty measures are larger than for (3a). Thus, this model may be rejected.

An alternative model formulation – not a submodel of (3a) – includes a second unobserved attribute, V, which underlies the judge's and jury's assessments of strength of evidence (C, D) and which takes values 1, 2, 3. In this model, the probability that a case has one of the 216 possible patterns of attributes is approximated by the hierarchical log-linear model

$$\log p^{UVABCD}_{uvabcdi} = \alpha + \lambda^{U}_{u} + \lambda^{V}_{v} + \lambda^{A}_{a} + \lambda^{B}_{b} + \lambda^{C}_{c} + \lambda^{D}_{d} + \lambda^{UA}_{ua} + \lambda^{UB}_{ub} + \lambda^{VC}_{vc} + \lambda^{VD}_{vd} + \lambda^{AC}_{ac} + \lambda^{BD}_{bd}, \quad (4)$$

with the constraints $\sum_u \lambda^U_u = \sum_v \lambda^V_v = \cdots = \sum_u \lambda^{UA}_{ua} = \sum_a \lambda^{UA}_{ua} = \cdots = \sum_b \lambda^{BD}_{bd} = \sum_d \lambda^{BD}_{bd} = 0$. This model allows the judge's classification of the case to depend on the true state (U) and on judge's assessed strength of evidence, and it allows the latter to depend on the unobserved underlying strength of evidence (V). The estimated standard errors for this model typically are larger than those for model (3a), which is not surprising as the model involves two latent variables. The penalty measures for (4) are all larger than for (3a) and (3b). At least for the NCSC data, this model does not appear to provide more trustworthy estimates than (3a) or (3b).

VI. SUMMARY AND CONCLUSIONS

We have shown that accuracy of jury verdicts can be studied empirically and systematically. Two different kinds of estimators were considered. A simple estimator based on the rate of agreement between judge and jury was developed and analysis indicated that it could tend to overestimate accuracy but would not tend to underestimate accuracy. The jury verdict was estimated to be accurate no more than 87% of the time for the cases in the recent study by the National Center for State Courts National Center (NCSC) and no more than 89% for the cases studied by Kalven and Zeisel in the 1950s. The NCSC cases should *not* be regarded as a representative sample with respect to jury accuracy, however.

Under a log-linear model based on different assumptions and utilizing additional components of the NCSC data, the estimate of accuracy dropped to 0.83 – 0.85. Some limitations of the log-linear analysis should be kept in mind: (i) the judge's classification and the jury's classification for a given case are assumed to be conditionally independent given the strength of the evidence, when in fact both classifications may be affected by presentation of false evidence; (ii) the specification of the log-linear model may not be correct; (iii) the classification of "correct" in the log-linear model may not correspond to our interpretation of "correct", (iv) sampling error may be underestimated. In light of these

limitations, the empirical estimates from the data analysis must be interpreted with great caution and in no event should be generalized beyond the NCSC study. That said, some *estimates* of accuracy of verdicts for the criminal cases included in the NCSC study are presented below, based on the log-linear model denoted (3b). The estimates are no basis for action other than future studies. Numbers below in parentheses are estimates of standard errors.

1. Jury verdicts were incorrect 15% of the time, with an estimated standard error of 4 percentage points. If the assumption of conditional independence between judge and jury error is incorrect, this estimate of error rate will tend to be too low.

2. The estimated proportion of defendants who were not guilty is around 28%, with an estimated standard error of 4 percentage points.

3. The conditional probability that the jury incorrectly convicts (a "type I error") was estimated at 0.25 (0.07), the conditional probability that a jury incorrectly acquits ("type II error") was estimated at 0.14 (0.03), and the difference was estimated at 0.12 (0.08). The estimated *number* of defendants in the NCSC cases who truly are not guilty but are convicted does seem to be smaller than the *number* who truly are guilty but are acquitted.

4. The type I error rate for the jury is estimated to be smaller than the type I error rate for the judge by 0.12 (0.10). The type I error rate for the judge was estimated at 0.37 (0.09), and the type II error rate for the judge was estimated at 0.02 (0.02).

5. The conditional probability of a wrongful conviction, given that the defendant was convicted, is estimated at 0.10 (0.03).

In conclusion, studies of judge-jury agreement are a powerful tool for assessing the performance of the jury system. The agreement rates themselves are statistical indicators and should be constructed and compared. [10] A variety of estimates of accuracy can be developed from the data. Additional studies of judge-jury agreement should be carried out to provide for each jury verdict in the study a second rating, ideally with equal or better

[10] Caution must be taken, however, that a statistical indicator is not used to regulate, reward, or punish – for that may well lead to corruption of the indicator. Wartime statistics on body counts is one well known example.

accuracy than the jury's. The rating could come from judge in the case, but it might also come from an observer (or observers) other than the sitting judge, for example by a retired judge or other expert; see Gastwirth and Sinclair (1998) for discussion and additional suggestions. Such studies would be strengthened by carefully worded questions assessing of strength of evidence and, perhaps, by confidence of judge and jury in their decisions. The studies could even be designed to dovetail with cases where the correct verdict would later be known. There are many possibilities for improving the studies, but the most important step is to carry out more studies of judge-jury agreement so that we can assess and compare accuracy rates over time, across jurisdictions, for different kinds of cases, for consistency between demographic characteristics of defendant and jury, etc.

REFERENCES

Agresti, A (1992) Modelling Patterns of Agreement and Disagreement. *Statistical Methods in Medical Research* 1, 201 – 218.

Agresti, A (2002) *Categorical Data Analysis*. 2nd ed. New York: Wiley.

Borchard EM (1932) *Convicting the Innocent: Errors of Criminal Justice.* New Haven: Yale University Press.

Efron B and Tibshirani RJ (1993) *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Eisenberg T, Hannaford-Agor PL, Hans VP, Waters NL, Munsterman GT, Schwab SJ, Wells MT (2005) Judge-Jury Agreement in Criminal Cases: A Partial Replication of Kalven and Zeisel's *The American Jury*. *Journal of Empirical Legal Studies* 2, 171-206.

Fleiss JL (1981) *Statistical Methods for Rates and Proportions*. 2nd ed. New York: Wiley.

Gastwirth JL and Sinclair MD (1998) Diagnostic Test Methodology in the Design and Analysis of Judge-Jury Agreement Studies. *Jurimetrics* 39, 59-78.

Gastwirth JL and Sinclair MD (2004) A Re-examination of the 1966 Kalven-Zeisel Study of Judge-Jury Agreements and Disagreements and Their Causes. *Law, Probability and Risk* 3, 169-191.

Gilula Z and Haberman SJ (1995) Prediction Functions for Categorical Panel Data. *The Annals of Statistics* 23, 1130-42.

Gilula Z and Haberman SJ (2001) Analysis of Categorical Response Profiles by Informative Summaries. *Sociological Methodology* 31, 193-211.

Haberman SJ (1979) *Analysis of Qualitative Data*, Vol 2. New York: Academic Press.

Haberman SJ (1988) A Stabilized Newton-Raphson Algorithm for Log-Linear Models for Frequency Tables Derived by Indirect Observation. *Sociological Methodology* 21, 129-187.

Hannaford-Agor PL, Hans VP, Mott NL, and Munsterman GT (2003) Evaluation of Hung Juries in Bronx County, New York, Los Angeles County, California, Maricopa County, Arizona, and Washington, DC, 2000-2001. National Center for State Courts User Guide. Williamsburg, VA: National Center for State Courts [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

Hui SL and Walter SD (1980) Estimating the Error Rates of Diagnostic Tests. *Biometrics* 36, 167-171.

Kalven H and Zeisel H (1966) *The American Jury*. Boston: Little, Brown. X

Kruskal W (1988) Miracles and Statistics: The Casual Assumption of Independence. *Journal of the American Statistical Association* 83, 929-940.

Lehmann EL (1966) Some Concepts of Dependence. *Annals of Mathematical Statistics* 37, 1137-1153.

Manski CF (1995) *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Qu YS, Tan M, Kutner MH (1996) Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests. *Biometrics* 52, 797-810.

Sinclair MD and Gaswirth JL (1996) On Procedures for Evaluating the Effectiveness of Reinterview Survey Methods: Application to Labor Force Data. *Journal of the American Statistical Association* 913, 965-969.

Vacek PM (1985) The Effect of Conditional Dependence on the Evaluation of Diagnostic Tests. *Biometrics* 41, 959-968.

Yang I and Becker MP (1997) Latent Variable Modeling of Diagnostic Accuracy. *Biometrics* 53, 948-958.

DERIVATIONS

To derive (1), note first that

$$\begin{aligned}\rho_i &= \omega_i + p_i^A p_i^B + \left(1 - p_i^A\right)\left(1 - p_i^B\right) \\ &= \omega_i + 2p_i^A p_i^B + 1 - p_i^A - p_i^B.\end{aligned}$$

Averaging over the N cases and substituting, we have

$$\begin{aligned}\rho &= \omega + 2\sigma_{AB} + 2p^A p^B + 1 - p^A - p^B \\ &= \gamma + 2p^A p^B + 1 - p^A - p^B \\ &= \gamma + 2(p^B + \delta)p^B + 1 - (p^B + \delta) - p^B \\ &= 2(p^B)^2 + 2(\delta - 1)p^B + 1 + \gamma - \delta.\end{aligned}$$

Proof of lower bound for ρ (as asserted in the second sentence following (1)).

The roots to (1) are given by $p^B = .5\left(1 - \delta \pm \sqrt{2(\rho - \gamma) - 1 + \delta^2}\right)$. Because the roots must be real, the quantity under the square root sign must be non-negative, and thus we have $\rho \geq \gamma + (1 - \delta^2)/2$.

Proof that identification error equals $f(\rho, \gamma, \delta) - p^B$.

Since the values of p^B, ρ, γ, and δ all occur, they must satisfy (1) and in particular $\rho \geq \gamma + (1 - \delta^2)/2$. It follows that lower constraint in (2) is unnecessary for evaluating $f(\rho, \gamma, \delta)$, and so $f(\rho, \gamma, \delta) = \min\left\{p_{high}^B, 1\right\}$. Thus, the identification error equals $\min\left\{p_{high}^B, 1\right\} - p^B = f(\rho, \gamma, \delta) - p_{high}^B$.

ADDENDUM

Jack Heinz asked about the extent to which the Type I error rate for judges is a function of the jury conviction rate. That is, if juries convicted more often (so that their rate was closer to that of judges), would the estimate of the judges' Type I rate would be lower, and by how much?

To address this question, I weighted the NCSC counts in Table 3, assigning a weight of 4 to each case with $(A,B)=(0,1)$ and assigning a weight of 2 to each case with $(A,B)=(1,1)$; all other cases received a weight of 1. Prior to the weighting, the judge conviction rate for Table 3 as 81.2% and the jury rate was 69.4%; after weighting, the rates were 80.4% and 83.0%, respectively. The agreement rate was 77.1% before weighting and 78.1% after.

The estimated probabilities of type I and type II error for judges changed from 37.5% and 1.8% before the weighting to 37.4% and 5.1% after. Thus, the type I error rate hardly changed. Similarly, the estimated proportion of not guilty cases $(U=0)$ increased from 28.0% to 29.3%.

ACKNOWLEDGMENT

The author thanks Juha Alho, Ron Allen, Kenworthey Bilz, Shari Seidman Diamond, Ted Eisenberg, Steve Fienberg, Joseph Gastwirth, Jack Heinz, Wenxin Jiang, Chuck Manski, Dorothy Roberts, and Sandy Zabell for helpful comments. He is deeply indebted to Shelby Haberman for suggestions and patient help with modeling. Responsibility for errors rests with the author. This research was supported by the Institute for Policy Research, Northwestern University.