On Measuring the Balance between Wrongful Convictions and Wrongful Acquittals in Criminal Trials[*]

Bruce D. Spencer[†]

November 7, 2007

Abstract

Verdicts in criminal trials can err by convicting an innocent person or acquitting a guilty defendant. Recent work shows how the rates and relative numbers of both types of errors can be measured statistically both for judge verdicts and for jury verdicts (Spencer 2007, Gastwirth and Sinclair 1998). The key is to have multiple ratings of the cases, e.g., also record the judge's belief of the correct verdict in a jury trial or an expert observer's belief of the correct verdict in a non-jury trial. Analysis of a non-random sample of state court cases in 2000-01 from four jurisdictions studied by the National Center for State Courts (Hannaford-Agor et al 2003) yields estimates that approximately 17% of the jury verdicts were incorrect: 7% were wrongful convictions and 10% were wrongful acquittals; it was estimated that 12% of the judges' verdicts were incorrect: 10% of judges' verdicts wrongful convictions and 1% wrongful acquittals. Those numbers are subject to a number of limitations, not least the non-representative nature of the sample of cases; this paper reviews those limitations not only for the analysis of the NCSC data but, more importantly, with an eye toward future studies.

KEY WORDS: latent class model; legal statistics; law

1. Introduction and Overview

Verdicts in criminal trials can and do err materially when an innocent person is convicted or a guilty defendant is acquitted. Verdicts rest on human processing of imperfect evidence and both kinds of errors do occur. In this paper we discuss how statistics can provide estimates of the various rates of error even though the correct verdict cannot be known on a case by case basis. We illustrate those methods with data from a recent empirical study (Sections 4.4, 6.3). The data are described in Section 2 and descriptive statistics are presented in Section 3. The quality of the statistical measurements is an important concern and will be discussed as well (Section 7).

1.1 Meaning of Correct Verdicts and Accuracy

To be clear about the meaning of "error", we need some clarity about what we mean by "truth". From a procedural or probatory perspective, the true state of a defendant is guilty if proof of the crime has been demonstrated to the standards required by the law, and otherwise the true state is not guilty. From an omniscient or material perspective, the true state is guilty if the defendant actually committed the crime (Laudan 2005). The latter perspective has the advantage of being independent of the courtroom proceedings, and casual observation leads me to believe that it conforms to popular notions of justice. Both perspectives are considered in this paper. A verdict is correct if it matches the defendant's true state (from whichever perspective). Some trials do not reach verdicts, as when there is a deadlocked or "hung" jury, and such outcomes may be treated in various ways (Gastwirth and Sinclair 1998, 69ff). In this paper, we exclude hung jury cases from consideration.

Accuracy refers to the proportion of verdicts that are correct, with reference to a group of cases. Groups can be based on geography, type of case, demographic characteristics of defendant and jury, etc.

1.2 Level of Error

Proportions of verdicts that are correct can be measured statistically when a second rating (replication) of the case is obtained (Section 2.1). Such replications are available for a non-random sample of criminal cases in 2000-01 studied by the National Center for State Courst (NCSC), as well as cases studied by Kalven and Zeisel (1966). Analysis of the NCSC data suggest that jury verdicts were correct in no more than about 87% of the cases in the more recent study

[*] Presented at Second Annual Conference on Empirical Legal Studies, NYU Law School, New York, November 9-10, 2007.

[†] Professor, Department of Statistics, and Faculty Fellow, Institute for Policy Research, Northwestern University. Mailing address: Department of Statistics, 2006 Sheridan Road, Evanston, IL 60201-4109.

(Section 4.4). Limitations of the estimate are discussed (e.g., Section 7), and the reader is cautioned not to apply that error rate to any cases outside the study itself.

1.3 Balance of Error

There has long been concern with the balance of two kinds of error, specifically that it is worse to convict one innocent person than to acquit n guilty people. As Volokh (1997) documents, various values of n have been advocated over the centuries. The most well-known value may be Blackstone's choice of n =10: it is better that ten guilty persons escape than that one innocent suffer" (Blackstone 1825, 358). There is no unique "right" value of n, as the choice of n reflects societal values, the type of crime and punishment under consideration, and the distribution of defendants among the guilty and not guilty; see Laudan (2005) for discussion and review of the literature.

There is, however, an empirical balance of errors that occurs in the courts, that is, there is some *actual* ratio between the number of erroneous convictions ("type I errors") and the number of erroneous acquittals ("type II errors"). The value of n in effect in the NCSC cases appears to be surprisingly small, although the estimates have large sampling error. The estimated value of n for the jury is 1.37 (standard error of 0.75) and for the judge it is only 0.13 (standard error of 0.14). Future studies will be needed to see if similar values of n apply more widely to larger sets of cases.

2. Data for Measuring Error Rates of Verdicts When Truth Is Not Observed

2.1 Replication Data

It is well known from sampling theory that sampling variance can be estimated even though the population value is unknown provided that the sampling method utilizes replication. Variance is only a partial measure of accuracy, but it provides a lower bound for the mean squared error. A somewhat similar idea applies for estimating the accuracy of verdicts when the truth is not observed. The key is to observe "replications" of the verdict, which could be provided by courtroom observer (such as a retired judge) in a non-jury trial or by the judge presiding at a jury trial. *Neither rater's verdict is assumed to be necessarily correct.*

2.2 Data Sources for Kalven-Zeisel and NCSC Studies

Judges' verdicts in jury trials have been recorded in the famous study by Kalven and Zeisel (1966), *The American Jury*, and much more recently in a study by the National Center for State Courts (NCSC), described in Hannaford-Agor et al. (2003). The Kalven-Zeisel study covered more than 3500 trials in 47 states and Washington D.C. in 1954-1955 (Sample I) and 1958 (Sample II). Our focus is mainly on the NCSC data, which although based on fewer cases is more current and includes the jury's as well as the judge's assessments of the strength of evidence in the case (Section 2.2.3, below). The NCSC data came from a non-random sample of four courts chosen for the purposes of studying hung juries: Los Angeles County Superior Court, District of Columbia Superior Court, Maricopa County Superior Court (whose jurisdiction includes Phoenix, AZ), and Bronx County Supreme Court. A number of courts had refused to participate in the survey, for reasons of administrative burden as well as "judicial sensitivity to the inviolability of jury deliberations" and "concern. . . that the information collected from jurors might provide a basis for convicted defendants to appeal the verdict" (Hannaford-Agor et al. 2003, v, vi). The NCSC study produced judge and jury data on 290 non-hung-jury trials for non-capital felony cases in 2000-2001. In addition to being a non-random sample of four areas, cases within the areas were selected with unequal (and unreported) rates from the four jurisdictions, and as a result analyses of the data cannot use sampling weights. For these various reasons (also see Section 7), *inferences about the cases in the study should not be extended to any larger population.*

2.2.1 Standard Errors

Standard errors are calculated for statistics computed from the NCSC sample, but the standard errors are themselves hypothetical for a number of reasons. For one thing, the cases are not a random sample. For another, the statistics from the study are not being used for making inferences about a larger population. The standard error calculations were designed to provide yardsticks of stability of the statistics with respect to the sample, and for that reason they were

calculated with a delete-one jackknife. Thus, these are "nominal" standard errors and they should not be interpreted in a formal sense of applying to a sampling distribution.[3]

2.2.2 Recording and Classifications of Verdicts

There may be some reporting error in the judge's verdict. Because the trials were jury trials, the judge's verdict recorded in the studies was not the actual verdict from the judge. Judges in the NCSC study were asked, "If you had decided this case in a bench trial, would you have rendered a verdict for the prosecution or for the defense?" Thus, the judge's verdict was a hypothetical one and might differ from the verdict the judge would have issued had the trial been a non-jury trial. Still, if the trial had been a non-jury trial then other aspects of the trial could have differed as well, such as how the witnesses were questioned by the attorneys.

The NCSC study asked questions about a variety of counts, and the overall classification of the verdict as "guilty" or "not guilty" involved a number of steps. The classifications being used here are adopted from the coding for the Eisenberg et al (2005) analysis, graciously provided by Professor T. Eisenberg.

2.2.3. Strength of Evidence

A decision by the judge or jury to acquit could indicate belief either in material innocence (the defendant did not commit the crime) or lack of probatory guilt (proof was not demonstrated beyond a reasonable doubt). Information about perceptions of strength of evidence is a promising way to help distinguish the two. The Kalven-Zeisel study asked the judges but not the juries in Sample II about strength of evidence, and the NCSC asked both judges and juries about strength of evidence.

Question 12 of Sample II (Kalven and Zeisel 1966, 532) asked the judge the following question.

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?

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 interpret the response to the NCSC question 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 (Eisenberg et al 2005, 186). 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. To obtain a single rating for the jury as a whole, the jurors' (7-point) ratings were averaged for the case and the average rating was classified as strong, medium or weak. Note also that the strength of evidence was reported for the trial as a whole and not necessarily for the count relating to the verdict being analyzed. For further critique, see Spencer (2007, 316ff).

2.2.4 Other Questionnaire Items

In the Kalven-Zeisel study, judges were also asked about other factors that could affect the verdict, including presence/absence of a prior record for the defendant, skill of defense counsel, difficulty of the evidence. Jurors were asked whether the defendant aroused sympathy and whether they disagreed with the law. See Gastwirth and Sinclair (2004) for insightful statistical analysis.

[3] If the cases had been selected randomly, the appropriate calculation of standard error would have accounted for the multi-stage design with courts as the primary sampling unit, rather than using a delete-one jackknife.

In the NCSC study, both jurors and judges were asked about "case complexity, attorney skill, evaluation of the evidence, formation of opinion, the dynamics of the deliberations including the first and final votes, juror participation, conflict, opinion about applicable law, assessment of criminal justice in the community, and demographic information" (Eisenberg et al, 2005, 176).

2.2.5 Other Considerations in NCSC Data Collection

It is unclear how aware of the study judges and attorneys were prior to and during the trial, and hence whether the study itself may after altered behavior as a "Hawthorne effect". Procedures were put into place to ensure jurors' privacy and all respondents' confidentiality, and to prevent use of the study data in adjudicatory proceedings (Hannaford-Agor et al. 2003, vii). These procedures are laudable but prevent further collection of data on the study cases beyond what was collected in the survey (no study of reversals of verdicts, for example, or later DNA analysis). For detailed discussion of the data see Eisenberg et al. (2005) and Hannaford-Agor et al. (2003).

3. Descriptive Statistics

3.1 Judge-Jury Agreement and Acquittal Rates

From the 2 x 2 table cross-classifying judge's verdict and jury's verdict we obtain acquittal rates (or their complements, conviction rates) for each rater, as well as agreement rates, as shown in Table 1. We see that juries acquitted more often than judges, at 30% versus 18% for NCSC cases and similarly at 32% versus 17% for Kalven-Zeisel.

The rates of judge-jury agreement (r) are 77% (NCSC) and 80% (Kalven-Zeisel). The consistency of results between the two studies decades apart and surveying different jurisdictions is remarkable. Yet, the levels of agreement are not high when one takes into account that the level of agreement by chance would be 62% for each study. The chance agreement rate (c) is calculated considering verdicts to be assigned randomly and independently by judge and jury at each rater's observed rate. A chance-corrected rate of agreement is Cohen's kappa, with values $(r-c)/(1-c)$ here equal to 0.47 (Kalven-Zeisel) and 0.38 (NCSC), indicating only fair-to-poor agreement (Fleiss 1981).

3.2. Evidentiary Strength

A rater's perception of the strength of evidence is a strong predictor of the rater's verdict. In the NCSC study, the rating of evidentiary strength for the prosecution varied between judge and jury, so that judge's and jury's assessments of the evidence as strong, medium, or weak matched only 56% of the time (Spencer 2007, 316). Judges and juries alike rated evidentiary strength as weak for only about 10% of the trials; just over half the trials were rated as having medium evidentiary strength (52% by judge, 54% by jury), and a little more than a third were rated as having strong evidence (39% by judge, 35% by jury). Conviction rates for judges (juries) were 27% (17%) for trials whose evidence they rated as weak, 78% (61%) for those whose evidence they rated as medium, and 99% (98%) whose evidence they rated as strong.

For NCSC trials classified by the judge's (jury's) assessment of the strength of evidence, the judge and jury agreed on acquittal for 54% (52%) of cases with weak evidence, 15% (14%) with medium evidence, and 0% (0%) for trials with strong evidence. In contrast, agreement on conviction varied depending on whose classification was used. For cases classified on the judge's (jury's) assessment of the strength of evidence, the judge and jury agreed on conviction for 27% (10%) of cases with weak evidence, 54% (54%) with medium evidence, and 86% (95%) for cases with strong evidence. (Eisenberg et al 2005, Table 4, p. 186)

4. Estimating Accuracy from Agreement

4.1 Upper Bound on Accuracy of Less Accurate Rater

The data on agreement yield a simple upper bound on the accuracy of the less accurate rater. Let r denote the rate of agreement between the two raters' verdicts for a collection of cases and let p denote the proportion of verdicts of a given type of rater (e.g., judge or jury) that are correct. Since the raters disagree on a proportion $1-r$ of the cases, it must be the case that an incorrect verdict is reached by one rater or the other on at least $1-r$ of the cases; thus, the

error rate for the less accurate of the raters (for the cases at hand) is at least $(1-r)/2$, and so we may infer that p for the less accurate rater is no greater than $(1+r)/2$. Substituting an estimate of the agreement rate, we obtain an estimate of an upper bound on p. From the data in Table 1, we obtain 0.885 as the estimate of p for the NCSC cases and 0.90 for p for the Kalven-Zeisel cases.

Table 1: Descriptive Statistics from Non-Hung-Jury Cases in NCSC and Kalven-Zeisel Studies.

	Agreement Rate			*Acquittal Rate*	
Study	*Observed*	*Chance*	*Kappa*	*Judge*	*Jury*
NCSC, 2000-01	77%	62%	0.38	18%	30%
Kalven-Zeisel, 1950s	80%	62%	0.47	17%	32%

NCSC and Kalven-Zeisel statistics are computed from Tables 3 and 2, respectively, in Eisenberg et al. (2005).

4.2 Agreement as Function of Accuracy and Dependence

We can improve on the estimator $(1+r)/2$ by adjusting for the extent of agreement that occurs when both verdicts are incorrect. Consider a study having N cases with a different jury for each case. For case i the probability of choosing a correct verdict is p_i^A for the judge and p_i^B for the jury. The probability that the judge and jury choose the same verdict (correctly or not) is ρ_i. If the judge and the jury for a case were to choose their verdicts independently, the probability of agreement would be $p_i^A p_i^B + (1-p_i^A)(1-p_i^B)$. 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, $\omega_i = \rho_i - p_i^A p_i^B - (1-p_i^A)(1-p_i^B)$. 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 $\sigma_{AB} = \sum_i (p_i^A - p^A)(p_i^B - p^B)/N$. 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 overall dependence, $\gamma = \omega + 2\sigma_{AB}$.

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

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

This equation allows us to interpret γ as the excess agreement that is expected when independence is not present.

4.3 A Sharper Upper Bound on Accuracy

If $\gamma \geq 0$ then

$$\hat{p} = \begin{cases} 0.5(1+\sqrt{2r-1}) & \text{if } 1/2 \leq r \\ 0.5 & \text{if } r < 1/2. \end{cases} \tag{2}$$

also provides an upper bound on the accuracy of the less accurate rater but is tighter than the bound in Section 4.1, in that $\hat{p} \leq (1+r)/2$ with strict inequality except when $r = 1$.[5] To see why it is plausible that $\gamma \geq 0$, consider the two

[5] This estimator is derived in Spencer (2007, 310ff). The essential idea is that the expected agreement rate ρ is equal to the probability of both raters being correct plus the probability of both being incorrect. Under simplifying assumptions of equal accuracy and independence, $\rho = p^2 + (1-p)^2$. Substituting $r \geq 1/2$ for ρ we can solve $r = p^2 + (1-p)^2$ for p. The upper bound properties under positive dependence are fortuitous.

components of γ, namely ω and σ_{AB}. The common environment and overlap in exposure to evidence and questioning make it plausible that $\omega \geq 0$. (If for all cases p_i^A and p_i^B take only the values 0 or 1, then each $\omega_i = 0$ and hence $\omega = 0$.) Even if ω_i were 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. The error in $\hat{p}$ can be decomposed into the sum of survey error (arising if the observed value of r is unequal to the population value or expected value) and non-survey error; Spencer (2007, 313) shows that the non-survey error is non-negative when $\gamma \geq 0$, and hence (2) provides an upper bound on accuracy when there is no error in the observed value of r.

The observed agreement rate for the Kalven-Zeisel study was $r = 0.80$, leading to $\hat{p} = 0.89$ for the less accurate rater, with estimated standard error (s.e.) < 0.01. The agreement rate for the NCSC cases (excluding hung juries, as always) was observed to be $r = 0.77$, leading to $\hat{p} = 0.87$, with s.e. = 0.02. These estimates of accuracy are slightly smaller than the estimates from Section 4.1.

4.4 Upper Bounds on Jury Accuracy

If external information is available to tell us which rater is less accurate, then we apply the estimates from Section 4.1 and 4.3 to that rater. There are reasons, both statistical and non-statistical, to believe that judges are more accurate on average than juries ($\delta \geq 0$) for the kinds of cases in the NCSC study. Statistical analyses based on log-linear models (discussed in Section 6.3) 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 material interpretation of correct verdict. Judges' greater experience and knowledge of the law also support the plausibility of $\delta \geq 0$ for the probatory interpretation of correct verdict.

With the assumption that judges are at least as accurate as juries on average, we may apply (2) to estimate that the overall error rate for juries in the NCSC cases was no better than 0.13. If a point value may be specified for δ, then the estimate in (2) for jury accuracy may be modified to

$$\hat{p}_\delta = \begin{cases} 1 & \text{if } 1+\delta < \hat{\rho} \\ \frac{1}{2}(1-\delta+\sqrt{2\hat{\rho}-1+\delta^2}) & \text{if } \frac{1}{2}(1-\delta^2) \leq \hat{\rho} \leq 1+\delta \\ \frac{1}{2}(1-\delta) & \text{if } \hat{\rho} < \frac{1}{2}(1-\delta^2), \end{cases} \tag{3}$$

which is not far from $\hat{p} - \delta/2$. The corresponding estimate for judge accuracy is obtained by replacing δ by $-\delta$, i.e., calculating $\hat{p}_{-\delta}$ from (3). For example, if we specify that $\delta = 0.05$, the estimate of jury accuracy drops from 0.87 to 0.84, and the estimate of judge accuracy becomes 0.89.

5. Types of Errors

A verdict can be incorrect in one of two ways: an innocent person can be convicted (*type I error*) or a guilty person can be acquitted (*type II error*). The distinction between type I and type II errors parallels that in statistical hypothesis testing where the null hypothesis is that the defendant is innocent. A type I error may also be called an wrongful conviction, and a type II error a wrongful acquittal.

Although the number of type I errors is by definition the same as the number of wrongful convictions, and the number of type II errors is the same as the number of wrongful acquittals, the interpretations of the error rates differ. Type I and type II error rates, say $P(\text{type I})$, $P(\text{type II})$, refer to conditional rates of error given the true state of the defendant, whereas wrongful conviction rates and wrongful acquittal rates, say $P(\text{WC})$, $P(\text{WA})$, are conditional on the verdict. A type I error rate of 0.25 indicates that an innocent defendant has a 25% chance of being convicted, whereas an erroneous conviction rate of 25% would mean that 25% of convictions are of innocent defendants. A type II error

rate of 0.10 indicates that a guilty defendant has a 10% chance of being acquitted, whereas an erroneous acquittal rate of 10% would mean that 10% of acquittals are of guilty defendants. One can also consider unconditional rates, e.g., the fraction of all verdicts that are errors of a given type.

The U.S. Constitution is interpreted as requiring fewer type I errors than type II (*In re Winship,* 397 U.S. 358, 1970). To relate error rates and numbers of errors by type, it is necessary to quantify the prevalence of guilt, $P(\text{guilt})$, i.e., the proportion of defendants whose true state is guilty. The prevalence of innocence is $P(\text{innocence}) = 1 - P(\text{guilt})$. For example, the overall error rate, $P(\text{error})$, may be expressed as

$$P(\text{error}) = P(\text{type I})P(\text{innocence}) + P(\text{type II})P(\text{guilt}).$$

The probability of a type I error relates to the probability of a wrongful conviction via

$$P(\text{type I}) = P(\text{WC})P(\text{convict}) / P(\text{innocence}),$$

where $P(\text{convict})$ denotes the fraction of verdicts that are convictions. Similarly, with parallel notation,

$$P(\text{type II}) = P(\text{WA})P(\text{acquit}) / P(\text{guilt}).$$

As discussed in Section 1, the true state (innocence or guilt) can be interpreted either from a probatory or from a material perspective.

6. Latent Class Models

6.1. Generalities

To estimate type I and type II error rates, we need more data than a single 2×2 table for judge and jury verdicts. In addition, some modeling assumptions will be needed. is insufficient. The models relate the observed, or *manifest,* variables such as verdict or evidentiary strength to unobserved categorical variables, or *latent classes*. In addition to the usual statistical modeling considerations, a key question for latent class models in the current context is how well the construct of the latent class within the model corresponds to the defendant's true state of innocence or guilt.

6.2 Hui-Walter Model Applied to Kalven-Zeisel Data

To estimate accuracy of verdicts in the Kalven-Zeisel data, which had two raters (judge and jury) only, Gastwirth and Sinclair (1998) used a model developed by Hui and Walter (1980). The model applies in the following circumstances: cases can be grouped, e.g., by type of crime, such that (i) for each rater $P(\text{type I})$ and $P(\text{type II})$ do not vary across groups but (ii) $P(\text{guilt})$ does vary across the groups. Although assumption (i) is not formally testable from the data, Gastwirth and Sinclair suggested it would apply to two subsets of crimes, auto theft and burglary. They estimated $P(\text{type I})$ at 0.009 for juries and 0.000 for judges, but with large standard errors (0.089 and 0.399, respectively); they estimated $P(\text{type II})$ to be 0.192 for juries and 0.012 for judges, with standard errors 0.44 and 0.015, respectively; they estimated the prevalence of guilt to be 0.921 for burglary and 0.810 for auto theft (Gastwirth and Sinclair 1998, 63). These translate to estimates of overall error rates for the jury of 0.18 (burglary) and 0.16 (auto theft) and for the judge of 0.01 for both, suggesting $\delta \geq 0$. If, as a sensitivity analysis, we increase the estimate of $P(\text{type I})$ for the judge to 0.4 (an increase of one standard error) but leave unchanged the other type I and type II error rates and the prevalence estimate, the estimated overall error rates for the judge increase to 0.04 (burglary) and 0.09 (auto theft). These results still are consistent with $\delta \geq 0$. The judge-jury agreement rates for these cases were 0.81 (burglary) and 0.82 (auto-theft), leading via (2) to "upper bound" estimates of jury accuracy of 0.90 for each type of case, somewhat higher than the Hui-Walter estimates of 0.82 and 0.84, respectively.

Even if the assumptions of the Hui-Walter model are valid for the burglary and auto theft cases in the Kalven-Zeisel study, the model's assumptions likely will not apply across broad categories of cases. With stronger data, however, those assumptions can weakened substantially.

6.3 Models Incorporating Observed Measures of Evidentiary Strength

The NCSC data had an important advantage over the Kalven-Zeisel data in that it included ratings by the jury and by the judge of the strength of the evidence for the prosecution. This enriches the set of data patterns and allows the fitting of statistical models with fewer assumptions than the Hui-Walter model. Spencer (2007) drew on coding of the data by Eisenberg et al (2005) to classify the cases into as many as 36 observable groups, based on judge's verdict (2 levels), jury's verdict (2 levels), judges rating of evidence (3 levels), and jury's rating of evidence (3 levels). Each observed group is an amalgam of the pair of latent classes, those defendants truly guilty and those truly not guilty. Thus, there are 72 possible classes, but only 36 possible pairings of classes can be observed. To analyze the data, Spencer (2007, 318ff) used a latent class model that assumed that the logarithm of the probability that a case falls into one of the 72 classes could be expressed by an ANOVA model. Various forms of the model were tried, and the preferred model (Spencer 2007, 321, model "3b") included as main effects the judge's verdict (*A*), the jury's verdict (*B*), the judge's rating of evidence (*C*), the jury's rating of evidence (*D*), the unobserved true state of the defendant (*U*), and pair-wise interactions (*AU*, *BU*, *CU*, *DU*, *AC*, *BD*, and *CD*). The key assumptions are that the general mathematical form of the model is known and that the cell classifications of the cases are independent. Estimates from the model are shown in Table 2; results are taken from Spencer (2007, 321) and supplementary calculations.

Table 2: Estimates of Accuracy for NCSC Cases (nominal s.e. in parentheses)

Quantity Being Estimated	*Jury*	*Judge*	*Difference*
P(correct)	0.83	0.88	0.05
	(0.03)	(0.04)	(0.05)
P(type I error)	0.25	0.37	-0.12
	(0.07)	(0.09)	(0.10)
P(type II error)	0.14	0.02	0.12
	(0.03)	(0.02)	(0.04)
P(type I error) – P(type II error)	0.12	0.36	-0.24
	(0.08)	(0.09)	(0.09)
P(wrongful conviction)	0.10	0.13	-0.03
	(0.03)	(0.05)	(0.40)
P(wrongful acquittal)	0.32	0.07	0.25
	(0.09)	(0.07)	(0.10)
wrongful convictions per 100 cases	7.1	10.5	-3.4
	(2.4)	(3.7)	(3.0)
wrongful acquittals per 100 cases	9.8	1.3	8.4
	(2.8)	(1.4)	(2.9)
n, ratio of wrongful acquittals to wrongful convictions	1.37	0.13	1.25
	(0.75)	(0.14)	(0.68)

From these numbers we can estimate the actual value of *n*, the ratio of the number of wrongful acquittals to the number of wrongful convictions, among the case studied. The proportion of wrongful convictions equals the product of 1 minus the prevalence of guilt times the probability of a type I error, and the proportion of wrongful acquittals equals the product of the prevalence of guilt times the probability of a type II error; the ratio of the two yields *n*. Thus, we estimate that *n* is 1.37 (1.7) for juries and 0.13 (0.20) for judges.

It is imperative to note that these statistics about accuracy are specific to the cases analyzed from the NCSC data and no generalizations from these analyses should be made to other cases, not even to other cases in the four jurisdictions. Spencer (2007, 327) cautioned that "The estimates are no basis for action other than future studies." With appropriately selected random samples of cases, however, generalizations could be justified (Section 9). There are additional limitations of the analysis (e.g., Spencer 2007, 322-323 and 327), including the question of interpretation of the latent class or "true state" in the model; these matters are discussed in Section 7.7.

6.4. Use of Three Raters

Gastwirth and Sinclair (1998, 69ff) developed a method discussed by Walter and Irwig (1988) that uses *three* raters, such as the jury, the judge, and a courtroom observer such as a retired judge, such that two raters assign cases to two classes (convict, acquit) and one rater can assign to three classes (convict, acquit, no verdict). Under the assumption

that the errors made by the three raters are mutually conditionally independent given the true state of the defendant, they show how to derive estimates of the type I and type II error rates for each rater, the conditional probability of no verdict (hung jury) given the true state, and the prevalence of guilt, i.e., the proportion of defendants whose true state is guilty. This is a powerful method, and can be further strengthened by the collection of data on strength of evidence. As with log-linear model (Section 6.3) and the Hui-Walter model (Section 6.2), there is a question about what "true state" means in this model – does it refer to material guilt or probatory guilt or something else? This question is discussed in Section 7.7.

7. Limitations of Empirical Analyses

7.1. Sources of Error

Assessments of error are themselves subject to error. Here we identify major kinds of error in past and possibly future studies and discuss associated concerns for estimates of accuracy: error from the sample, error from nonresponse to the study, misreporting to the study, imperfect questionnaires, Hawthorne effects, faulty assumptions in the statistical analysis, and invalidity of the interpretation of latent class. Identifying the sources of error is preliminary to mitigating these sources of error, as feasible, and then quantifying the effects of the remaining sources of error as in a total error analysis (Mulry and Spencer 1990). These sources of error are not unique to estimation of accuracy of verdicts, but rather are common kinds of error in statistical studies in the social sciences.

7.2. Sampling Error

If the sample of cases being analyzed is scientifically selected random sample, the statistical results may be generalized to a larger population of cases that were not selected but that had a chance of being in the study. (A caveat here arises from non-response, discussed below.) An advantage of random sampling is that we can quantify sampling error for inferences extended outside the sample itself. In the NCSC study, the exact manner of sampling was not specified in the documentation, and the extent to which statistical inferences can be extended to non-study
cases even within the four jurisdictions is unclear. What is clear, nonetheless, is that only jury trials were considered, and no inferences should be made about non-jury trials, and only trials resulting in verdicts were analyzed, so inferences should not be made about hung juries. When cases are selected with unequal rates, statistical weighting may be needed to allow such generalizations. The sampling method used in the NCSC study appears to have used clustering of cases (e.g., by judges), which calls for special methods when sampling errors (standard errors) are computed. The analysis in Spencer (2007) was not attempting to generalize beyond the cases in the study, and for that reason (and others) it used simplified techniques for estimating standard errors; such simplifications would be inappropriate, however, if statistical results were being generalized to apply beyond the study cases. The simplified methods of estimating standard error often give higher estimates of precision than the methods appropriate when the sample is being generalized, but even so some of the estimated standard errors were quite large. For example, the estimated difference in overall accuracy between the judge and the jury was 0.05, with an estimated standard error of 0.05, and the latter is so large that the apparent difference in accuracy could just reflect randomness in the study. On the other hand, the estimates of n (ratio of number of wrong acquittals to wrong convictions) were far away from Blackstone's prescription of 10, relative to the nominal standard errors. The limitations to the nominal standard errors (Section 2.2.1) mean that we cannot quantify our uncertainty about the estimates of n, except that the estimates are low enough to cause concern and warrant further study.

7.3. Non-response Error

Error from non-response in a study is an endemic problem in survey research. In the NCSC study, if a judge was asked to participate but did not, will the statistical findings be systematically different than if the judge did participate? The answer depends on whether the judge was more (or less) accurate than average, on whether the juries in the judge's courtroom perform issue more (or less) accurate verdicts than average, and on whether the prevalence of guilt is higher (or lower) than average. Non-response can also occur on a partial basis, for example a judge could participate but neglect to report the perceived strength of evidence for the prosecution. Such a case would have been excluded from the latent class analysis (Section 6.3).

7.4. Reporting Error

Reporting error, whether arising from misunderstanding of the question or other reason, can also affect the analysis. In the NCSC study, the cases were jury trials and the jury verdicts were actual verdicts, but the judges' verdicts were only the judges' reports of what verdicts they would have issued at trial. The two are not the same, as has been earlier noted for the Kalven-Zeisel and the NCSC cases (Eisenberg et al 2005, 178). If the judge was trying to guess the jury's verdict rather than try to report a correct verdict, the assumptions behind the statistical analysis may fail to some extent. The estimates of judge accuracy would then be invalid, and it is unclear what the impact would be on the estimates of jury accuracy. This is a potentially important point for future studies if they are used in any way to reward or punish judges, because then there would be incentives to misreport verdicts.

Another source of reporting error, commingled with question wording issues, arises in questions about evidentiary strength, as discussed in Section 2.2.3). For example, the questions ask about the case as a whole, but analyses will be focusing on particular counts. A further source of possible mismatch is that jurors on the same case varied in their ratings, and averages of the ratings were used to obtain a rating for the jury as a whole. For example, for 5 cases the jury reported weak evidence but convicted anyway. If those 5 cases are omitted, the estimates in Table 2 change (full data standard error estimates are in parentheses): for judges, P(correct) changes from 0.831 (0.03) to 0.853, P(type I error) changes from 0.254 (0.07) to 0.173, and P(type II error) changes from at 0.136 (0.03) to 0.139. The change for P(type I) error is about 1 standard error, which is not trivial. However, we are not only deleting 5 observations from 271 – we are reducing the number of observed cells in the 2x2x3x3 table from 25 to 21. Thus, the question about the sensitivity of sensitivity of the estimates to data misreporting of evidentiary strength remains unresolved.

Reporting error also relates to model misspecification (Sections 7.6, 7.7). The statistical analysis does not assume that the responses to the question on evidentiary strength are correct or valid. Rather, the log-linear model makes an assertion about a structural relationship involving the latent (true) state, the manifest verdicts, and the *manifest* responses rather than the *theoretically correct* responses to the question. One might expect that more "correct" responses to the question on evidentiary strength would lead to more precision in the estimates of error, but it is not clear whether or how much the current estimates are biased as a result of imperfections in the observed ratings of evidentiary strength. Misclassification errors could affect the validity of the interpretation of latent class, however. Future studies should include variants of the question, and likely more than one question.

7.5. Hawthorne Effects

There is as well the possibility of a "Hawthorne effect", namely that jurors and judges behaved differently for cases in the study because they knew they were being studied. This could be an important issue if such studies are carried out for monitoring purposes and the results of the monitoring affect the judges.

7.6. Incorrectness of Assumptions Underlying Models

The statistical estimators are based on data, as described above, but they also use certain assumptions. Although a model does not have to be perfectly correct to be useful, if the errors in assumptions are severe enough, the estimates can be way off. That is one reason that using multiple estimation models is a good thing to do when feasible. For example, the model in Section 4 yields estimates of jury accuracy of 0.84 for the NCSC cases when the differential accuracy is delta = 0.05, and this is close to the estimate of 0.83 obtained from the model in Section 6.3, which did not make assumptions about delta but rather estimated delta to equal 0.05.

Estimator (2) assumes that δ is equal to 0 (equal accuracy of judge and jury). For an example of sensitivity of the estimates to different assumptions about δ, note that if δ were really know, the estimate of accuracy of the jury (judge) would decreased (increased) by about $\delta/2$, Spencer 2007, 314, Table 2). If in fact there is negative dependence between errors by the judge and the jury, the estimates from (2) will not necessarily be "optimistic", but rather could overstate the error rate; see Spencer (2007, 314, Table 2) for numerical results in a sensitivity analysis. The nature of the dependence can be studied statistically if there are three raters, but otherwise non-numerical information about behavior may be needed to validate or invalidate the assumption. A similar point applies to the other models, which all use some form of independence assumption.

The log linear model of Spencer (2007), as discussed in Section 6, assumed a linear form for the logarithm of the probabilities. If the linearity does not hold, or if certain interactions have been omitted from the model, the estimates could be biased.

7.7. Error in Interpreting Latent Class

A more severe challenge to the estimates from the latent class models (e.g., the models in Sections 6.2-6.4) is the validity of the interpretation of latent class as representing the true state of the defendant. We want the latent class to represent the true state of the defendant in the material sense, but given that the analysis is based on agreement, and the judge and the jury are supposed to acquit if the evidence is weak, it is reasonable to suppose that the latent class represents a mixture of the material interpretation of true state and the probatory interpretation. One partial check on the interpretation is provided by the consistency of the overall error rates under the models of Sections 4.3 and 6.3, because the relatively transparent model of Section 4.3 does support the material interpretation of true state if the assumptions underlying the model are correct.

8. Benefit from Monitoring Error Rates

The kinds of data collection and statistical analysis discussed in Section 7 can allow us as a society to monitor the levels of error and the actual balance between the two types of error in verdicts. For the NCSC cases studied, and we emphasize that the numerical findings for these cases should not be extrapolated, the ratio n of erroneous acquittals to erroneous convictions, appears to be far from Blackstone's 10:1 ratio for n – the estimated rates correspond to n less than 2 for juries and n less than 0.2 for judges. Are such values of n societally acceptable? Studies could be done to better understand what society values for the ratio. Then the results of those studies could be compared to the empirical estimates of n. Perhaps the feedback of performance measures such as n can be used to correct the system if it is out of balance, and perhaps feedback on other error measures can also be used to improve performance.

9. Suggestions for Future Data Collection/Analysis

If additional empirical studies with multiple raters for verdicts were carried out, we could accomplish a number of things. We could see if the findings of Spencer (2007) for the NCSC cases apply more generally; if they do not we could learn about patterns of variation in the error rates. The error rates can be estimated for groups of cases, where the groups could be formed by area, by demographic characteristics of defendants, by type of crime, and so forth; it may be feasible to carry out statistical modeling for the groups rather than carrying out separate large studies for each group (Rao 2003).

We could also improve the survey and estimation methodology to better control and quantify the kinds of statistical error discussed in Section 7. Using three raters, not just two, would allow for more stable estimation and less sensitivity to assumptions of independence, and could allow better understanding of rates of hung juries when the true state of the defendant is guilty and not guilty. The results could also be used to check the assumptions of models using just two raters. That is important, because if studies with three raters are more expensive to conduct but they can show that the models with only two raters provide accurate estimates of accuracy, then the less expensive two-rater studies could carried out in the future. Including courtroom observers (perhaps retired judges) could allow estimation of error rates for nonjury trials. The studies could be done for civil trials as well as for criminal trials.

10. Conclusions

An upper bound on the accuracy of jury verdicts can be estimated rather simply, if the judges are also asked their views about the cases. Using more complicated statistical models, the rates of two types of errors (erroneous acquittals and erroneous convictions) can be estimated. Some additional validation of the statistical models is warranted, and various sources of inaccuracy of the estimates have been discussed.

Given that the ratio of erroneous acquittals to erroneous convictions can be measured, albeit imperfectly, society should invest the resources to make, analyze, and report the measurements. If the empirical balance is skewed away from societal values, then there likely will be some pressures to shift the balance, much the same way that information about poverty, educational achievement, and health status has led to pressures for change and improvement. One important point not to be overlooked is that measurement of a system's performance can be detrimental if the measurement process leads to changes in behavior that do not improve the underlying performance but only improve the measured performance Weisbrod (1988, 43ff). For example, in education there is concern that widespread attention to test scores has led to a distortion of educational goals towards what is measured and away from what is not or cannot be measured. That should not deter us from further measuring the performance of the court system, viz., estimating the accuracy of verdicts and the balance between erroneous convictions and erroneous acquittals.

Acknowledgements

The author is appreciative for suggestions, advice, and criticism from Ron Allen, Shari Seidman Diamond, Joseph Gastwirth, Shelby Haberman, and Jack Heinz. Responsibility for the views herein and any errors should be attributed to the author only.

References

Blackstone W (1825) *Commentaries on the Laws of England. Book the Fourth. 16 ed.* London: Strahan.

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.

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.

Laudan L (2006) *Truth, Error, and Criminal Law: An Essay in Legal Epistemology*. New York: Cambridge University Press.

Mulry MH and Spencer BD (1991) Total Error in PES Estimates of Population. *Journal of the American Statistical Association* 86, 839-855

Rao JNK (2003) *Small Area Estimation.* New York: Wiley.

Spencer BD (2007) Estimating the Accuracy of Jury Verdicts. *Journal of Empirical Legal Studies* 4, 305-329.

Volokh A (1977) n Guilty Men. *University of Pennsylvania Law Review* 146, 173-216.

Walter SD and Irwig LM (1988) Estimation of Test Error Rates, Disease Prevalence, and Relative Risk from Misclassified Data: A Review. *Journal of Clinical Epidemiology* 41, 923-937.

Weisbrod BA (1988) *The Nonprofit Economy*. Cambridge: Harvard University Press.