Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 22

Estimating Classification Consistency for Complex Assessments*

Lei Wan
Robert L. Brennan
Won-Chan Lee†

January 2007

*The authors would like to thank the National Conference of Bar Examiners for supporting this study.

†Lei Wan is Research Assistant, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: lei.wan@pearson.com). Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, CASMA, University of Iowa. Won-Chan Lee is Research Scientist, CASMA, University of Iowa.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

The purpose of this study is to investigate the performance of five procedures for estimating classification consistency for assessments containing both dichotomous and polytomous items. The procedures include a normal approximation procedure(NM), the Breyer-Lewis procedure(BL), the Livingston-Lewis procedure(LL), a bootstrap procedure(BW) and a compound multinomial procedure(CM). Both simulated and real data were used to demonstrate the behavior of these procedures. The simulation incorporated the following testing conditions: eight test lengths, three degrees of cross-format equivalence, three cut score positions, and two sets of performance categories. The real data were taken from the Multistate Bar Examination and bar essay examination. The procedures were evaluated according to how accurately they estimated classification consistency in the simulations and how well their assumptions were met for the real data.

The results showed that with the simulated data, the accuracy of the procedures varies across different testing conditions. With the real data, the assumptions of the procedures were reasonably well satisfied. In general, the NM and LL procedures yielded relatively accurate decision consistency estimates, whereas the BW and CM procedures yielded less accurate estimates. When a bias correction method was employed for the BW and CM procedures, much more accurate estimates were obtained.

1 Introduction

In many assessment contexts, it is necessary to classify examinees into non-overlapping performance categories according to a set of predetermined standards. For example, mastery/non-mastery decisions or pass/fail decisions have been an integral part of testing for decades not only in educational contexts but also in licensure and certification contexts. Also in recent years, the National Assessment of Educational Progress and No Child Left Behind Act have focused attention on classifications into multiple categories. In such circumstances, classical approaches to addressing reliability concerns (see Feldt & Brennan, 1989) may still be relevant, but almost always users want statistics that more directly address the consistency of classification decisions instead of the consistency of test scores per se.

There are two commonly used classification consistency measures. The first is the agreement index P proposed by Hambleton and Novick (1973), which is the proportion of examinees consistently classified on alternate administrations of a test.

$$P = \sum_{j=1}^{J} p_{jj}, \tag{1}$$

where p_{jj} is the proportion of examinees classified consistently into the j-th performance category on the alternate administrations, and J is the total number of performance categories. The other measure is coefficient κ (kappa) (Swaminathan, Hambleton, & Algina, 1974). The index P does not account for chance agreement, whereas coefficient kappa was proposed to account for chance agreement in classifications.

$$\kappa = \frac{P - p_c}{1 - p_c}, \tag{2}$$

where

$$p_c = \sum_{j=1}^{J} p_{j\bullet} p_{\bullet j}. \tag{3}$$

The symbols $p_{j\bullet}$ and $p_{\bullet j}$ represent the marginal proportions of examinees assigned to the j-th performance category on the first and the second administration, respectively, and the symbol p_c represents the total proportion of agreement due to chance.

Since the 1970s, procedures for estimating decision consistency based on single-administration data have been proposed. Huynh (1976) used a beta-binomial model to estimate classification consistency directly for a group of examinees. Later, Hanson and Brennan (1990) extended Huynh's procedure by using a four-parameter beta-binomial model. In contrast, Subkoviak (1976) suggested another type of procedure in which a binomial distribution was assumed for errors given true score, but no marginal distribution was assumed for

true scores. Subkoviak first estimated classification consistency for each individual examinee and then averaged the individual estimates to obtain an overall decision consistency estimate for the group.

The above-mentioned procedures for estimating classification consistency have been extensively discussed in the literature(for other less known procedures, see Berk, 1980; Traub & Rowley, 1980). However, these early procedures deal with tests consisting of dichotomous items, only. The prevalence of such procedures is due to the fact that, in the past, a standardized assessment typically included only multiple-choice items. Currently more and more test publishers are using constructed response (CR) and multiple choice (MC) items together to access the desirable features of both formats. It is not surprising, therefore, that in the recent past there have been some new procedures proposed for estimating decision consistency for assessments containing both dichotomous and polytomous items (called *complex assessments* in this report). What is somewhat surprising is that, for the most part, these new procedures have not been systematically studied and compared. Therefore, the primary purpose of this study is to examine how several procedures for estimating classification consistency for complex assessments work and compare under various testing conditions. Specifically, a normal approximation procedure (Peng & Subkoviak, 1980), the Breyer-Lewis procedure (Breyer & Lewis, 1994), the Livingston-Lewis procedure (Livingston & Lewis, 1995), a bootstrap procedure (Brennan & Wan, 2004) and a compound multinomial procedure (Lee, 2005b) are considered.

1.1 Normal Approximation Procedure (NM)

For complex assessments, a rather simple solution is to assume that observed scores from two alternate test administrations have a bivariate normal distribution. Peng and Subkoviak (1980) employed KR-21 of the actually administered test as the correlation for the bivariate distribution, whereas Woodruff and Sawyer (1989) proposed dividing a test into parallel halves so that a stepped-up reliability estimate (using the Spearman-Brown formula) could be used as the correlation.

Originally, the Peng-Subkoviak normal procedure was developed to simplify the computation for estimating decision consistency for dichotomously-scored data. However, since this procedure does not impose restrictions on data type, it is worth examining how well this procedure can work for mixed-format data. If this procedure is applicable for complex assessments as well, its simplicity could save considerable resources in practice.

1.2 Breyer-Lewis Procedure (BL)

The Breyer-Lewis (1994) procedure requires dividing a test into two comparable half-tests. A cut score needs to be given to each half, and the sum of the half-test cut scores should be equal to the cut score for the full-length test. A bivariate normal distribution is assumed underlying for the half-tests. This procedure

first estimates the tetrachoric correlation for the half-test contingency table, taking it as a reliability estimate for the half-tests. Then the procedure obtains a reliability estimate for the full-length test using the Spearman-Brown formula. This new reliability estimate is then used as the correlation coefficient for the bivariate normal distribution for the full-length test. Overall, this procedure involves a complicated and somewhat ad hoc line of reasoning. Breyer and Lewis provided detailed computational steps for the procedure.

1.3 Livingston-Lewis Procedure (LL)

The essence of this procedure is that Livingston and Lewis (1995) created a so called "*effective test length*" (denoted as $\tilde{n}$ here) to model complex data. The term refers to the number of discrete, dichotomously-scored, locally independent test items necessary to produce total scores having the same precision as the scores being actually used. The formula for effective test length suggested by the authors is

$$\tilde{n} = \frac{(\mu_x - X_{min})(X_{max} - \mu x) - r\sigma_x^2}{\sigma_x^2(1-r)},$$

where $\tilde{n}$ is rounded to the nearest integer, X_{min} is the lowest possible score, X_{max} is the highest possible score, μ_x is the mean, σ_x^2 is the variance, and r is the reliability of the test. In this procedure, X refers to the reported score, which could be either raw scores or scaled scores. Using the effective test length, the test score X can be transformed onto a new scale $X^{'}$ that extends from 0 to $\tilde{n}$ (Livingston & Lewis, 1995).

The distribution of true scores is then estimated by fitting a four-parameter beta distribution whose parameters can be estimated from the observed distribution of $X^{'}$. Also, the distribution of conditional errors is estimated by fitting a binomial model with regard to $X^{'}$ and $\tilde{n}$. With parameters for both distributions known, classification consistency is computed in the same way that Hanson and Brennan (1990) computed classification consistency. The results still need to be adjusted so that the predicted marginal category proportions match those observed for the actual test. That is, in determining decision consistency, Livingston and Lewis suggested taking the classifications on the actual test into account.

1.4 Bootstrap Procedure (BW)

The bootstrap algorithm involves taking multiple random samples with replacement from the original sample. The Brennan-Wan (2004) procedure begins by generating a bootstrap sample of the item response vector, and then applies the scoring/scaling rules used with the original data to the bootstrap sample. For each examinee, a consistent decision is made when he or she is classified into the same performance category over two test samples. This process is repeated a large number of times. The proportion of consistent decisions over all replications can be computed for each examinee, and then these individual P estimates can be averaged to obtain an overall estimate of P and kappa for the group. In

this sense, the BW procedure is a descendant of Subkoviak's procedure (1976). With complex assessments, replication repeats a stratified bootstrap sampling procedure. That is, a bootstrap sample is taken separately from each distinct section built into the design of the assessments (Brennan & Wan, 2004).

Brennan and Wan (2004) explicitly distinguished two approaches to determining classification consistency indexes. In one approach, decision consistency is determined based on any two random forms of a test. The forms can be generated through item sampling as in the bootstrap procedure or predicted from a presumed distributional model (e.g. beta-binomial) as in the Livingston-Lewis procedure. Test results on these two hypothetical forms are compared to obtain classification consistency. This approach is commonly seen in the literature. In the other approach, decision consistency is determined on the basis of the test actually administered and one random form of the test. Brennan and Wan argued that it is often reasonable to take into account test results on the actual test, because in operational settings where an examinee passes or fails a particular test, the only consistent decision that could be made is a corresponding pass or fail decision on an alternate form. Brennan and Wan focused on the second approach in designing the bootstrap algorithm, but they considered the first approach, too.

1.5 Compound Multinomial Procedure (CM)

Lee (2005b) proposed a multinomial error model for a test with undifferentiated polytomous items, and a compound multinomial error model for a test containing a mixture of item sets. The multinomial procedure reduces to Subkoviak's procedure (1976) when items are dichotomously scored. Suppose a test contains n polytomous items, each with $h\,(h > 2)$ score points, $g_1 < g_2 < \ldots < g_h$. Let $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_h\}$ denote the proportions of items in the universe such that an examinee can get scores of $g_1, g_2, \ldots, g_h$, respectively. Further suppose that $X_1, X_2, \ldots, X_h$ are the random variables representing the numbers of items scored with each of the possible h points. For each individual examinee, these random variables follow a multinomial distribution:

$$Pr(X_1 = x_1, X_2 = x_2, \ldots, X_h = x_h|\boldsymbol{\pi}) = \frac{n!}{x_1!x_2!\ldots x_h!}\pi_1^{x_1}\pi_2^{x_2}\ldots\pi_h^{x_h},$$

where $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_h\}$ can be estimated by the observed proportions of items scored with the corresponding points. Since it is possible that many different sets of values of $X_1, X_2, \ldots, X_h$ can lead to a particular total score of y, the probability density function (PDF) of Y can be obtained by summing over all sets of $X_1, X_2, \ldots, X_h$ that make the total score of y:

$$Pr(Y = y|\boldsymbol{\pi}) = \sum_{c_1x_1+c_2x_2+\cdots+c_hx_h=y} Pr(X_1 = x_1, X_2 = x_2, \ldots, X_h = x_h|\boldsymbol{\pi}).$$

When this PDF is known, it is easy to calculate the P index for each individual examinee, and then the overall P and kappa index for the group.

When a test is composed of different sets of items (e.g., items taken from fixed content categories or items having different numbers of score points), the multinomial model needs to be replaced by a compound multinomial model. An important assumption of the compound multinomial model is that conditional on proficiency, errors over the different sets of items are uncorrelated so that the joint PDF of the subtotal scores is the product of the marginal PDFs of each subtotal score. Similarly, note that there can be different sets of subtotals leading to a particular grand total score.

The CM procedure also accommodates the two approaches to determining classification consistency that Brennan and Wan (2004) considered. Moreover, the CM procedure can estimate classification consistency for scale scores.

In the previous paragraphs, an overview of five procedures for estimating decision consistency for complex assessments has been provided. In order to demonstrate the behavior of these procedures, both simulated and real data were used in this study. Next, methodologies for conducting the simulations are outlined, followed by a description of the real data analyses. Then results of the simulated and real data analyses are presented. This report concludes with a summary of the findings and a discussion of the limitations of the study.

2 Method

2.1 Simulated Data

Simulations are important in this study for two principal reasons: (1) the accuracy of the procedures can be evaluated only when prior "true" P or kappa values are available, and (2) different measurement conditions can be easily built into simulations so that we can investigate the behavior of the procedures in a variety of situations.

In this study, the three-parameter logistic IRT model (3PL) was used to generate dichotomous items, and the generalized partial credit model (GPC) was used to generate polytomous items.

Under the 3PL model (Lord, 1980), the probability that an examinee with ability θ_i answers item j correctly is defined as

$$P_{ij} = P(\theta_i|a_j, b_j, c_j) = c_j + (1 - c_j)\frac{exp[Da_j(\theta_i - b_j)]}{1 + exp[Da_j(\theta_i - b_j)]},$$

where a_j is a discrimination parameter, b_j is a difficulty parameter, c_j is a lower asymptote for the item, and D is a scaling constant (typically 1.7). Under the GPC model (Muraki, 1997), if item j has score categories $1, 2, \ldots, H$, the probability that an examinee obtains a particular score category h is

$$P_{ijh} = P(\theta_i|a_j, b_{j1}, \ldots, b_{jh}, \ldots, b_{jH}) = \frac{exp[\sum_{v=1}^{h} Da_j(\theta_i - b_{jv})]}{\sum_{m=1}^{H} exp[\sum_{v=1}^{m} Da_j(\theta_i - b_{jv})]},$$

where a_j is a discrimination parameter, b_{jh} parameters are item-category parameters, and v and m are two dummy variables representing score categories.

The first goal of data generation was to create a pool of item parameters for 10,000 dichotomous and 1,000 polytomous items scored 0–4 (i.e. five score categories). The simulated item parameters were generated on the basis of several distributional-form assumptions and real-test item parameters (see Wan, 2006, for details). The means and standard deviations of the item parameter distributions are those for the Multivariate Bar Examination and the bar essay examination, which will be described later.

Parameters a, b, and c for the 3PL model were generated such that $a \sim$ *Lognormal* (0.4, 0.2), $b \sim$ *Normal* (-0.8, 1.8), and $c \sim$ *Beta* (2.2, 8.3). In addition, for practical purposes the following restrictions were imposed: $0.1 \leq a \leq 1$ and $-4 \leq b \leq 4$. For the GPC model, the parameter a was generated such that $a \sim Lognormal(0.3, 0.1)$, and parameters b_2 through b_5 were generated from *Normal* (-1.5, 0.8), *Normal* (-1.6, 0.8), *Normal* (0.6, 1.1) and *Normal* (1.4, 1.1), respectively. Similarly, restrictions were imposed so that $0.1 \leq a \leq 1$ and $-4 \leq b_2, b_3, b_4, b_5 \leq 4$. For the ability parameters, two theta vectors (one denotes the latent trait measured by dichotomous items, and the other denotes the trait measured by polytomous items) for a sample of 1,000 simulees were generated, using a bivariate normal distribution with a mean of 0, a standard deviation of 1, and three degrees of correlations r=0.5, 0.8 and 1.0.

Given the item and ability parameters, it is easy to generate responses to the items. For each dichotomous response, the probability of a correct answer was based on the 3PL model. A random number u was drawn from a uniform distribution [0, 1]. If u was less than P_{ij}, the corresponding response was 1; otherwise it was 0. For each polytomous response, the probability of obtaining a particular score was based on the GPC model. Let $\tilde{P}_{ij(-1)} = 0$ and $\tilde{P}_{ijg} = \sum_{g=0}^{g} P_{ijg}$. If $\tilde{P}_{ij(g-1)} \leq u < \tilde{P}_{ijg}$, then the corresponding response was set to g, where $g = 0, 1, \ldots, 4$.

2.2 Real Data

The real data used in this study are from the Multistate Bar Examination (MBE) and the bar essay examination. The MBE is developed by the National Conference of Bar Examiners (NCBE) and is widely administered in the United States twice a year. The MBE is an objective six-hour examination containing 200 multiple-choice questions. The examination is divided into two periods of three hours each, one in the morning and one in the afternoon, with 100 items in each period. The two sessions are developed to be parallel to each other.

A majority of the jurisdictions in the United States also administer an essay test of legal knowledge as part of the process for determining a candidate's competence to practice law. The bar essay test is often constructed locally. Consequently, the essay tests administered in various jurisdictions are usually different in terms of the number and the content of the questions. The essay test is always administered in participating jurisdictions one day before the MBE, and it is often given in one continuous three-hour time period.

Each jurisdiction determines its own policy regarding the relative weights for the MBE and the essay test. It is common that the MBE and the essay

raw scores undergo some score transformation for licensure decisions. In order to investigate the applicability of the decision-consistency procedures for scale scores, two hypothetical types of scales were constructed in this study. They are discussed later in the subsection of "Scale Transformation".

This study collected two sets of bar examination data. Both sets are from a recent year's administration. Jurisdiction #1 used the 200-item MBE and 10 essay questions, each of which was scored 1–12. Jurisdiction #2 used the same MBE and 9 essay questions, each of which was scored 1–5. Descriptive statistics for the data sets are reported in Table 1 (*composite* and *normalized* refer to scale scores to be discussed).

2.3 Factors Investigated

For the simulations, the factors that were examined are: test length, degree of construct equivalence, and number of cut scores. For the real data analyses, two hypothetical scaling functions were considered.

2.3.1 Test Length

Eight conditions of test length were considered in the simulations: 200/10, 200/5, 100/10, 100/5, 50/5, 50/2, 25/5 and 25/2. The value before the slash denotes the number of dichotomous items, and the value after the slash denotes the number of polytomous items. As can be observed, the first two conditions have approximately the same test length as the bar examinations, and the next two conditions are shortened approximately by half, and so on.

2.3.2 Degree of Construct Equivalence

Three degrees of cross-format correlation were considered in the simulations: $r_{\theta_1\theta_2}$= 0.5, 0.8, and 1.0, where θ_1 denotes the ability measured by dichotomous items and θ_2 denotes the ability measured by polytomous items. These three degrees of construct equivalence were picked because they roughly cover the range of possible construct correlations between MC and CR items in real settings. [1]

2.3.3 Number of Cut Scores

For the simulated data, this study considers both binary classifications (J = 2) and multi-level classifications (J = 4). When there are four performance categories, the cut scores are specified as 50%, 65% and 80% of the maximum possible score, and they are applied simultaneously. When there are only two performance categories, the three cut scores are applied separately, thus the influence of the position of the cut score can be observed.

[1] It was found that for the various Advanced Placement tests, the typical disattenuated correlations between MC and CR scores were 0.56–1.0 (Lukhele, Thissen, & Wainer, 1994). For the bar examinations, the disattenuated correlation between the MBE and the essay test is 0.904 for jurisdiction # 1 and 0.902 for jurisdiction # 2.

For the real data, only one cut score (65% of the maximum score) was considered, because the licensure decision involves only pass or fail.

2.3.4 Scale Transformation

There are two common ways that raw scores can be transformed to scale scores: (1) subtest raw scores are transformed separately to subtest scale scores, and then these subtest scale scores are combined to form a composite scale score; (2) a single raw total score is transformed to a single scale score. This report considers both types of transformation in the context of the bar examinations.

The first type of transformation somewhat reflects the transformation employed by the bar examinations in the real world. It is not feasible to obtain exact information about the scaling process actually used, but general guidelines are available (Case, 2005; Klein, 1995; Ripkey, personal communication, Nov. 2005). The basic idea is that the MBE score and the essay score are respectively standardized, and then linearly transformed to the same scale with a mean of 140 and a standard deviation of 15. Then the scaled MBE and essay scores are summed to form a composite scale score. With the MBE and essay test means and standard deviations known for each jurisdiction, it is easy to construct the scaling functions as follows:

$$CSS_1 = \frac{(MBE_{raw} - 132.975)}{15.755} + 140 + \frac{(Essay_{raw} - 64.309))}{6.038} + 140, \qquad (4)$$

$$CSS_2 = \frac{(MBE_{raw} - 133.834)}{14.1} + 140 + \frac{(Essay_{raw} - 29.011)}{6.038} + 140, \qquad (5)$$

where the subscripts for CSS refer to jurisdiction #1 or jurisdiction #2. The final composite scale scores were rounded to integers.

The second type of transformation is a normalizing transformation. It is considered here because normalizing transformations are quite common in practice. The basic idea is to determine the percentile ranks of the raw scores, identify z-values in a normal distribution which have the same percentile ranks, and then linearly transform these z-values to a new scale with a certain mean and standard deviation (Kolen & Brennan, 2004). In this study, the mean was chosen to be 280 and standard deviation 28 in order to approximate the characteristics of the composite scale score.

A scatter plot of the normalizing transformation function is presented in Appendix A for each jurisdiction. Note that though normalization is a non-linear function, it does not cause dramatic non-linearity for the real data sets in this study (except perhaps in the tails of the score distributions), implying that the raw scores are approximately normally distributed. No scatter plot is provided for the composite scale scores, because there are too many different combinations of the MBE and essay scores.

The raw cut scores were directly converted to scaled cut scores for the normalized scale scores. In contrast, the raw cut scores could not be directly converted for the composite scale scores because, at least occasionally, it is likely that one raw score will convert to many different composite scale scores. In this situation, the maximum composite scale score was first computed using the maximum MBE and essay scores, and then 65% of it was set as the corresponding scale cut score. In Appendix B, score ranges as well as cut scores for the real and simulated data sets are presented. Remember that both the scaling functions and the cut scores used for the bar examinations are completely hypothetical in this study.

In summary, Table 2 displays all the factors investigated in this study. The factors are crossed, thus resulting in three ($1 \times 1 \times 3 \times 1$) conditions for each bar examination data set and 96 ($8 \times 3 \times 1 \times 4$) conditions for the simulated data.

2.4 Implementation of the Procedures

The procedures for estimating classification consistency for complex assessments have been reviewed in the previous section. This section highlights the specifics of implementing the procedures in the present study.

A reliability coefficient is crucial for implementing the normal approximation procedure. In decision consistency contexts, it is the absolute magnitude of an examinee's score that is of interest rather than the examinee's relative ranking, so it is sensible to choose a reliability coefficient related to absolute error variance $\sigma^2(\Delta)$, using the terminology and notation in generalizability theory (Brennan, 2001). In this study, multivariate Φ (phi) coefficients for raw scores were computed from the multivariate generalizability design $p^\bullet \times i^\circ$, where dichotomous items and polytomous items were differentiated as two fixed levels. For the composite scale score, the estimation of reliability also employed multivariate generalizability theory, as it is easy to compute the phi coefficient for linearly transformed scale scores (See details in Brennan, 2001, chapters 9 and 10). For the normalized scale scores, the estimation of reliability employed a compound multinomial method developed by Lee (2005a). In using this method, the conditional standard errors of measurement (CSEM) for scale scores are first computed for each person, and then an error variance is obtained by averaging the squared CSEMs over all the persons. The error variance is then divided by the observed scale score variance, and 1 minus the result is regarded as the final reliability estimate.

The implementation of the BL procedure requires splitting a test into two half-tests that are as comparable as possible. The manner in which a test is split can potentially exert considerable influence over the final results (Wan, 2006). Because a sensible split needs to be based on both content and statistical considerations, and content is not part of the simulations, the BL procedure is not part of the simulation study. The BL procedure is studied, however, in the context of the bar examinations, since the design of the MBE has a built-in split. In order to balance the effect of content differences in half tests, three random pairs of bar examination half tests were constructed for each jurisdiction. The

process for constructing each random pair was as follows: (1) MBE-am plus a random half of the essay items, and (2) MBE-pm plus the remaining essay items. It is the average P and kappa estimates over the three random pairs that are reported. The application of the BL procedure is confined to raw scores only, since the scaling functions for half-tests are not available.

The LL procedure requires four pieces of information as input: (a) a score distribution for the test actually administered, (b) a reliability coefficient, (c) the maximum and minimum possible scores, and (d) cut score(s). As explained earlier, reliability coefficients can be estimated either using generalizability theory methods (Brennan, 2001) or the compound multinomial method (Lee, 2005a). The program BB-CLASS (Brennan, 2004) was used to do the computation.

For the BW and CM procedures, dichotomous and polytomous items were differentiated as two strata. For both procedures, decision consistency was derived following the two approaches to determining classification consistency for the real data, where there truly was an "actual" test. For the simulations, however, decision consistency was examined using two hypothetical forms, only, mainly because in a simulation there is no "actual" test. A C program was written to do the bootstrap computations, and the program MULT-CLASS (Lee, 2005c) was used to do computations for the CM procedure. As a summary, Table 3 lists all the classification consistency analyses involving each procedure.

2.5 Evaluation Criteria

Examining how accurately each procedure can estimates classification consistency is one step in evaluating the adequacy of the procedures. It is impossible to obtain true P or kappa values for the real data, but a simulation permits prespecification of these true values. So in this study the simulations provide a basis for assessing the accuracy of all but the BL procedure, which is not considered in the simulations. Accuracy was evaluated through bias, standard errors (SE), and root mean square errors (RMSE). Suppose β is a parameter, then the bias, SE, and RMSE of an estimator of β are given by

$$Bias(\hat{\beta}) = \bar{\hat{\beta}}_i - \beta, \tag{6}$$

$$SE(\hat{\beta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_i - \bar{\hat{\beta}}_i)^2}, \tag{7}$$

$$RMSE(\hat{\beta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_i - \beta)^2}, \tag{8}$$

where n is the number of replications. Obviously, in order to conduct these analyses, true values of P and kappa need to be known.

As traditionally understood, a direct estimate of classification consistency needs to be derived from comparing examinees' classifications on two classically

parallel forms (Hambleton & Novick, 1973; Swaminathan, Hambleton, & Algina, 1974). However, data on classically parallel forms are rarely available in reality. It is even difficult to generate classically parallel forms through simulations. In practice, criterion-referenced tests are usually viewed as consisting of items selected as a random sample from a well-defined domain. It follows that expected values of classification consistency indexes, with the expectation taken over all possible pairs of randomly parallel forms, can be regarded as the true values. Therefore, in this study, the following steps were carried out to derive the parameter values for P and kappa: (1) draw two random test forms from the item pool; (2) generate item response data for the two forms using the item and ability parameters; (3) for each form, assign the simulees into appropriate performance categories with respect to the cut scores; (4) compute P and κ, considering results obtained in step (3); and (5) repeat the above steps 1000 times and obtain the average P and κ, which can be regarded as the parameter values (i.e., β in Equations 6-8). Since 96 conditions were created for the simulated data, there are 96 sets of true values of P and kappa.

In order to conduct the analyses of bias, SE and RMSE, repeated applications of the procedures are also necessary. Each procedure was repeated as follows: (1) draw one random form from the item pool; (2) generate item response data for the form using the item and ability parameters; (3) apply a procedure to the form and obtain one set of estimates of P and κ (i.e., $\hat{\beta}_i$ in Equations 6-8); and (4) repeat the above steps 100 times and calculate the average P and κ (i.e., $\bar{\hat{\beta}}_i$ in Equations 6-8). Note that each procedure was applied to the same 100 random forms.

Examining how the assumptions of the various procedures can be met in practice is another step in evaluating the usefulness of the procedures. Since only the bar examination raw scores are real data, the assumptions were investigated for the raw scores only.

Since the raw scores come from a single administration, there is no direct way to check the bivariate normality assumption for the NM and BL procedures. However, the parallel nature of the MBE makes it possible to construct approximately parallel half tests that are proportionally representative of the full-length test in terms of content and statistical characteristics. Thus we are able to examine the bivariate normality of the full-length test through examining the bivariate normality of these half-tests. A chi-square plot is a multivariate statistical method for judging bivariate normality. If bivariate normality holds, the plot should resemble a straight line through the origin having slope of 1 (Johnson & Wichern, 2002, p.185). For each jurisdiction, chi-square plots were constructed for the three pairs of half-tests used in the implementation of the BL procedure. In addition, Q-Q plots were constructed to examine if the marginal distribution of the raw total scores is normal. These plots depict the sample quantile versus the quantile one would expect to observe if the observations are truly normally distributed. When the points lie very nearly along a straight line, the normality assumption is tenable (Johnson & Wichern, 2002, p.180). Although there is a risk that the univariate examination will miss some features

that can be revealed only in a higher dimension, many types of non-normality are indeed often reflected in marginal distributions.

The fit of the four-parameter beta-binomial distribution in the LL procedure was evaluated by comparing plots and central moments of observed and fitted score distributions of effective test length. Also chi-square goodness-of-fit tests were carried out to facilitate the comparison by providing a formal test statistic. Finally, as an indirect way of checking the assumptions, pass rates derived from the P and kappa estimates produced by each procedure were compared to the observed pass rates in both jurisdictions. The more similar they are, the more likely it is that the assumptions are satisfied.

2.6 Research Questions

The purpose of this study is to better understand how several procedures for estimating classification consistency perform for complex assessments. This study mainly addresses the following four research questions:

1. How accurate are the selected procedures? For simulations in which true decision consistency indexes are defined a priori, what are the bias, standard errors, and root mean square errors of the estimates yielded by each procedure?

2. For the MBE and bar essay data, how well are the assumptions of each procedure justified?

3. How do various measurement factors affect decision consistency estimates in general? How do various measurement factors affect the behavior of each procedure?

4. What are the advantages and disadvantages of each procedure? What guidelines may be provided to practitioners for using these procedures?

3 Results

3.1 Simulated Data Analyses

3.1.1 True Values of P and kappa

The true values of P and kappa under the different conditions are reported in Tables 4 and 5, respectively. In the tables, r represents the degree of cross-format correlation between θ_1 and θ_2. The percentages listed in the leftmost columns represent the positions of the cut scores: 50% of the maximum possible score, and so on; "All" indicates a multi-level classification situation.

In Table 4, four findings can be observed with respect to the four manipulated factors. (1) All other things being equal, P values tend to increase as test length increases. However, this is not always true for P values generated from adjacent levels of test length. For example, in the first row of Table 4 with $r = 0.5$ and

cut score = 50%, P is 0.8473 for the 25/2 condition and 0.8311 for the 25/5 condition. Similarly, P is 0.9142 for the 50/2 condition and 0.8981 for the 50/5 condition. In both cases, the P values for the shorter tests are slightly larger than the P values for the longer tests. This pattern continues for the rest of the first row, and can be found with other rows as well. (2) All other things being equal, in general P values do not vary much across different degrees of format equivalence. (3) All other things being equal, the position of the cut score substantially influences the magnitude of P. P values are markedly larger when the 50% and 80% cut scores are used than when the 65% cut score is used; (4) All other things being equal, P values are substantially higher for binary classifications than for multi-level classifications.

The last two findings are consistent with the literature (Berk, 1980; Lee, et al., 2002; Subkoviak, 1980; Subkoviak, 1988; Wan, Lee, Brennan, & Chien, 2006) and they are relatively easy to understand. Classification consistency decreases as the cut score moves towards the middle because there is greater score density around the middle scores, and in the same neighborhood conditional measurement errors tend to peak. Thus there is a greater possibility of making classification errors. Classification consistency is higher when there are two performance categories than when there are four performance categories, because multi-level classifications allow for greater opportunities for classification errors.

However, the first two findings may seem counterintuitive. It is often claimed in the literature that classification consistency should increase with longer tests and/or higher reliability. Here, though the P index displays this tendency, P sometimes decreases as a test gets longer or as reliability gets higher. It is suggested here that the principal reason for this apparent anomaly is that while studying the effect of test length, we often ignore the fact that the relative position of the cut scores is actually changing. For example, the 50% cut score is 17 for the 25/2 condition and 23 for the 25/5 condition. In an absolute sense, these two cut scores represent the same proportion correct points for their respective tests, but in a relative sense, the positions of the cut scores are different in their respective score distributions. Since P values are influenced by both test length and the position of cut score, a longer test alone does not necessarily lead to a higher P value.

In order to demonstrate this, the percentile rank of each possible score point in the simulated data types was calculated and averaged over 1,000 replications. Results related to the 25/2 and 25/5 conditions with $r = 0.5$ are presented in Appendix C. Since two random samples of the test were drawn in each replication, there are two columns of percentile ranks in the appendix, one for each sample. The values in these two columns are very close, as they should be. The output shows that for the 25/2 condition, the first cut score (17) has a percentile rank of 16, and the second cut score (22) has a percentile rank of 55. The mean of this distribution is 22. For the 25/5 condition, the first cut score (23) has a percentile rank of 21, and the second cut score (30) has a percentile rank of 63. The mean is 28, which happens to have a percentile rank of 50. In terms of percentile ranks, the first cut score for the 25/2 condition is farther away from its mean than the first cut score for the 25/5 condition is from its

mean. Thus, although reliability for the 25/2 condition is lower, the effect of the more extreme cut score overtakes the effect of reliability and leads to a slightly higher P value. Also in terms of percentile ranks, the second cut score for the 25/2 condition is closer to its mean than the second cut score for the 25/5 condition is. Thus, the 25/5 condition has not only a higher reliability but also a more extreme cut score, which leads to a higher P value. The same trend can be observed for other situations where a longer test appears to have a lower P value. Similarly, this logic explains why sometimes slightly lower P values are associated with higher cross-format correlations. Furthermore, examining the output files in the appendix, we can find that for the simulated data sets, the 65% cut score is usually near the mean.

From Table 5, five observations about kappa are evident. (1) The kappa values are much smaller than their corresponding P values presented in Table 4. (2) The kappa values always increase as test length increases, and the increases are more substantial than the increases shown for the P values. (3) The kappa values always increase as the degree of cross-format correlation increases. (4) For binary classifications, the kappa values are obviously affected by the position of the cut score, but the direction is somewhat inverse to that shown for the P values; that is, kappa values drop as the cut score moves away from the mean score. (5) The kappa values are noticeably higher for binary classifications than for multi-level classifications.

The above findings indicate that kappa is far more sensitive to changes in test length and reliability than is P. That is, P values do not necessarily increase with test length or cross-format correlation, whereas kappa values tend to increase appreciably. There tends to be an inverse relationship between kappa and cut score because, when the cut score moves away from the mean, the overall decision consistency increases, but a large proportion of the increase is due to chance agreement. The P index does not take chance agreement into account, so it increases anyway; kappa corrects for chance agreement, so it drops instead.

3.1.2 Estimated Values of P and Kappa

For each procedure, classification consistency estimates were computed by averaging estimates over 100 random samples of each data type. Tables 6 and 7 present the average P and kappa estimates, respectively.

Table 6 shows that the four procedures (NM, LL, BW, and CM) produce roughly comparable P estimates. For the long tests (e.g. 200/5 and 200/10), the results produced by the four procedures are almost indistinguishable. So in order to investigate the differences among the procedures, the following discussion makes reference mainly to the results for the short tests.

The most striking observation is that the procedures fall into two groups: the BW and CM procedures produce nearly identical estimates, and the NM and LL procedures produce relatively close estimates. Difference between these two groups of procedures varies with the position of the cut score. As the cut score moves towards the mean, the difference often expands. For the 50% and 65% cut scores, the NM and LL procedures produce lower estimates, whereas

for the 80% cut scores, they produce slightly higher estimates. By contrast, difference between the two groups of procedures appears to remain stable across the different degrees of cross-format correlation.

As for kappa, the NM procedure usually produces the smallest estimates, the LL procedure produces somewhat larger estimates, and the BW and CM procedures yield nearly identical estimates which are much higher than those yielded by the two former procedures. The differences between the two groups of procedures are markedly larger than the corresponding differences in the P estimates. These large differences indicate that either some of the procedures are not accurate in estimating kappa or all of the procedures give inaccurate kappa estimates. The forthcoming bias analyses will help resolve the uncertainty.

3.1.3 Bias, Standard Errors, and RMSE

Bias

As mentioned earlier, substantial differences or unusual trends in classification consistency estimates often indicate problems, but without comparing such results to true values, we never know the nature of the problems. That is why Tables 8 and 9, which report the bias for the P and kappa estimates, are necessary. In the tables, the smallest bias under each condition is boldfaced. For the sake of convenience, it is the absolute values of bias that are addressed in the main body of this report.

Table 8 shows that the NM and LL procedures generally produce quite small bias for P estimates: a majority of the bias is smaller than 0.02. For the BW and CM procedures, the position of the cut score affects the magnitude of bias in P substantially. When the 50% and 80% cut scores were used, these two procedures produce quite small bias (e.g. 0.001-0.023), whereas when the 65% cut score was used, these two procedures produce remarkably larger bias—the bias can be as much as 0.07 or 0.08 for the shortest test. Moreover, the BW and CM bias is rather large for multi-level classifications. In the literature, researchers (Algina & Noe, 1978; Subkoviak, 1978) reported similar trends in the context of estimating decision consistency for dichotomous data using the Subkoviak procedure, which is the ancestor of the BW and CM procedures.

Table 9 shows that in general all the procedures yield larger bias for kappa than for P. Intuitively, this is understandable, because kappa is derived from P and p_c (chance agreement), and it is likely that errors in estimating P and p_c accumulate and cause larger errors in estimating kappa.

The NM procedure often produces rather small bias for kappa (e.g. 0.001-0.03). The LL procedure also usually produces relatively small bias for kappa, but for the 80% cut score, the LL procedure tends to substantially underestimate kappa. That happens, however, only when r = 0.5 and r = 0.8. When r = 1.0, the magnitude of bias, including the bias for the 80% cut score, drops so much that the LL bias becomes almost negligible. By contrast, the BW and CM procedures produce very large bias for kappa regardless of the cross-format correlation or the position of the cut score. For short tests, the bias can be as

large as 0.10–0.18.

So far the discussion about bias leads to the impression that the BW and CM procedures are less accurate than the NM and LL procedures. However, a bias correction method will be introduced later, and results will show that after incorporating the correction, the BW and CM procedures yield much more accurate estimates of decision consistency.

Standard Errors

Standard errors reflect the random errors that may influence the results for a procedure, as opposed to the systematic errors reflected by bias. Small standard errors indicate greater accuracy. Standard errors for P and kappa yielded by each procedure are reported in Tables 10 and 11, respectively.

From Table 10, it can be seen that the NM and LL procedures usually produce relatively similar standard errors for P. The BW and CM procedures produce nearly identical standard errors for P, which tend to be smaller than those produced by the NM and LL procedures. Furthermore, standard errors produced by all four procedures share some common tendencies. First, standard errors decrease as the tests get longer. Second, standard errors for P increase as the cut score moves away from the mean. Third, the degree of cross-format correlation does not make much difference in the magnitude of the standard errors.

The same trends seen in Table 10 are also present in Table 11. In general, the standard errors for kappa appear to be somewhat larger than their counterparts for P, and they do not drop as much as the standard errors for P when the tests get longer. The standard errors of kappa for the LL procedure are exceptionally large for the 80% cut scores. Since the LL procedure also yields particularly large bias for kappa with the same cut scores, more research needs to be conducted to explain why these 80% cut scores cause trouble for the LL procedure in estimating kappa.

RMSE

The last step in evaluating the adequacy of the procedures is to summarize the magnitude of bias and standard errors, so that users may have an overall understanding of the accuracy of the procedures. For this purpose, the root mean square error (RMSE) values for P and kappa are presented in Tables 12 and 13, respectively. Given that the trends in bias and standard errors have been identified, it is not difficult to identify the trends in the RMSE values.

First, the NM and LL procedures produce relatively similar RMSE values, whereas the BW and CM procedures produce almost exactly the same RMSE values. Second, test length has an apparent effect on RMSE values: as the tests become longer, RMSE values for P and kappa become smaller, and the differences among the various procedures tend to diminish. Third, the degree of cross-format correlation does not have much impact on the RMSE values for P or kappa. In general, the RMSE values for kappa are noticeably larger than the corresponding RMSE values for P, which is a natural consequence of the larger bias and standard errors associated with kappa.

The choice of cut score also affects the RMSE values for P and kappa. Examining the tables vertically, we can see that for the NM and LL procedures, RMSE values for P and kappa tend to be small for the 65% cut score and when multiple cut scores are applied. For the BW and CM procedures, smaller RMSE values for P appear for the 50% and 80% cut scores, whereas smaller RMSE values for kappa appear for the 65% cut score and when multiple cut scores are applied. Examining the tables horizontally, we can see that in terms of the P index, the NM and LL procedures produce slightly higher RMSE values than the BW and CM procedures when the cut score is in the tails, and markedly lower RMSE values when the cut score is near the mean or when multiple cut scores are used. Therefore, in terms of the P index, it is hard to conclude that a particular procedure is definitely superior in estimating decision consistency, since the conclusion is dependent on the choices of cut score(s). In terms of the kappa coefficient, the NM and LL procedures seem to be superior, since they consistently produce much smaller RMSE values.

3.1.4 Bias Correction Results

A possible reason for the large bias produced by the BW and CM procedures is that these two procedures employ observed proportion correct score(s) for each person as estimates of his or her true proportion correct score(s). Subkoviak (1976) realized that this might be a limitation for his procedure [2], so he suggested using Kelley's (1947) regressed score estimates, which incorporate collateral information to estimate true proportion correct scores. However, Subkoviak did not empirically show that this would work. Later Algina and Noe (1978) followed his suggestion and compared the decision consistency results based on the regressed-score estimates to the results based on the observed proportion estimates. They considered many situations, but in general they found that the regressed-score estimates did not reduce the magnitude of bias for P estimates, yet they often changed the sign of bias.

Recently, Brennan and Lee (2006) re-addressed the issue of estimating proportion correct true scores in the context of estimating decision consistency. They focused principally on the dichotomous-data case, but they also considered extensions to polytomous data and complex assessments. For dichotomous data, they found that there is considerable bias in estimates of P, p_c and kappa using both observed- and regressed-score estimates. The absolute value of the bias tends to be about the same using these two estimators of true scores, but the direction of the bias tends to be reversed. Inspired by the opposite signs of the bias, Brennan and Lee weighted the observed- and the Kelly's regressed-score estimates such that the variance of the resulting weighted estimates equals true score variance, which can be estimated if there is an estimate of reliability. They used this "optimally" weighted estimator of true score to compute decision consistency. The weights that they found are $\frac{1}{1+\sqrt{\rho^2}}$ for the regressed score and

[2] The BW and CM procedures can be viewed as generalized versions of the Subkoviak procedure for estimating classification consistency for complex assessments.

$\frac{\sqrt{\rho^2}}{1+\sqrt{\rho^2}}$ for the observed proportion score, where ρ^2 is a reliability estimate for the test. For dichotomous data, the optimally weighted estimator of true score for a person (person subscript is suppressed) can be shown as:

$$\hat{\pi}^{'} = \frac{M_x}{n} + \sqrt{\rho^2}(\frac{x}{n} - \frac{M_x}{n}), \tag{9}$$

where n is the number of items, x is the person's number correct score, and M_x is the mean score across examinees. Brennan and Lee (2006, Equations 10-14) provided a derivation of the optimal weights. They illustrated that this new estimator of true score can substantially reduce bias in decision consistency estimates using a bootstrap procedure or a binomial procedure for dichotomous data.

In this study, this method was extended to estimate decision consistency for mixed-format data using the current BW and CM procedures. In the original CM procedure, values of $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_h\}$ for polytomous items are estimated by observed proportions of items scored for each score point. With the Brennan-Lee bias correction method, however, the proportion correct true scores (for each score point) for an examinee are estimated as

$$\hat{\pi}_h^{'} = \frac{M_h}{n} + \sqrt{\rho^2}(\frac{x_h}{n} - \frac{M_h}{n}), \tag{10}$$

where x_h is the observed number of items scored with a score point h for the examinee, and M_h is the mean number of items scored with the score point h across all examinees. Using Equation 10 for polytomous items and Equation 9 for dichotomous items, the CM procedure yields corrected results for P and kappa. In this study, the overall reliability coefficient ρ^2, which is a two-strata multivariate phi coefficient (See Brennan, 2001), was used for all the different score categories.

To employ the Brennan-Lee bias correction method for the BW procedure, the ordinary bootstrap algorithm needs to be replaced by a parametric algorithm (Brennan & Wan, 2004). For dichotomous items, assuming that the estimated true proportion correct score is $\pi^{'}$ for an examinee, the parametric bootstrap involves generating a sample in the following manner:

- Draw a uniform random number u.
- If $u < \pi^{'}$, set the item response to 1.
- If $u \geq \pi^{'}$, set the item response to 0.

For polytomous items with five score points (0–4), assuming that the true proportion for each score category has been estimated as $\pi_1^{'}, \pi_2^{'}, \ldots, \pi_5^{'}$ for an examinee using Equation 10, the parametric bootstrap involves generating a sample in the following manner:

- Draw a uniform random number u.

- If u is in the interval $[0, \pi_1^{'})$, set the item response to 0.
- If u is in the interval $[\pi_1^{'}, \pi_1^{'} + \pi_2^{'})$, set the item response to 1 ...
- If u is in the interval $[\pi_1^{'} + \pi_2^{'} + \pi_3^{'} + \pi_4^{'}, 1]$, set the item response to 4.

The above steps were repeated twice for each examinee in a replication, thus creating two randomly parallel forms. From the two forms, one set of P and kappa estimates could be computed. After 1,000 replications of this process, average P and kappa estimates were regarded as the bias-corrected results produced by the BW procedure.

Since bias in P and kappa is particularly large for the short tests, the bias correction method was applied to the 25/2 and 25/5 conditions as an illustration, and the results are presented in Tables 14–21.

Tables 14 and 15 report the corrected P and kappa estimates, respectively. It is clear from the tables that the adjusted BW and CM procedures still produce very similar estimates. The corrected P and kappa estimates are usually smaller than the P and kappa estimates yielded by the original BW and CM procedures, which is good because the original P and kappa values tend to overestimate the true values.

Tables 16 and 17 report the bias in the corrected P and kappa estimates. In general, the effect of the bias correction method is overwhelming. In terms of the index P, the adjusted bias is smaller than the bias produced by the original BW and CM procedures, and the adjusted bias is also somewhat smaller than the bias yielded by the NM and LL procedures. The most effective correction takes place for the most biased P estimates, i.e. those yielded when the cut score is near the mean. In terms of kappa, the bias is reduced even more noticeably. It is not uncommon to find that bias drops by a factor of 20. Even in relation to the NM and LL procedures, which produce quite small bias in kappa, the adjusted results appear to be slightly superior.

Tables 18 and 19 report the standard errors for the adjusted P and kappa estimates. In general, the standard errors yielded by the adjusted procedures are slightly larger than their counterparts produced by the original BW and CM procedures. This is not surprising. Often there are trade-offs between systematic and random errors. Even so, the standard errors yielded by the adjusted BW and CM procedures are not inferior to those yielded by the NM and LL procedures.

Finally, Tables 20 and 21 summarize the RMSE values yielded by the adjusted BW and CM procedures. In terms of the index P, the adjusted RMSE values are markedly smaller than the original RMSE values for cut scores near the mean and when multiple cut scores are used. For cut scores near the tails, the adjusted RMSE values are slightly larger than the original RMSE values, resulting from the fact that under these situations, the bias decreases less than the standard errors increase. Moreover, the adjusted RMSE values for P are slightly smaller than those produced by the NM and LL procedures. In terms of kappa, the adjusted RMSE values are substantially smaller than the RMSE values produced by the original BW and CM procedures across all testing conditions, and

the adjusted RMSE values for kappa are comparable to those produced by the NM and LL procedures.

3.2 Real Data Analyses

3.2.1 Raw Scores

The real data are taken from the Multistate Bar Examination (MBE) and the bar essay examination from two jurisdictions. Table 22 reports the classification consistency estimates for jurisdiction #1, and Table 23 reports the estimates for jurisdiction #2. Some cells in these two tables are left blank indicating that a procedure is not applicable for a situation.

In both jurisdictions, raw scores are derived by summing the MBE and the essay test scores. Again it was found that the NM and LL procedures produce similar estimates for P and kappa, and the BW and CM procedures produce almost identical estimates of P and kappa, except for the kappa estimates for jurisdiction #2. The slight differences are likely due to the fact that there are a few non-integer essay scores in jurisdiction #2. The bootstrap program can process non-integer values, whereas the current version of MULT-CLASS can not process non-integer values. In order to implement the CM procedure, the non-integer values were rounded up, thus resulting in the differences.

The bond between the BW and CM procedures is not surprising. Though the BW and CM procedures appear to be different, they actually follow the same resampling plan when the number of bootstrap replications approaches infinity. In this study, the bootstrap process was repeated 1000 times, and the results given by the BW and CM procedures are close. The bond between the NM and the LL procedures makes some intuitive sense, too, considering that both procedures assume a score distribution for the group, and both compute decision consistency directly for the group. Though the assumptions involve different distributional models (bivariate normal vs. beta-binomial), as long as both models fit the raw data reasonably well, the results should be similar.

By contrast, the results of the BL procedure are somewhat unexpected. Since the BL procedure assumes bivariate normality and it computes decision consistency directly for the group, it is natural to expect its results to be more similar to those yielded by the NM and LL procedures. However, the BL results are quite close to those yielded by the BW and CM procedures. Further research needs to be conducted to explain this finding.

From the tables, we can also see that the LL procedure produces almost the same results using the two approaches to obtaining classification consistency (actual/hypo vs. hypo/hypo), whereas the BW and CM procedures produce higher estimates when classifications on the actual test are taken into account. The most likely reason for this is that these three procedures make use of the actual test results in different ways. In the BW and CM procedures, each person's classification status on the hypothetical form is compared with his or her classification status on the actual form. If the classifications are consistent, it is a consistent decision; otherwise, it is an inconsistent decision. This

actual/hypothetical comparison leads to higher consistency than when comparison is made between two hypothetical forms, because random errors have double the influence for two hypothetical forms. By contrast, in the LL procedure, no attention is given to a person's original classification; instead it is the predicted marginal classification rates (e.g. the marginal proportions of examinees who pass or fail) that are adjusted to the actual marginal classification rates.

3.2.2 Scale Scores

Estimates of P and kappa for scale scores are presented in Tables 22 and 23 as well. For the scale scores, the NM and LL procedures still produce similar P and kappa estimates, and the BW and CM procedures produce very similar estimates. Also, the BW and CM procedures produce higher P and kappa estimates when taking the actual test form into account, whereas the LL procedure produces almost unchanged estimates whether or not the actual test form is involved.

In terms of raw scores, classification consistency indexes for jurisdiction #1 are in general higher than the indexes for jurisdiction #2, which is expected because jurisdiction #1 has a higher reliability coefficient (0.8662 vs. 0.8529). Nevertheless, in terms of the composite scale scores (CSS), P estimates for jurisdiction #1 are lower than P estimates for jurisdiction #2, even though jurisdiction #1 still has a slightly higher reliability coefficient (0.8586 vs. 0.8575). By contrast, kappa estimates for jurisdiction #1 are higher than those for jurisdiction #2. This seems to confirm that first, P estimates are not simply determined by the magnitude of reliability. Second, kappa estimates are more sensitive to the magnitude of reliability. Third, the relationship between P and kappa is complicated in the sense that higher P values are not always associated with higher kappa values.

On the other hand, in terms of the normalized scale scores (NSS), except for a few cases in the tails where two or more raw scores correspond to the same scale score, the conversion between the raw and NSS is basically one-to-one. Furthermore, since the NSS cut score is directly converted from the raw cut score, in theory decision consistency for the NSS should remain the same as the decision consistency for the raw scores.

However, as shown in the tables, whether decision consistency remains unchanged depends on which procedure is used. The NM and LL procedures produce somewhat different P and kappa estimates, whereas the BW and CM procedures produce nearly unchanged results. The differences found with the NM and LL procedures are attributable to the reliability estimates used. In both jurisdictions, the reliability coefficients estimated for the NSS are different from those estimated for the raw scores (0.8501 vs. 0.8662 in jurisdiction #1 and 0.8339 vs. 0.8529 in jurisdiction #2). This finding suggests that users of the NM and LL procedures need to be aware of that these procedures are sensitive to the magnitude of reliability estimates.

3.2.3 Bias Correction Results

The bias correction method was also applied to the real data. With the real data, though we can not calculate the bias for P and kappa, changes in P and kappa estimates can be assumed to reflect changes in bias. Corrected decision consistency estimates are reported at the bottom of Tables 22 and 23. From the tables, it can be seen that the adjusted BW and CM procedures yield lower decision consistency results than the original BW and CM procedures, and the corrected results are closer to the results yielded by the NM and LL procedures, indicating that the bias correction method has reduced inflation in the original BW and CM estimates. It seems that the adjustments with the real data are not as dramatic as those for the simulated data. This is expected, because compared to the bias for the simulated 25/2 and 25/5 conditions, the bias for the real data, which is based on many more test items (200/10 and 200/9), should be much smaller, so there is not much room for improvement.

3.2.4 Assumption Check

Figure 1 presents the chi-square plots for the two jurisdictions. From the plots, it can be seen that in both jurisdictions, no matter how the test is split, the chi-square plot resembles a straight line through the origin having an approximate slope of 1, suggesting that bivariate normality holds for the half-tests. Figure 2 presents the Q-Q plots depicting the marginal distributions for the two jurisdictions. In both jurisdictions, the points lie very nearly along a straight line, indicating that the marginal normality assumption is tenable.

For the LL procedure, observed and fitted distributions of effective test length scores are plotted in Figure 3, one plot for each jurisdiction. Due to the small sample sizes, there are many irregularities in the observed distributions, yet roughly the observed and fitted distributions share the same shape. In addition, the raw and fitted moment statistics are presented in Table 24, as well as the likelihood ratio chi-square statistics. In both jurisdictions, the first three moments fit perfectly, which is expected because the 4-parameter beta-binomial model should fit the first three moments, and there is only slight discrepancy in the fourth moments, indicating a generally adequate fit of the LL model. The chi-square statistics appear to be significant at the 0.01 level, but given the fact that chi-square statistics are very sensitive to large sample size, this significance is not unusual. Actually, a χ^2/df ratio statistic is often used to adjust for the effect of large sample size on χ^2 statistics. The smaller this ratio is, the better a model fits. A ratio in the neighborhood of 2.5 is widely considered acceptable (Kline, 2004). It is easy to calculate that the ratio is 1.58 for jurisdiction #1 and 1.61 for jurisdiction #2, which indicates favorable model fit.

Finally, pass-rate results are presented in Table 25. Except for the NM procedure, the other procedures yielded very close estimated and observed pass rates.

Considering the above analyses, it seems reasonable to conclude that the assumptions for the procedures are basically satisfied with the raw data used in

this study.

4 Discussion and Conclusion

This section first summarizes the answers to the four research questions stated earlier. Then it provides a discussion of the limitations of the study and possibilities for future research.

4.1 Accuracy

The analyses of bias show that in general the NM and LL procedures produce fairly small bias in estimating decision consistency. The original BW and CM procedures can occasionally yield small bias for P, but when the cut score is near the mean, the BW and CM procedures noticeably overestimate P. Moreover, the original BW and CM procedures substantially overestimate kappa no matter where the cut score is. The analyses of standard errors show that all these procedures perform quite stably over item sampling. Since the standard errors yielded by the various procedures are not markedly different, the trends in RMSE values are mainly consistent with the trends in bias.

A method proposed by Brennan and Lee (2006) was used to reduce bias for the BW and CM procedures. The new estimator of true scores is a weighted composite of the observed proportion correct score and the Kelly's regressed score. Overall, the bias correction method substantially improves the accuracy of the BW and CM procedures. Though it slightly increases standard errors, it reduces bias and RMSE values for decision consistency estimates, and the effect is particularly dramatic for kappa estimates. The accuracy statistics yielded by the adjusted BW and CM procedures are not only much better than the original BW and CM statistics, but also slightly better than those yielded by the NM and LL procedures.

4.2 Assumption Check

The primary usefulness of analyzing the real data is twofold: first, it provides a chance to observe how the various procedures behave and compare under realistic circumstances; second, it provides a chance to examine the extent to which the assumptions of the various procedures are satisfied with the real data. Since the BW and CM procedures do not heavily depend on distributional-form assumptions, the main focus here is to check assumptions for the NM, BL, and LL procedures.

For the bivariate normality assumption, we examined the chi-square plots and Q-Q plots. For the four-parameter beta-binomial assumption, we examined the fitted score distributions, fitted central moments, as well as the χ^2/df ratios. None of these analyses reveal evidence to suggest serious violation of the assumptions. In another word, the analyses indicate that these assumptions are fairly realistic. But remember that the bar examination is a rather long test,

and score distributions for long tests often tend to be somewhat bell-shaped. For short tests, scores are more likely to have an irregular distributional shape, and how these procedures work for those non-normal score distributions needs to be further investigated.

The distributional-form assumptions were not checked for the simulated data. However, given that the distributional-form assumptions were reasonably well satisfied for the real data, and that the simulated data were generated based on real data item parameters, it is natural to expect that the distributional-form assumptions apply for the simulated data as well. Likely this is why the NM and LL procedures appear to be relatively accurate in estimating classification consistency in the simulations. We need to be cautious about generalizing this observation to other situations, where the distributional-form assumptions do not necessarily hold.

4.3 Impact of the Factors

Test length has a powerful influence on the estimation of classification consistency. As test length increases, decision consistency estimates increase, and the estimates tend to be more accurate. The influence of test length is associated with the fact that test length is a major element in affecting reliability—all other things being equal, the longer a test is, the more reliable the scores are. By contrast, the impact of cross-format correlation on classification consistency is minimal, though this factor can also affect reliability. The position of the cut score has a substantial influence on the estimation of classification consistency, but the impact is complicated, depending on the statistic estimated and the procedure used. In short, a cut score near the mean leads to lower P estimates but higher kappa estimates. For the BW and CM procedures, a cut score near the mean also leads to substantially less accurate P estimates. Finally, P and kappa estimates are much higher when there are two performance categories than when there are four categories. However, considering RMSE values (for the BW and CM procedures, consider the corrected RMSE values), the P and kappa estimates in this study are always more accurate for multi-level classifications than for binary classifications.

In the above paragraphs, we have analyzed the impact of the factors individually. In practice, however, these factors are often intertwined, and their effects on P or kappa estimates may cancel out each other, thus making it hard to predict the direction of change in P or kappa estimates.

4.4 Advantages and Disadvantages

The previous paragraphs have basically considered how each of the procedures would perform under various testing conditions. In this section, attention is focused on comparing the advantages and disadvantages of each procedure. The objective of the following discussion is not to throw any procedure out of our tool box but to highlight the possible gains and problems of using a procedure.

The NM procedure is easy to implement if a reliability estimate is available. It has a straightforward assumption, and the application of it does not require every examinee's item response vector but only the total scores. When the normality assumption is reasonably well satisfied, the NM procedure tends to produce fairly accurate P and kappa estimates. One problem with the NM procedure is that sometimes an appropriate reliability coefficient is not readily available, in particular for scale scores, and different reliability estimates will lead to different NM estimates.

The LL procedure has many of the same advantages and disadvantages as the NM procedure, although the LL procedure is more flexible in that the beta-binomial assumption can fit a much more diverse set of distributions than the normality assumption does. The LL procedure is easy to implement if a reliability coefficient is available. When the beta-binomial assumption is reasonably well satisfied, the LL procedure tends to produce fairly accurate classification consistency estimates. However, the LL procedure, as well as the NM procedure, is very sensitive to reliability estimates. Another problem with the LL procedure is that though it takes the actual test results into consideration, it makes adjustments according to the marginal distribution of the actual test, only, and it can not take account of each examinee's original classification status.

The BL procedure has not been studied much in this study or elsewhere, but its advantages and disadvantages can still be discussed. The advantage of this procedure is that by splitting a test into approximately parallel halves, a reliability coefficient can be conveniently obtained. However, this procedure has many more disadvantages than advantages. The biggest problem is that it is not easy to split a test into approximately parallel halves, and different ways to split the test cause different decision consistency results. Besides, this procedure can not be used for multi-level classifications or scale scores. Furthermore, though the accuracy of the BL procedure has not been investigated in the simulations, this procedure is likely to overestimate classification consistency, since its results are closer to the results yielded by the original BW and CM procedures, which always give overestimated results.

The biggest advantage of the BW procedure is that it has no distributional-form assumption, so this procedure has great flexibility. It is applicable to many types of complex tests, and it can deal with different types of scale scores with ease. Two favorable features can be programmed easily into the procedure. First, both approaches to determining classification consistency (hypo-hypo vs. hypo-actual) can be accommodated. Second, classification consistency for each examinee can be computed if necessary. The results indicate that the BW procedure is stable in producing P or kappa estimates, but in general its estimates of P and kappa tend to be biased. Fortunately, a bias correction method (Brennan & Lee, 2006) can be incorporated into the BW procedure, and the adjusted BW procedure dramatically improves the estimation accuracy. Two potential problems with the BW procedure are that it will break down when there is only one item in a content or format category, and item response vectors for all examinees must be available.

The CM procedure is based on the same sampling plan as the BW procedure

when the number of replications in the latter approaches infinity. Actually, as shown in this study, with only 1,000 replications, the results yielded by the two procedures are nearly identical, as they should be. Therefore, the CM procedure shares almost the same advantages and disadvantages as the BW procedure. The bias correction strategy can be employed also for the CM procedure. A disadvantage peculiar to the CM procedure is that the current version of the program MULT-CLASS requires a large amount of computing time to process large data sets. For example, it took the program 14-16 hours to compute classification consistency estimates for 100 simulated samples for the 200/10 condition, whereas it took the bootstrap program only about 30-40 minutes to finish the same amount of work. The performance of the CM procedure may be enhanced by using faster computers or more efficient programming routines, but the computing algorithm for the CM procedure is indeed more complicated than the algorithm for the BW procedure.

In summary, five procedures for estimating classification consistency for complex assessments are discussed in this study. The NM procedure seems promising from a practical point of view, but more effort should be taken to verify its robustness in non-normal situations. The LL procedure is mathematically sophisticated, but its adequacy also needs to be verified for more extreme score distributions. The BL procedure can be very useful in a situation where two parallel half-sections have been built into a test and where only binary classifications are of interest, but other than that, it is not likely to be very useful. The BW and CM procedures are two important procedures: they have great flexibility in dealing with complex data, and with the bias correction method, they can produce more accurate results than the other procedures—at least for the situations characterized by the simulated data in this study.

4.5 Limitations and Future Research Possibilities

This study has several limitations. First, only limited types of score distributions were studied. It is obvious that the data from the two jurisdictions have distributions that are not too different from normal distributions. Because the simulated item parameters were generated on the basis of the real test item parameters, the simulated data also have approximately normal distributions. This limitation restricts our ability to generalize the findings here to other situations, where score distributions can be severely skewed, for instance. Presumably the shape of score distributions will have greater influence over the NM and LL procedures than over the BW and CM procedures, since the former two procedures make direct distributional-form assumptions. It is desirable that future researchers employ a wider variety of score distributions and investigate the adequacy of these procedures under non-normal situations.

Second, in the simulations, various testing conditions were incorporated so that their effects on estimating classification consistency could be investigated. Most of the results in relation to the testing conditions are clear and understandable, but when these conditions interact with each other and with the procedures, explanations can become complicated and even obscure. For ex-

ample, there seems to be a strange interaction between the LL procedure and the 80% cut score in estimating kappa. For other cut scores, the LL and NM procedures tend to produce similar results, whereas for the 80% cut score, the LL procedure always yields much less accurate results than the NM procedure. Hopefully, further research will help solve this mystery.

Finally, though the primary purpose of this study is to examine the adequacy of several procedures for estimating classification consistency for complex assessments, it would be informative to investigate the performance of these procedures for purely dichotomous-item cases and purely polytomous-item cases. It will be particularly revealing to compare the Livingston-Lewis procedure with the Hanson-Brennan procedure (Hanson & Brennan, 1990) in dichotomous-data cases. In this situation, these two procedures are almost the same, except that the LL procedure employs an effective test length formula. Thus, it provides a good chance to check the reasonableness of the LL procedure. If the LL and Hanson-Brennan procedures produce very similar decision consistency estimates, more confidence can be put on the effective test length formula. Otherwise, the justification of the formula may be suspect. Studying the purely polytomous-data cases would be useful, because in this report only conditions which have many more dichotomous items than polytomous items were examined. Another way to address the issue is to increase the proportion of polytomous items in the simulated data conditions so that the influence of polytomous items will become more prominent.

Table 1: Descriptive Statistics for Raw and Scale Scores

	Mean	S.D.	Skew.	Kurt.	Min.	Max.
Juris #1						
Raw	197.285	26.286	-0.116	2.666	113	264
Composite	280.001	27.815	-0.117	2.658	191	351
Normalized	280.014	28.018	0.013	3.138	182	387
Juris #2						
Raw	162.845	18.852	-0.290	2.716	94	206
Composite	279.994	27.701	-0.333	2.627	183	339
Normalized	279.937	28.100	-0.006	3.151	173	376

Table 2: Factors Investigated

Factors	Real data (2 jurisditions)	Simulated data
Test Length	200/10 and 200/9 (fixed)	25/2, 25/5, 50/2, 50/5, 100/5, 100/10, 200/5, 200/10
Construct Equivalence	0.904 and 0.902 (fixed)	0.5, 0.8, 1
Score Scale	raw, CSS, NSS	raw
Classification Categories	J=2(65%)	J=2(50%), J=2(65%), J=2(80%) J=4(50%, 65%, and 80% together)

Table 3: Analyses Performed Using Each Procedure

Procedure	MBE/Essay Raw Scores	MBE/Essay Scale Scores	Simulated Raw Scores
NM	√	√	√
BL	√	—	—
LL	√ (actual-hypo/hypo-hypo)	√ (actual-hypo/hypo-hypo)	√ (hypo-hypo)
BW	√ (actual-hypo/hypo-hypo)	√ (actual-hypo/hypo-hypo)	√ (hypo-hypo)
CM	√ (actual-hypo/hypo-hypo)	√ (actual-hypo/hypo-hypo)	√ (hypo-hypo)

Table 4: True P Values for the Simulated Data

	25/2	25/5	50/2	50/5	100/5	100/10	200/5	200/10
$r = 0.5$								
50%	0.8473	0.8311	0.9142	0.8981	0.9374	0.9288	0.9569	0.9551
65%	0.6957	0.7443	0.7520	0.7709	0.8150	0.8299	0.8549	0.8618
80%	0.8725	0.9004	0.8982	0.9215	0.9303	0.9507	0.9441	0.9560
All	0.4773	0.5222	0.5767	0.6013	0.6842	0.7103	0.7558	0.7729
$r = 0.8$								
50%	0.8315	0.8529	0.9055	0.8864	0.9322	0.9225	0.9561	0.9510
65%	0.6964	0.7664	0.7658	0.7877	0.8248	0.8416	0.8641	0.8719
80%	0.8674	0.8881	0.8950	0.9101	0.9283	0.9412	0.9447	0.9529
All	0.4825	0.5328	0.5785	0.5955	0.6870	0.7061	0.7649	0.7758
$r = 1.0$								
50%	0.8414	0.8492	0.9025	0.8915	0.9291	0.9197	0.9535	0.9500
65%	0.7475	0.7838	0.7810	0.8068	0.8442	0.8627	0.8805	0.8870
80%	0.8558	0.8862	0.8843	0.8970	0.9152	0.9323	0.9363	0.9446
All	0.4919	0.5427	0.5837	0.6018	0.6872	0.7154	0.7695	0.7816

Table 5: True Kappa Values for the Simulated Data

	25/2	25/5	50/2	50/5	100/5	100/10	200/5	200/10
$r = 0.5$								
50%	0.3491	0.4675	0.3827	0.4682	0.5413	0.6027	0.6091	0.6539
65%	0.3970	0.4655	0.4887	0.5393	0.6169	0.6509	0.7084	0.7255
80%	0.2821	0.3390	0.3802	0.4183	0.4974	0.5211	0.6091	0.6138
All	0.2281	0.3018	0.3366	0.3810	0.4856	0.5316	0.6029	0.6311
$r = 0.8$								
50%	0.3962	0.4920	0.4371	0.5020	0.5728	0.6323	0.6646	0.6906
65%	0.4445	0.5146	0.5248	0.5684	0.6518	0.6882	0.7262	0.7446
80%	0.3253	0.3823	0.4128	0.4633	0.5432	0.5766	0.6416	0.6500
All	0.2523	0.3330	0.3562	0.3977	0.5121	0.5576	0.6264	0.6458
$r = 1.0$								
50%	0.4198	0.5312	0.4987	0.5720	0.6294	0.6703	0.6949	0.7257
65%	0.4817	0.5555	0.5699	0.6244	0.6942	0.7163	0.7571	0.7725
80%	0.3406	0.4184	0.4194	0.4715	0.5524	0.5807	0.6362	0.6493
All	0.2797	0.3613	0.3787	0.4381	0.5400	0.5702	0.6415	0.6615

Table 6: Estimated P Yielded by Each Procedure

	$r = 0.5$				$r = 0.8$				$r = 1.0$			
	NM	LL	BW	CM	NM	LL	BW	CM	NM	LL	BW	CM
25/2												
50%	0.837	0.837	0.860	0.86	0.828	0.829	0.855	0.854	0.830	0.836	0.860	0.860
65%	0.707	0.706	0.768	0.768	0.722	0.722	0.776	0.776	0.739	0.738	0.790	0.790
80%	0.893	0.883	0.871	0.871	0.894	0.883	0.875	0.875	0.887	0.871	0.868	0.868
All	0.487	0.482	0.536	0.536	0.494	0.489	0.542	0.543	0.504	0.498	0.553	0.553
25/5												
50%	0.818	0.822	0.848	0.848	0.827	0.831	0.854	0.854	0.829	0.836	0.858	0.858
65%	0.766	0.758	0.796	0.796	0.777	0.771	0.806	0.806	0.793	0.788	0.819	0.819
80%	0.918	0.910	0.899	0.899	0.908	0.895	0.893	0.893	0.906	0.890	0.889	0.889
All	0.536	0.527	0.571	0.571	0.544	0.534	0.580	0.580	0.559	0.549	0.592	0.593
50/2												
50%	0.915	0.909	0.911	0.911	0.901	0.897	0.902	0.902	0.902	0.898	0.903	0.903
65%	0.752	0.760	0.793	0.793	0.762	0.769	0.798	0.798	0.776	0.784	0.813	0.813
80%	0.907	0.897	0.891	0.891	0.912	0.900	0.897	0.897	0.902	0.887	0.886	0.886
All	0.587	0.580	0.606	0.606	0.588	0.581	0.608	0.608	0.593	0.584	0.613	0.613
50/5												
50%	0.897	0.893	0.900	0.900	0.887	0.884	0.893	0.893	0.888	0.888	0.896	0.896
65%	0.775	0.777	0.803	0.803	0.792	0.795	0.817	0.817	0.807	0.810	0.831	0.831
80%	0.925	0.917	0.912	0.912	0.921	0.910	0.910	0.910	0.914	0.900	0.900	0.900
All	0.607	0.598	0.623	0.623	0.610	0.600	0.628	0.628	0.618	0.608	0.635	0.635
100/5												
50%	0.938	0.932	0.933	0.933	0.930	0.925	0.928	0.928	0.929	0.923	0.927	0.927
65%	0.812	0.819	0.831	0.831	0.824	0.831	0.840	0.840	0.836	0.843	0.855	0.855
80%	0.937	0.931	0.930	0.930	0.935	0.925	0.928	0.928	0.929	0.919	0.919	0.919
All	0.689	0.683	0.694	0.694	0.691	0.683	0.697	0.697	0.696	0.688	0.702	0.702
100/10												
50%	0.928	0.924	0.927	0.927	0.922	0.919	0.922	0.922	0.920	0.918	0.921	0.921
65%	0.829	0.833	0.842	0.842	0.843	0.847	0.854	0.854	0.855	0.859	0.869	0.869
80%	0.951	0.948	0.946	0.946	0.947	0.939	0.941	0.941	0.942	0.934	0.932	0.932
All	0.709	0.705	0.716	0.716	0.712	0.706	0.717	0.717	0.718	0.712	0.723	0.723
200/5												
50%	0.964	0.959	0.957	0.957	0.958	0.954	0.954	0.954	0.955	0.949	0.950	0.950
65%	0.858	0.865	0.868	0.868	0.865	0.872	0.874	0.874	0.872	0.878	0.884	0.884
80%	0.948	0.941	0.943	0.943	0.947	0.937	0.943	0.943	0.946	0.939	0.937	0.937
All	0.770	0.764	0.768	0.768	0.770	0.763	0.771	0.771	0.772	0.766	0.771	0.771
200/10												
50%	0.958	0.954	0.954	0.954	0.953	0.949	0.950	0.950	0.950	0.946	0.948	0.948
65%	0.862	0.867	0.870	0.870	0.871	0.878	0.879	0.879	0.881	0.886	0.892	0.892
80%	0.957	0.953	0.954	0.954	0.954	0.946	0.950	0.950	0.950	0.945	0.942	0.942
All	0.777	0.774	0.778	0.778	0.778	0.772	0.778	0.778	0.782	0.777	0.782	0.782

Table 7: Estimated Kappa Yielded by Each Procedure

	$r = 0.5$				$r = 0.8$				$r = 1.0$			
	NM	LL	BW	CM	NM	LL	BW	CM	NM	LL	BW	CM
25/2												
50%	0.318	0.368	0.472	0.472	0.363	0.403	0.495	0.495	0.407	0.457	0.536	0.536
65%	0.379	0.382	0.521	0.521	0.415	0.419	0.542	0.542	0.453	0.457	0.570	0.570
80%	0.281	0.176	0.460	0.459	0.320	0.227	0.487	0.486	0.369	0.247	0.496	0.496
All	0.227	0.224	0.341	0.341	0.252	0.249	0.358	0.358	0.281	0.276	0.379	0.379
25/5												
50%	0.435	0.470	0.546	0.546	0.478	0.508	0.571	0.571	0.519	0.551	0.607	0.607
65%	0.462	0.460	0.566	0.566	0.504	0.506	0.595	0.595	0.539	0.542	0.621	0.621
80%	0.361	0.249	0.480	0.480	0.419	0.325	0.524	0.524	0.464	0.365	0.537	0.537
All	0.305	0.302	0.393	0.392	0.335	0.330	0.415	0.414	0.368	0.361	0.437	0.437
50/2												
50%	0.386	0.428	0.496	0.496	0.423	0.466	0.521	0.521	0.463	0.508	0.554	0.554
65%	0.493	0.507	0.574	0.574	0.511	0.526	0.586	0.586	0.544	0.559	0.617	0.617
80%	0.397	0.284	0.501	0.501	0.415	0.303	0.514	0.514	0.464	0.356	0.528	0.529
All	0.345	0.336	0.410	0.410	0.358	0.351	0.421	0.421	0.383	0.373	0.439	0.439
50/5												
50%	0.453	0.494	0.547	0.547	0.507	0.538	0.577	0.578	0.549	0.583	0.618	0.618
65%	0.530	0.539	0.597	0.598	0.566	0.577	0.627	0.627	0.600	0.610	0.656	0.656
80%	0.426	0.296	0.508	0.508	0.477	0.385	0.552	0.552	0.527	0.429	0.563	0.563
All	0.382	0.376	0.439	0.439	0.409	0.403	0.462	0.462	0.439	0.430	0.483	0.483
100/5												
50%	0.522	0.548	0.585	0.586	0.566	0.595	0.617	0.617	0.601	0.629	0.648	0.648
65%	0.619	0.632	0.655	0.655	0.644	0.658	0.676	0.676	0.668	0.681	0.705	0.705
80%	0.526	0.423	0.569	0.569	0.562	0.463	0.599	0.599	0.602	0.527	0.609	0.609
All	0.497	0.490	0.525	0.525	0.516	0.508	0.542	0.542	0.537	0.529	0.560	0.560
100/10												
50%	0.578	0.607	0.636	0.636	0.624	0.649	0.659	0.660	0.661	0.682	0.699	0.699
65%	0.650	0.661	0.680	0.680	0.679	0.690	0.705	0.706	0.705	0.715	0.735	0.735
80%	0.550	0.403	0.575	0.575	0.598	0.495	0.616	0.616	0.640	0.569	0.629	0.630
All	0.534	0.531	0.560	0.560	0.557	0.551	0.577	0.577	0.579	0.574	0.598	0.598
200/5												
50%	0.614	0.628	0.642	0.643	0.645	0.666	0.677	0.677	0.672	0.691	0.699	0.699
65%	0.711	0.723	0.727	0.727	0.726	0.739	0.743	0.743	0.741	0.751	0.765	7645
80%	0.641	0.569	0.651	0.651	0.662	0.585	0.673	0.673	0.685	0.632	0.664	0.664
All	0.621	0.613	0.626	0.626	0.632	0.623	0.641	0.642	0.644	0.636	0.649	0.649
200/10												
50%	0.638	0.653	0.673	0.673	0.674	0.695	0.710	0.701	0.707	0.724	0.733	0.733
65%	0.721	0.732	0.737	0.737	0.740	0.753	0.755	0.755	0.761	0.770	0.782	0.782
80%	0.643	0.550	0.647	0.647	0.675	0.585	0.676	0.676	0.708	0.657	0.680	0.680
All	0.633	0.628	0.643	0.643	0.648	0.640	0.655	0.655	0.665	0.659	0.671	0.671

Table 8: Bias for P Yielded by Each Procedure

	$r = 0.5$				$r = 0.8$				$r = 1.0$			
	NM	LL	BW	CM	NM	LL	BW	CM	NM	LL	BW	CM
25/2												
50%	-0.011	**-0.010**	0.013	0.013	**-0.003**	**-0.003**	0.023	0.023	-0.012	**-0.006**	0.019	0.018
65%	0.011	**0.010**	0.072	0.072	0.026	**0.025**	0.080	0.080	-0.008	-0.009	0.042	0.042
80%	0.021	0.011	**-0.002**	**-0.002**	0.027	0.016	**0.007**	0.008	0.032	0.015	**0.012**	**0.012**
All	0.010	**0.005**	0.059	0.059	0.012	**0.006**	0.060	0.060	0.012	**0.006**	0.061	0.061
25/5												
50%	-0.013	**-0.009**	0.017	0.017	-0.026	-0.022	**0.001**	**0.001**	-0.020	-0.013	**0.009**	**0.009**
65%	0.022	**0.014**	0.052	0.052	0.010	**0.005**	0.040	0.040	0.009	**0.004**	0.035	0.036
80%	0.017	0.009	**-0.001**	**-0.001**	0.020	0.007	**0.005**	**0.005**	0.019	0.004	**0.003**	**0.003**
All	0.014	**0.005**	0.049	0.049	0.011	**0.001**	0.047	0.047	0.016	**0.006**	0.050	0.050
50/2												
50%	**0.001**	-0.006	-0.003	-0.003	**-0.004**	-0.009	**-0.004**	**-0.004**	**-0.001**	-0.005	**0.001**	**0.001**
65%	**0.000**	0.008	0.041	0.041	-0.004	**0.003**	0.032	0.033	-0.005	**0.003**	0.032	0.032
80%	0.009	**-0.001**	-0.007	-0.007	0.017	0.005	**0.002**	**0.002**	0.018	0.003	**0.001**	**0.001**
All	0.010	**0.003**	0.030	0.030	0.009	**0.002**	0.030	0.029	0.009	**0.000**	0.029	0.029
50/5												
50%	**-0.001**	-0.005	0.002	0.002	**0.001**	-0.002	0.006	0.006	**-0.003**	-0.004	0.004	0.004
65%	**0.004**	0.006	0.032	0.032	**0.005**	0.007	0.029	0.030	**0.000**	0.003	0.024	0.024
80%	**0.004**	**-0.004**	-0.010	-0.010	0.011	**-0.001**	**-0.001**	**-0.001**	0.017	**0.003**	**0.003**	**0.003**
All	0.005	**-0.003**	0.022	0.022	0.015	**0.005**	0.033	0.033	0.017	**0.006**	0.033	0.033
100/5												
50%	**0.001**	-0.005	-0.005	-0.005	**-0.002**	-0.007	-0.004	-0.004	**0.000**	-0.004	-0.002	-0.002
65%	**-0.003**	0.004	0.016	0.016	**0.000**	0.006	0.015	0.016	-0.008	**-0.002**	0.011	0.011
80%	0.007	**0.000**	-0.001	-0.001	0.007	-0.004	**-0.001**	**-0.001**	0.014	0.004	**0.003**	0.004
All	0.005	**-0.001**	0.010	0.010	**0.004**	**-0.004**	0.010	0.010	0.009	**0.001**	0.015	0.015
100/10												
50%	**-0.001**	-0.005	-0.002	-0.002	**0.000**	-0.004	-0.001	-0.001	**0.000**	-0.002	0.002	0.002
65%	**-0.001**	0.003	0.012	0.012	**0.001**	0.005	0.013	0.013	-0.008	**-0.004**	0.006	0.006
80%	**0.000**	-0.003	-0.005	-0.005	0.005	-0.002	**-0.001**	**-0.001**	0.009	0.001	**0.000**	**0.000**
All	**-0.002**	-0.005	0.005	0.005	0.006	**0.000**	0.011	0.011	**0.002**	-0.004	0.008	0.008
200/5												
50%	0.007	0.002	**0.000**	**0.000**	0.002	-0.002	-0.002	-0.002	**0.001**	-0.004	-0.004	-0.004
65%	**0.003**	0.010	0.013	0.013	**0.001**	0.008	0.010	0.010	-0.009	**-0.003**	0.004	0.004
80%	0.003	-0.003	**-0.001**	**-0.001**	**0.002**	-0.007	**-0.002**	**-0.002**	0.009	0.002	**0.001**	**0.001**
All	0.014	**0.008**	0.012	0.012	0.005	**-0.002**	0.006	0.006	**0.002**	-0.004	**0.002**	**0.002**
200/10												
50%	0.003	-0.002	**-0.001**	**-0.001**	0.002	-0.002	**-0.001**	**-0.001**	**0.000**	-0.004	-0.002	-0.002
65%	**0.000**	0.006	0.008	0.008	**-0.001**	0.006	0.007	0.007	-0.006	**-0.001**	0.005	0.005
80%	**0.001**	-0.003	-0.002	-0.002	**0.001**	-0.007	-0.003	-0.003	0.006	**0.000**	-0.002	-0.002
All	0.004	**0.001**	0.005	0.005	0.003	-0.004	**0.002**	**0.002**	**0.000**	-0.005	**0.000**	**0.000**

Table 9: Bias for Kappa Yielded by Each Procedure

	$r = 0.5$				$r = 0.8$				$r = 1.0$			
	NM	LL	BW	CM	NM	LL	BW	CM	NM	LL	BW	CM
25/2												
50%	-0.032	**0.019**	0.123	0.123	-0.033	**0.007**	0.099	0.099	-0.013	**-0.006**	0.116	0.116
65%	-0.018	**-0.015**	0.124	0.124	-0.030	**-0.025**	0.098	0.098	-0.029	**-0.009**	0.088	0.088
80%	**-0.001**	-0.106	0.178	0.177	**-0.006**	-0.098	0.162	0.161	0.028	**0.015**	0.156	0.155
All	**-0.001**	-0.004	0.113	0.113	**0.000**	-0.004	0.106	0.105	**0.001**	0.006	0.100	0.099
25/5												
50%	-0.032	**0.002**	0.078	0.078	**-0.014**	0.016	0.079	0.079	**-0.012**	-0.013	0.076	0.075
65%	**-0.004**	-0.005	0.101	0.100	-0.011	**-0.009**	0.080	0.080	-0.017	**0.004**	0.066	0.066
80%	**0.022**	-0.090	0.141	0.141	**0.036**	-0.057	0.142	0.142	0.046	**0.004**	0.118	0.119
All	0.003	**0.000**	0.091	0.090	**0.002**	-0.003	0.082	0.081	**0.006**	**0.006**	0.076	0.076
50/2												
50%	**0.003**	0.045	0.113	0.113	**-0.014**	0.028	0.084	0.084	-0.036	**-0.005**	0.055	0.055
65%	**0.004**	0.019	0.085	0.085	-0.014	**0.001**	0.061	0.061	-0.026	**0.003**	0.047	0.047
80%	**0.016**	-0.096	0.121	0.121	**0.002**	-0.110	0.102	0.101	0.045	**0.003**	0.109	0.109
All	0.008	**-0.001**	0.074	0.073	**0.002**	-0.005	0.065	0.065	0.005	**0.000**	0.061	0.060
50/5												
50%	**-0.016**	0.026	0.078	0.078	**0.005**	0.036	0.075	0.076	-0.023	**-0.004**	0.046	0.046
65%	-0.009	**-0.001**	0.058	0.058	**-0.002**	0.008	0.058	0.058	-0.025	**0.003**	0.032	0.032
80%	**0.008**	-0.123	0.089	0.090	**0.014**	-0.079	0.089	0.089	0.055	**0.003**	0.091	0.092
All	**0.001**	-0.005	0.058	0.058	0.012	**0.006**	0.064	0.064	**0.001**	0.006	0.045	0.045
100/5												
50%	-0.019	**0.007**	0.044	0.044	**-0.007**	0.022	0.044	0.044	-0.028	**-0.004**	0.018	0.019
65%	**0.002**	0.015	0.038	0.039	-0.008	**0.006**	0.024	.024	-0.026	**-0.002**	0.011	0.011
80%	**0.028**	-0.075	0.071	0.072	**0.019**	-0.081	0.056	0.056	0.050	**0.004**	0.057	0.057
All	0.012	**0.005**	0.039	0.039	**0.004**	-0.005	0.030	0.030	-0.003	**0.001**	0.020	0.020
100/10												
50%	-0.025	**0.004**	0.033	0.033	**-0.008**	0.016	0.027	0.027	-0.010	**-0.002**	0.028	0.029
65%	**-0.001**	0.010	0.029	0.029	-0.009	**0.002**	0.017	0.017	-0.012	**-0.004**	0.019	0.019
80%	**0.029**	-0.118	0.054	0.054	**0.021**	-0.082	0.039	0.040	0.059	**0.001**	0.048	0.049
All	0.002	**-0.001**	0.028	0.028	**-0.001**	-0.006	0.019	0.019	0.009	**-0.004**	0.027	0.027
200/5												
50%	**0.005**	0.019	0.033	0.033	-0.020	**0.002**	0.012	0.012	-0.023	**-0.004**	**0.004**	0.005
65%	**0.002**	0.015	0.019	0.019	**-0.001**	0.013	0.016	0.016	-0.017	**-0.003**	0.007	0.008
80%	**0.032**	-0.040	0.042	0.042	**0.021**	-0.057	0.032	0.032	0.049	**0.002**	0.028	0.028
All	0.018	**0.010**	0.023	0.023	0.005	**-0.004**	0.015	0.015	**0.002**	-0.004	0.008	0.008
200/10												
50%	-0.016	**-0.001**	0.019	0.019	-0.017	**0.005**	0.010	0.010	-0.019	**-0.004**	0.007	0.007
65%	**-0.005**	0.006	0.012	0.012	**-0.004**	0.008	0.010	0.010	-0.012	**-0.001**	0.010	0.010
80%	**0.029**	-0.064	0.033	0.033	**0.025**	-0.065	0.026	0.026	0.059	**0.000**	0.031	0.031
All	**0.002**	-0.003	0.012	0.012	**0.002**	-0.005	0.009	0.009	**0.004**	-0.005	0.010	0.010

Table 10: Standard Errors for P Yielded by Each Procedure

	$r = 0.5$				$r = 0.8$				$r = 1.0$			
	NM	LL	BW	CM	NM	LL	BW	CM	NM	LL	BW	CM
25/2												
50%	0.050	0.047	0.033	**0.032**	0.041	0.040	**0.028**	**0.028**	0.038	0.035	**0.025**	**0.025**
65%	0.021	0.022	**0.013**	**0.013**	0.020	0.022	**0.013**	**0.013**	0.018	0.019	**0.014**	**0.014**
80%	0.037	0.049	**0.032**	**0.032**	0.030	0.041	**0.026**	**0.026**	0.029	0.041	**0.025**	**0.025**
All	**0.011**	0.012	**0.011**	**0.011**	0.012	0.013	**0.011**	**0.011**	0.013	0.014	**0.012**	**0.012**
25/5												
50%	0.035	0.034	**0.024**	**0.024**	0.027	0.027	**0.021**	**0.021**	0.027	0.026	**0.020**	**0.020**
65%	0.024	0.023	**0.014**	**0.014**	0.018	0.018	**0.013**	**0.013**	0.019	0.016	**0.011**	**0.011**
80%	**0.026**	0.036	0.027	0.027	0.023	0.032	**0.022**	**0.022**	0.024	0.034	**0.023**	**0.023**
All	0.014	0.018	**0.013**	**0.013**	0.013	0.016	**0.012**	**0.012**	0.016	0.019	**0.014**	0.015
50/2												
50%	0.026	0.024	**0.019**	**0.019**	0.029	0.028	**0.023**	**0.023**	0.025	0.024	**0.019**	**0.019**
65%	0.011	0.014	**0.010**	**0.010**	0.012	0.013	**0.010**	**0.010**	**0.010**	0.012	0.011	0.011
80%	0.022	0.031	**0.021**	0.022	0.024	0.032	**0.023**	**0.023**	**0.021**	0.030	**0.021**	**0.021**
All	0.007	**0.006**	**0.006**	**0.006**	0.007	0.007	0.007	0.007	0.008	**0.007**	**0.007**	**0.007**
50/5												
50%	0.027	0.025	**0.021**	**0.021**	0.023	0.022	**0.019**	**0.019**	0.022	0.020	**0.017**	**0.017**
65%	0.012	0.013	**0.010**	**0.010**	0.012	0.011	**0.010**	**0.010**	0.010	**0.009**	**0.009**	**0.009**
80%	**0.021**	0.030	**0.021**	**0.021**	**0.019**	0.027	**0.019**	**0.019**	**0.018**	0.025	**0.018**	**0.018**
All	**0.007**	0.008	**0.007**	**0.007**	**0.007**	0.009	**0.007**	**0.007**	**0.007**	0.008	**0.007**	**0.007**
100/5												
50%	0.015	0.015	**0.012**	**0.012**	0.016	0.015	**0.014**	**0.014**	0.013	0.012	**0.011**	**0.011**
65%	**0.008**	0.009	0.009	**0.008**	**0.006**	0.007	0.007	0.007	**0.007**	0.008	**0.007**	**0.007**
80%	**0.014**	0.019	**0.014**	**0.014**	**0.013**	0.019	0.014	0.014	**0.012**	0.018	0.014	0.014
All	**0.004**	**0.004**	0.005	0.005	**0.004**	**0.004**	0.005	0.005	**0.005**	**0.005**	0.006	0.006
100/10												
50%	0.014	0.014	**0.012**	**0.012**	0.011	0.011	0.011	0.011	0.012	**0.011**	**0.011**	**0.011**
65%	0.007	**0.006**	0.007	0.007	0.006	**0.005**	0.006	0.006	**0.005**	0.006	0.006	0.006
80%	**0.010**	0.015	0.012	0.012	**0.009**	0.013	0.010	0.010	**0.009**	0.014	0.012	0.011
All	**0.004**	0.005	0.005	0.005	**0.004**	**0.004**	0.006	0.006	**0.005**	**0.005**	0.006	0.006
200/5												
50%	0.009	0.009	**0.008**	**0.008**	0.009	0.009	0.009	0.009	0.007	0.007	0.007	0.007
65%	**0.005**	0.006	0.006	0.006	**0.004**	0.005	0.005	0.005	**0.004**	0.005	0.006	0.006
80%	**0.009**	0.013	0.011	0.011	**0.008**	0.012	**0.008**	**0.008**	**0.007**	0.010	0.009	0.009
All	0.003	**0.002**	0.005	0.005	0.003	**0.002**	0.005	0.005	**0.003**	**0.003**	0.005	0.005
200/10												
50%	0.009	0.009	**0.007**	**0.007**	**0.008**	**0.008**	0.009	0.009	0.007	0.007	0.007	**0.006**
65%	**0.004**	**0.004**	0.005	0.005	**0.004**	**0.004**	0.005	0.005	**0.003**	**0.003**	0.005	0.005
80%	**0.007**	0.011	0.009	0.009	**0.007**	0.011	0.008	0.008	**0.006**	0.009	0.008	0.008
All	**0.002**	**0.002**	0.005	0.005	**0.002**	**0.002**	0.005	0.005	**0.002**	**0.002**	0.005	0.005

Table 11: Standard Errors for Kappa Yielded by Each Procedure

	$r = 0.5$				$r = 0.8$				$r = 1.0$			
	NM	LL	BW	CM	NM	LL	BW	CM	NM	LL	BW	CM
25/2												
50%	0.040	0.037	**0.027**	**0.027**	0.041	0.042	**0.032**	**0.032**	0.040	0.038	**0.029**	**0.029**
65%	0.029	0.037	**0.019**	**0.019**	0.030	0.036	**0.023**	**0.023**	0.031	0.041	**0.025**	**0.025**
80%	0.046	0.057	**0.036**	**0.036**	0.045	0.068	**0.036**	**0.036**	0.046	0.074	**0.040**	**0.040**
All	0.021	0.022	**0.017**	**0.017**	0.023	0.023	**0.019**	**0.019**	0.024	0.025	**0.019**	**0.019**
25/5												
50%	0.036	0.035	**0.027**	**0.027**	0.030	0.028	**0.021**	**0.021**	0.031	0.032	**0.027**	**0.027**
65%	0.028	0.032	**0.022**	**0.022**	0.024	0.030	**0.021**	**0.021**	0.023	0.028	**0.020**	**0.020**
80%	0.043	0.059	**0.032**	**0.032**	0.038	0.069	**0.036**	**0.036**	0.035	0.063	**0.033**	**0.033**
All	0.022	0.023	0.017	**0.016**	0.020	0.022	0.017	**0.016**	0.022	0.024	**0.018**	**0.018**
50/2												
50%	0.042	0.043	**0.035**	**0.035**	0.040	0.040	**0.029**	**0.029**	0.031	0.026	**0.024**	**0.024**
65%	0.022	0.025	**0.017**	**0.017**	0.022	0.025	**0.018**	0.019	**0.019**	0.021	**0.019**	**0.019**
80%	0.037	0.066	0.036	**0.035**	0.040	0.066	**0.033**	**0.033**	0.035	0.073	0.033	**0.032**
All	0.016	0.017	**0.013**	0.014	0.017	0.017	**0.014**	**0.014**	0.017	0.016	**0.014**	**0.014**
50/5												
50%	0.031	0.031	**0.024**	**0.024**	0.031	0.030	**0.023**	**0.023**	0.025	0.023	**0.020**	**0.020**
65%	0.020	0.026	**0.018**	**0.018**	0.020	0.023	**0.019**	**0.019**	**0.017**	0.019	**0.017**	**0.017**
80%	0.040	0.066	**0.034**	**0.034**	0.036	0.059	**0.029**	**0.029**	**0.031**	0.067	0.035	0.035
All	0.017	0.015	**0.014**	**0.014**	0.017	0.018	**0.014**	**0.014**	0.015	0.015	**0.013**	**0.014**
100/5												
50%	**0.031**	**0.031**	0.032	0.032	0.025	0.022	**0.021**	**0.021**	**0.022**	0.023	**0.022**	**0.022**
65%	**0.014**	0.016	**0.014**	**0.014**	**0.012**	0.013	0.014	0.014	**0.013**	0.014	**0.013**	**0.013**
80%	0.027	0.052	0.027	**0.026**	**0.025**	0.057	0.028	0.028	**0.025**	0.050	0.031	0.031
All	**0.011**	**0.011**	0.012	0.012	**0.010**	0.011	**0.010**	**0.010**	0.012	**0.011**	**0.011**	**0.011**
100/10												
50%	0.025	**0.022**	0.024	0.024	**0.018**	**0.018**	0.019	0.019	**0.016**	0.017	0.017	0.017
65%	**0.011**	0.012	0.014	0.014	**0.010**	0.011	0.012	0.012	**0.011**	0.012	0.013	0.012
80%	**0.023**	0.071	0.030	0.030	**0.020**	0.054	0.028	0.028	0.029	0.041	**0.027**	**0.027**
All	**0.009**	**0.009**	0.010	0.010	**0.009**	**0.009**	0.010	0.010	0.010	0.010	0.010	0.010
200/5												
50%	**0.022**	0.024	0.034	0.033	0.019	0.019	0.019	0.019	**0.017**	**0.017**	0.022	0.022
65%	**0.009**	**0.009**	0.010	0.010	**0.007**	0.008	0.010	0.010	**0.009**	**0.009**	0.012	0.012
80%	**0.018**	0.042	0.023	0.022	**0.014**	0.034	0.022	0.022	**0.015**	0.030	0.023	0.023
All	**0.007**	**0.007**	0.009	0.009	**0.006**	0.007	0.009	0.009	**0.008**	**0.008**	0.011	0.011
200/10												
50%	**0.018**	0.019	0.026	0.026	**0.016**	**0.016**	0.017	0.017	**0.013**	**0.013**	0.016	0.016
65%	**0.007**	0.008	0.009	0.009	**0.007**	0.008	0.010	0.010	**0.006**	**0.006**	0.009	0.009
80%	**0.016**	0.047	0.020	0.020	**0.015**	0.044	0.025	0.025	**0.011**	0.025	0.020	0.020
All	**0.006**	**0.006**	0.010	0.010	**0.006**	**0.006**	0.009	0.009	**0.005**	0.006	0.009	0.009

Table 12: RMSE for P Yielded by Each Procedure

	$r = 0.5$				$r = 0.8$				$r = 1.0$			
	NM	LL	BW	CM	NM	LL	BW	CM	NM	LL	BW	CM
25/2												
50%	0.051	0.048	**0.035**	**0.035**	0.041	0.040	**0.036**	**0.036**	0.040	0.036	**0.031**	**0.031**
65%	**0.024**	**0.024**	0.073	0.073	**0.033**	**0.033**	0.081	0.081	**0.020**	0.021	0.044	0.044
80%	0.042	0.050	**0.032**	**0.032**	0.040	0.044	**0.027**	**0.027**	0.043	0.044	**0.028**	**0.028**
All	0.015	**0.013**	0.060	0.060	0.017	**0.014**	0.061	0.061	0.017	**0.015**	0.062	0.062
25/5												
50%	0.037	0.035	**0.029**	0.030	0.038	0.035	**0.021**	**0.021**	0.034	0.029	**0.022**	**0.022**
65%	0.033	**0.027**	0.054	0.053	0.021	**0.018**	0.041	0.042	0.021	**0.017**	0.037	0.037
80%	0.031	0.037	**0.027**	**0.027**	0.030	0.032	**0.023**	**0.023**	0.031	0.034	**0.024**	**0.024**
All	0.020	**0.018**	0.050	0.050	0.017	**0.016**	0.049	0.049	0.023	**0.020**	0.052	0.052
50/2												
50%	0.026	0.025	**0.019**	**0.019**	0.030	0.029	**0.023**	**0.023**	0.025	0.024	**0.019**	**0.019**
65%	**0.011**	0.015	0.042	0.042	**0.012**	0.013	0.034	0.034	**0.011**	0.012	0.034	0.034
80%	0.024	0.031	**0.022**	0.023	0.029	0.033	**0.023**	**0.023**	0.027	0.030	**0.021**	**0.021**
All	0.012	**0.007**	0.030	0.030	0.012	**0.007**	0.030	0.030	0.012	**0.007**	0.030	0.030
50/5												
50%	0.027	0.025	**0.021**	**0.021**	0.023	0.022	**0.020**	**0.020**	0.022	0.020	**0.018**	**0.018**
65%	**0.012**	0.014	0.033	0.033	**0.012**	0.013	0.031	0.031	0.010	**0.009**	0.026	0.026
80%	**0.021**	0.031	0.024	0.024	0.022	0.027	**0.019**	**0.019**	0.025	0.025	**0.019**	**0.018**
All	**0.009**	**0.009**	0.023	0.023	0.017	**0.010**	0.033	0.033	0.018	**0.010**	0.034	0.034
100/5												
50%	0.015	0.016	**0.013**	**0.013**	0.016	0.016	**0.014**	**0.014**	0.013	0.013	**0.011**	**0.011**
65%	**0.008**	0.010	0.018	0.018	**0.006**	0.010	0.017	0.017	0.010	**0.008**	0.013	0.013
80%	0.015	0.019	**0.014**	**0.014**	0.015	0.019	**0.014**	**0.014**	0.019	0.018	**0.015**	**0.015**
All	0.006	**0.004**	0.012	0.012	**0.006**	**0.006**	0.012	0.012	0.010	**0.005**	0.016	0.016
100/10												
50%	0.014	0.015	**0.012**	**0.012**	0.011	0.011	0.011	0.011	0.012	**0.011**	**0.011**	**0.011**
65%	**0.007**	**0.007**	0.014	0.014	**0.006**	0.007	0.014	0.014	0.009	**0.007**	0.009	0.009
80%	**0.010**	0.015	0.013	0.013	**0.010**	0.013	**0.010**	**0.010**	0.013	0.014	0.012	**0.011**
All	**0.004**	0.007	0.008	0.008	0.007	**0.004**	0.013	0.013	**0.005**	0.006	0.010	0.010
200/5												
50%	0.011	0.009	**0.008**	**0.008**	0.009	0.009	0.009	0.009	**0.007**	0.008	0.008	0.008
65%	**0.006**	0.012	0.014	0.014	**0.004**	0.009	0.011	0.011	0.010	**0.006**	0.007	0.007
80%	**0.010**	0.014	0.011	0.011	**0.008**	0.014	**0.008**	**0.008**	0.011	0.010	**0.009**	**0.009**
All	0.014	**0.009**	0.013	0.013	0.006	**0.003**	0.008	0.008	**0.003**	0.005	0.006	0.006
200/10												
50%	0.009	0.009	**0.007**	**0.007**	0.008	0.008	0.009	0.009	**0.007**	0.008	**0.007**	**0.007**
65%	**0.004**	0.007	0.010	0.010	**0.004**	0.007	0.009	0.009	0.007	**0.004**	0.007	0.007
80%	**0.007**	0.011	0.009	0.009	**0.007**	0.013	0.008	0.008	**0.008**	0.009	0.009	0.009
All	0.005	**0.002**	0.007	0.007	**0.003**	0.004	0.006	0.006	**0.002**	0.006	0.005	0.005

Table 13: RMSE for Kappa Yielded by Each Procedure

	$r = 0.5$				$r = 0.8$				$r = 1.0$			
	NM	LL	BW	CM	NM	LL	BW	CM	NM	LL	BW	CM
25/2												
50%	0.051	**0.042**	0.126	0.126	0.052	**0.042**	0.104	0.104	0.042	**0.039**	0.119	0.120
65%	**0.034**	0.040	0.125	0.125	**0.042**	0.044	0.100	0.100	0.043	**0.042**	0.092	0.092
80%	**0.046**	0.120	0.182	0.181	**0.045**	0.119	0.166	0.165	**0.054**	0.075	0.161	0.160
All	**0.021**	0.022	0.114	0.114	**0.023**	**0.023**	0.108	0.107	**0.024**	0.025	0.101	0.101
25/5												
50%	0.048	**0.035**	0.083	0.083	0.033	**0.032**	0.082	0.082	**0.034**	**0.034**	0.080	0.080
65%	**0.028**	0.033	0.103	0.102	**0.027**	0.031	0.083	0.083	**0.028**	**0.028**	0.069	0.069
80%	**0.048**	0.108	0.144	0.145	**0.053**	0.090	0.146	0.147	**0.058**	0.064	0.123	0.123
All	**0.022**	0.023	0.092	0.091	**0.020**	0.022	0.083	0.083	**0.023**	0.024	0.078	0.078
50/2												
50%	**0.042**	0.062	0.118	0.118	**0.042**	0.049	0.089	0.089	0.047	**0.027**	0.060	0.060
65%	**0.023**	0.031	0.086	0.087	0.026	**0.025**	0.063	0.064	0.032	**0.021**	0.051	0.051
80%	**0.040**	0.116	0.126	0.126	**0.040**	0.128	0.107	0.106	**0.057**	0.073	0.113	0.114
All	0.018	**0.017**	0.075	0.074	**0.017**	0.018	0.066	0.066	0.017	**0.016**	0.062	0.062
50/5												
50%	**0.035**	0.040	0.082	0.082	**0.031**	0.047	0.079	0.079	0.034	**0.023**	0.050	0.050
65%	**0.022**	0.026	0.061	0.061	**0.020**	0.024	0.061	0.061	0.030	**0.020**	0.036	0.036
80%	**0.040**	0.140	0.096	0.096	**0.038**	0.098	0.094	0.094	**0.063**	0.067	0.098	0.098
All	0.017	**0.015**	0.060	0.060	0.021	**0.018**	0.065	0.066	**0.015**	0.016	0.046	0.047
100/5												
50%	0.036	**0.031**	0.055	0.055	**0.026**	0.031	0.049	0.049	0.036	**0.023**	0.028	0.029
65%	**0.014**	0.022	0.041	0.041	**0.015**	**0.015**	0.027	0.028	0.029	**0.014**	0.017	0.017
80%	**0.039**	0.091	0.076	0.076	**0.031**	0.099	0.062	0.062	0.056	**0.050**	0.064	0.065
All	0.016	**0.012**	0.041	0.041	**0.011**	0.012	0.031	0.031	0.012	**0.011**	0.022	0.022
100/10												
50%	0.035	**0.023**	0.041	0.041	**0.019**	0.024	0.033	0.033	0.019	**0.017**	0.033	0.033
65%	**0.011**	0.016	0.032	0.032	0.014	**0.012**	0.021	0.021	0.016	**0.012**	0.023	0.023
80%	**0.037**	0.138	0.061	0.061	**0.029**	0.098	0.048	0.048	0.066	**0.041**	0.055	0.056
All	**0.009**	**0.009**	0.030	0.030	**0.009**	0.011	0.021	0.022	0.014	**0.011**	0.029	0.029
200/5												
50%	**0.023**	0.030	0.047	0.047	0.028	**0.019**	0.022	0.022	0.028	**0.017**	0.023	0.023
65%	**0.009**	0.017	0.021	0.021	**0.007**	0.015	0.019	0.019	0.019	**0.009**	0.014	0.014
80%	**0.036**	0.058	0.047	0.048	**0.025**	0.066	0.038	0.039	0.051	**0.030**	0.036	0.036
All	0.019	**0.012**	0.025	0.025	**0.008**	**0.008**	0.017	0.017	**0.008**	0.009	0.013	0.013
200/10												
50%	0.024	**0.019**	0.032	0.032	0.023	**0.017**	0.020	0.020	0.023	**0.013**	0.018	0.018
65%	**0.008**	0.010	0.015	0.015	**0.008**	0.011	0.014	0.014	0.013	**0.006**	0.013	0.013
80%	**0.033**	0.079	0.039	0.039	**0.029**	0.078	0.036	0.036	0.060	**0.025**	0.036	0.036
All	**0.006**	**0.006**	0.015	0.015	**0.006**	0.008	0.013	0.013	**0.006**	0.008	0.013	0.013

Table 14: Corrected and Original P Estimates

	Original BW and CM Procedures						Corrected BW and CM Procedures						Other Procedures					
	$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$	
	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	NM	LL	NM	LL	NM	LL
25/2																		
50%	0.860	0.860	0.855	0.854	0.860	0.860	0.855	0.855	0.847	0.847	0.850	0.850	0.837	0.837	0.828	0.829	0.830	0.836
65%	0.768	0.768	0.776	0.776	0.790	0.790	0.706	0.700	0.722	0.718	0.743	0.740	0.707	0.706	0.722	0.722	0.739	0.738
80%	0.871	0.871	0.875	0.875	0.868	0.868	0.873	0.874	0.876	0.877	0.866	0.867	0.893	0.883	0.894	0.883	0.887	0.871
All	0.536	0.536	0.542	0.543	0.553	0.553	0.487	0.484	0.496	0.493	0.507	0.505	0.487	0.482	0.494	0.489	0.504	0.498
25/5																		
50%	0.848	0.848	0.854	0.854	0.858	0.858	0.833	0.832	0.841	0.841	0.845	0.845	0.818	0.822	0.827	0.831	0.829	0.836
65%	0.796	0.796	0.806	0.806	0.819	0.819	0.756	0.755	0.772	0.771	0.792	0.791	0.766	0.758	0.777	0.771	0.793	0.788
80%	0.899	0.899	0.893	0.893	0.889	0.889	0.905	0.905	0.894	0.894	0.889	0.889	0.918	0.910	0.908	0.895	0.906	0.890
All	0.571	0.571	0.580	0.580	0.592	0.593	0.531	0.529	0.542	0.541	0.557	0.556	0.536	0.527	0.544	0.534	0.559	0.549

Table 15: Corrected and Original Kappa Estimates

	Original BW and CM Procedures						Corrected BW and CM Procedures						Other Procedures					
	$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$	
	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	NM	LL	NM	LL	NM	LL
25/2																		
50%	0.472	0.472	0.495	0.495	0.536	0.536	0.323	0.312	0.369	0.360	0.426	0.419	0.318	0.368	0.363	0.403	0.407	0.457
65%	0.521	0.521	0.542	0.542	0.570	0.570	0.386	0.375	0.427	0.419	0.471	0.465	0.379	0.382	0.415	0.419	0.453	0.457
80%	0.460	0.459	0.487	0.486	0.496	0.496	0.275	0.263	0.319	0.308	0.351	0.344	0.281	0.176	0.320	0.227	0.369	0.247
All	0.341	0.341	0.358	0.358	0.379	0.379	0.228	0.220	0.257	0.251	0.288	0.283	0.227	0.224	0.252	0.249	0.281	0.276
25/5																		
50%	0.546	0.546	0.571	0.571	0.607	0.607	0.444	0.439	0.488	0.485	0.539	0.537	0.435	0.470	0.478	0.508	0.519	0.551
65%	0.566	0.566	0.595	0.595	0.621	0.621	0.470	0.465	0.518	0.514	0.557	0.555	0.462	0.460	0.504	0.506	0.539	0.542
80%	0.480	0.480	0.524	0.524	0.537	0.537	0.341	0.335	0.406	0.402	0.433	0.430	0.361	0.249	0.419	0.325	0.464	0.365
All	0.393	0.392	0.415	0.414	0.437	0.437	0.306	0.303	0.340	0.338	0.374	0.372	0.305	0.302	0.335	0.330	0.368	0.361

Table 16: Corrected and Original Bias for P

	Original BW and CM Procedures						Corrected BW and CM Procedures						Other Procedures					
	$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$	
	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	NM	LL	NM	LL	NM	LL
25/2																		
50%	0.013	0.013	0.023	0.023	0.019	0.018	0.007	0.008	0.016	0.016	0.009	0.008	-0.011	-0.010	-0.003	-0.003	-0.012	-0.006
65%	0.072	0.072	0.080	0.080	0.042	0.042	0.010	0.005	0.026	0.022	-0.004	-0.007	0.011	0.010	0.026	0.025	-0.008	-0.009
80%	-0.002	-0.002	0.007	0.008	0.012	0.012	0.000	0.001	0.008	0.009	0.011	0.011	0.021	0.011	0.027	0.016	0.032	0.015
All	0.059	0.059	0.060	0.060	0.061	0.061	0.009	0.007	0.013	0.011	0.015	0.013	0.010	0.005	0.012	0.006	0.012	0.006
25/5																		
50%	0.017	0.017	0.001	0.001	0.009	0.009	0.002	0.001	-0.012	-0.012	-0.004	-0.004	-0.013	-0.009	-0.026	-0.022	-0.020	-0.013
65%	0.052	0.052	0.040	0.040	0.035	0.036	0.012	0.010	0.006	0.005	0.008	0.007	0.022	0.014	0.010	0.005	0.009	0.004
80%	-0.001	-0.001	0.005	0.005	0.003	0.003	0.004	0.005	0.006	0.006	0.003	0.003	0.017	0.009	0.020	0.007	0.019	0.004
All	0.049	0.049	0.047	0.047	0.050	0.050	0.008	0.007	0.009	0.008	0.015	0.014	0.014	0.005	0.011	0.001	0.016	0.006

Table 17: Corrected and Original Bias for Kappa

	Original BW and CM Procedures						Corrected BW and CM Procedures						Other Procedures					
	$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$	
	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	NM	LL	NM	LL	NM	LL
25/2																		
50%	0.123	0.123	0.099	0.099	0.116	0.116	-0.026	-0.037	-0.027	-0.036	0.006	-0.001	-0.032	0.019	-0.033	0.007	-0.013	-0.006
65%	0.124	0.124	0.098	0.098	0.088	0.088	-0.011	-0.022	-0.018	-0.026	-0.011	-0.017	-0.018	-0.015	-0.030	-0.025	-0.029	-0.009
80%	0.178	0.177	0.162	0.161	0.156	0.155	-0.007	-0.020	-0.007	-0.018	0.011	0.003	-0.001	-0.106	-0.006	-0.098	0.028	0.015
All	0.113	0.113	0.106	0.105	0.100	0.099	0.000	-0.008	0.004	-0.002	0.008	0.004	-0.001	-0.004	0.000	-0.004	0.001	0.006
25/5																		
50%	0.078	0.078	0.079	0.079	0.076	0.075	-0.024	-0.029	-0.004	-0.007	0.008	0.006	-0.032	0.002	-0.014	0.016	-0.012	-0.013
65%	0.101	0.100	0.080	0.080	0.066	0.066	0.004	-0.001	0.003	0.000	0.002	-0.001	-0.004	-0.005	-0.011	-0.009	-0.017	0.004
80%	0.141	0.141	0.142	0.142	0.118	0.119	0.002	-0.004	0.023	0.020	0.015	0.012	0.022	-0.090	0.036	-0.057	0.046	0.004
All	0.091	0.090	0.082	0.081	0.076	0.076	0.004	0.001	0.007	0.005	0.013	0.011	0.003	0.000	0.002	-0.003	0.006	0.006

Table 18: Corrected and Original Standard Errors for P

	Original BW and CM Procedures						Corrected BW and CM Procedures						Other Procedures					
	$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$	
	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	NM	LL	NM	LL	NM	LL
25/2																		
50%	0.033	0.032	0.028	0.028	0.025	0.025	0.045	0.046	0.039	0.040	0.035	0.036	0.050	0.047	0.041	0.040	0.038	0.035
65%	0.013	0.013	0.013	0.013	0.014	0.014	0.022	0.023	0.020	0.022	0.020	0.021	0.021	0.022	0.020	0.022	0.018	0.019
80%	0.032	0.032	0.026	0.026	0.025	0.025	0.045	0.045	0.036	0.037	0.036	0.036	0.037	0.049	0.030	0.041	0.029	0.041
All	0.011	0.011	0.011	0.011	0.012	0.012	0.012	0.012	0.013	0.013	0.013	0.013	0.011	0.012	0.012	0.013	0.013	0.014
25/5																		
50%	0.024	0.024	0.021	0.021	0.020	0.020	0.034	0.034	0.027	0.027	0.026	0.026	0.035	0.034	0.027	0.027	0.027	0.026
65%	0.014	0.014	0.013	0.013	0.011	0.011	0.021	0.022	0.017	0.018	0.016	0.016	0.024	0.023	0.018	0.018	0.019	0.016
80%	0.027	0.027	0.022	0.022	0.023	0.023	0.032	0.033	0.028	0.028	0.029	0.030	0.026	0.036	0.023	0.032	0.024	0.034
All	0.013	0.013	0.012	0.012	0.014	0.015	0.014	0.015	0.013	0.014	0.017	0.017	0.014	0.018	0.013	0.016	0.016	0.019

Table 19: Corrected and Original Standard Errors for Kappa

	Original BW and CM Procedures						Corrected BW and CM Procedures						Other Procedures					
	$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$	
	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	NM	LL	NM	LL	NM	LL
25/2																		
50%	0.027	0.027	0.032	0.032	0.029	0.029	0.039	0.042	0.043	0.046	0.042	0.044	0.040	0.037	0.041	0.042	0.040	0.038
65%	0.019	0.019	0.023	0.023	0.025	0.025	0.032	0.036	0.035	0.038	0.037	0.040	0.029	0.037	0.030	0.036	0.031	0.041
80%	0.036	0.036	0.036	0.036	0.040	0.040	0.050	0.053	0.051	0.054	0.054	0.056	0.046	0.057	0.045	0.068	0.046	0.074
All	0.017	0.017	0.019	0.019	0.019	0.019	0.023	0.025	0.025	0.027	0.026	0.028	0.021	0.022	0.023	0.023	0.024	0.025
25/5																		
50%	0.027	0.027	0.021	0.021	0.027	0.027	0.038	0.039	0.029	0.031	0.036	0.037	0.036	0.035	0.030	0.028	0.031	0.032
65%	0.022	0.022	0.021	0.021	0.020	0.020	0.031	0.033	0.029	0.031	0.027	0.028	0.028	0.032	0.024	0.030	0.023	0.028
80%	0.032	0.032	0.036	0.036	0.033	0.033	0.044	0.046	0.048	0.049	0.043	0.043	0.043	0.059	0.038	0.069	0.035	0.063
All	0.017	0.016	0.017	0.016	0.018	0.018	0.022	0.023	0.021	0.022	0.024	0.024	0.022	0.023	0.020	0.022	0.022	0.024

Table 20: Corrected and Original RMSE for P

	Original BW and CM Procedures						Corrected BW and CM Procedures						Other Procedures					
	$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$	
	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	NM	LL	NM	LL	NM	LL
25/2																		
50%	0.035	0.035	0.036	0.036	0.031	0.031	0.046	0.047	0.042	0.043	0.036	0.036	0.051	0.048	0.041	0.040	0.040	0.036
65%	0.073	0.073	0.081	0.081	0.044	0.044	0.024	0.024	0.033	0.031	0.020	0.022	0.024	0.024	0.033	0.033	0.020	0.021
80%	0.032	0.032	0.027	0.027	0.028	0.028	0.045	0.045	0.037	0.038	0.037	0.038	0.042	0.050	0.040	0.044	0.043	0.044
All	0.060	0.060	0.061	0.061	0.062	0.062	0.015	0.014	0.018	0.017	0.020	0.018	0.015	0.013	0.017	0.014	0.017	0.015
25/5																		
50%	0.029	0.030	0.021	0.021	0.022	0.022	0.034	0.034	0.029	0.030	0.026	0.026	0.037	0.035	0.038	0.035	0.034	0.029
65%	0.054	0.053	0.041	0.042	0.037	0.037	0.024	0.024	0.018	0.018	0.017	0.017	0.033	0.027	0.021	0.018	0.021	0.017
80%	0.027	0.027	0.023	0.023	0.024	0.024	0.033	0.033	0.028	0.029	0.030	0.030	0.031	0.037	0.030	0.032	0.031	0.034
All	0.050	0.050	0.049	0.049	0.052	0.052	0.017	0.016	0.016	0.016	0.022	0.022	0.020	0.018	0.017	0.016	0.023	0.020

Table 21: Corrected and Original RMSE for Kappa

	Original BW and CM Procedures						Corrected BW and CM Procedures						Other Procedures					
	$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$		$r = 0.5$		$r = 0.8$		$r = 1.0$	
	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	BW	CM	NM	LL	NM	LL	NM	LL
25/2																		
50%	0.126	0.126	0.104	0.104	0.119	0.120	0.047	0.056	0.051	0.059	0.043	0.044	0.051	0.042	0.052	0.042	0.042	0.039
65%	0.125	0.125	0.100	0.100	0.092	0.092	0.034	0.042	0.039	0.046	0.038	0.043	0.034	0.040	0.042	0.044	0.043	0.042
80%	0.182	0.181	0.166	0.165	0.161	0.160	0.051	0.057	0.051	0.057	0.055	0.056	0.046	0.120	0.045	0.119	0.054	0.075
All	0.114	0.114	0.108	0.107	0.101	0.101	0.023	0.026	0.026	0.027	0.027	0.028	0.021	0.022	0.023	0.023	0.024	0.025
25/5																		
50%	0.083	0.083	0.082	0.082	0.080	0.080	0.045	0.048	0.030	0.031	0.037	0.037	0.048	0.035	0.033	0.032	0.034	0.034
65%	0.103	0.102	0.083	0.083	0.069	0.069	0.031	0.033	0.030	0.031	0.027	0.028	0.028	0.033	0.027	0.031	0.028	0.028
80%	0.144	0.145	0.146	0.147	0.123	0.123	0.044	0.046	0.053	0.052	0.046	0.045	0.048	0.108	0.053	0.090	0.058	0.064
All	0.092	0.091	0.083	0.083	0.078	0.078	0.023	0.023	0.022	0.023	0.027	0.026	0.022	0.023	0.020	0.022	0.023	0.024

Table 22: Corrected and Original Decision Consistency Estimates for Jurisdiction #1

	Hypothetical and Hypothetical Forms						Hypothetical and Actual Forms					
	Raw		*CSS*		*NSS*		Raw		*CSS*		*NSS*	
	P	Kappa	*P*	Kappa	*P*	Kappa	*P*	Kappa	*P*	Kappa	*P*	Kappa
Using the Original Procedures												
NM	0.847	0.661	0.847	0.648	0.837	0.640	—	—	—	—	—	—
LL	0.846	0.667	0.849	0.658	0.837	0.643	0.846	0.667	0.848	0.659	0.836	0.645
BL	0.866	0.714	—	—	—	—	—	—	—	—	—	—
BW	0.867	0.714	0.854	0.676	0.867	0.716	0.905	0.795	0.898	0.773	0.905	0.796
CM	0.867	0.714	0.856	0.677	0.867	0.714	0.905	0.795	0.898	0.771	0.905	0.795
Using the Bias-Corrected Procedures												
BW	0.857	0.691	0.846	0.651	0.857	0.690	0.897	0.778	0.889	0.751	0.897	0.778
CM	0.857	0.690	0.846	0.648	0.857	0.690	0.897	0.778	0.889	0.751	0.897	0.778

Table 23: Corrected and Original Decision Consistency Estimates for Jurisdiction #2

	Hypothetical and Hypothetical Forms						Hypothetical and Actual Forms					
	Raw		*CSS*		*NSS*		Raw		*CSS*		*NSS*	
	P	Kappa	*P*	Kappa	*P*	Kappa	*P*	Kappa	*P*	Kappa	*P*	Kappa
Using the Original Procedures												
NM	0.829	0.649	0.890	0.619	0.820	0.625	—	—	—	—	—	—
LL	0.836	0.659	0.889	0.634	0.821	0.629	0.836	0.658	0.889	0.636	0.822	0.628
BL	0.853	0.695	—	—	—	—	—	—	—	—	—	—
BW	0.856	0.701	0.895	0.673	0.856	0.702	0.899	0.790	0.924	0.761	0.898	0.788
CM	0.854	0.697	0.897	0.663	0.854	0.697	0.897	0.786	0.925	0.745	0.897	0.786
Using the Bias-Corrected Procedures												
BW	0.842	0.672	0.892	0.634	0.842	0.672	0.889	0.767	0.916	0.726	0.889	0.767
CM	0.842	0.670	0.895	0.632	0.842	0.670	0.888	0.767	0.922	0.728	0.888	0.767

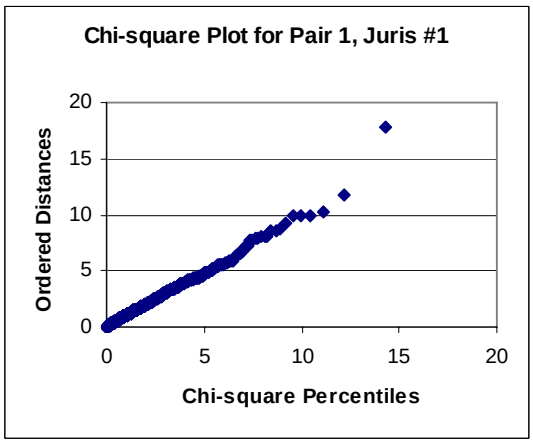
Table 24: Model Fit for the LL Procedure

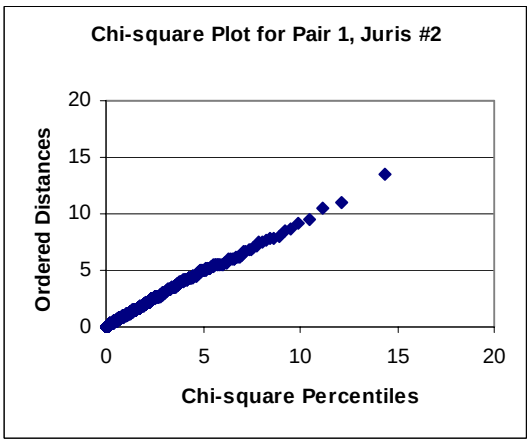
	Mean	S.D.	Skew.	Kurt.
Juris #1				
Observed	30.1415	9.3932	-0.1162	2.6594
Fitted	30.1415	9.3932	-0.1162	2.4728
Likelihood Ratio Chi-Square = 82.1210 (with df = 52), $p < 0.01$				
Juris #2				
Observed	31.3489	8.5777	-0.2887	2.7041
Fitted	31.3489	8.5777	-0.2887	2.5949
Likelihood Ratio Chi-Square = 78.7502 (with df = 49), $p < 0.01$				

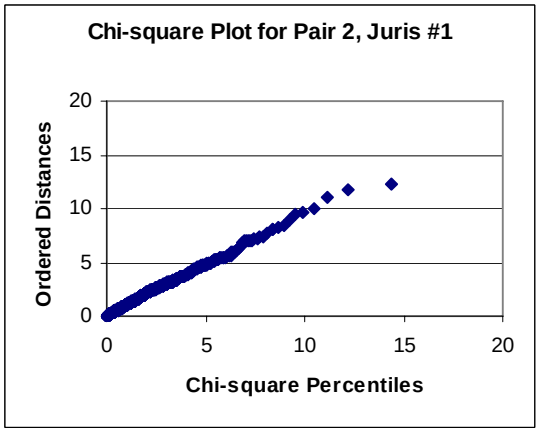
Table 25: Observed and Predicted Pass Rates

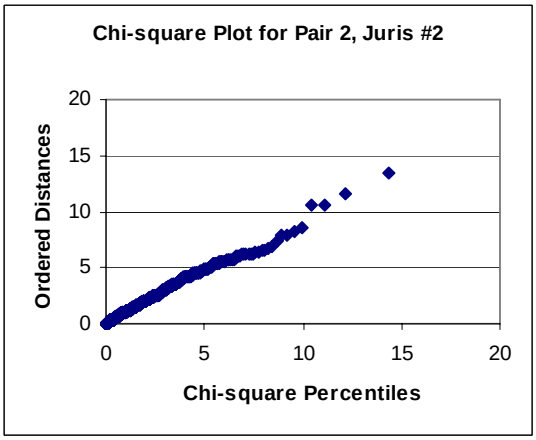
Two Hypothetical Forms					Hypothetical and Actual Forms				
NM	LL	BL	BW	CM	NM	LL	BL	BW	CM
Juris #1: observed pass rate = 0.365									
0.343	0.361	0.371	0.371	0.371	—	0.364	—	0.368	0.368
Juris #2: observed pass rate = 0.602									
0.580	0.596	0.599	0.598	0.597	—	0.600	—	0.598	0.599

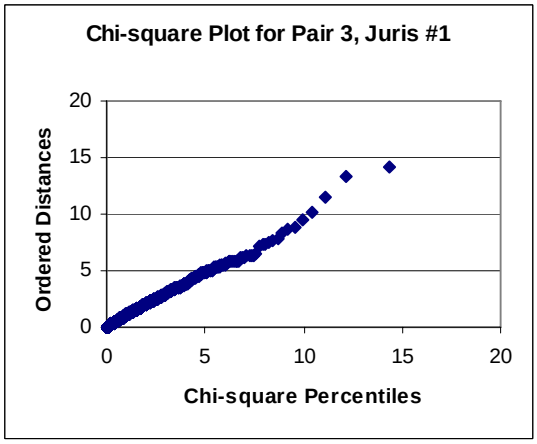
Figure 1: Chi-square Plots for the Jurisdictions











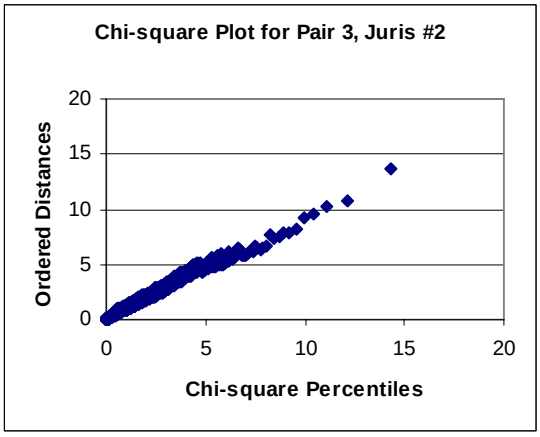
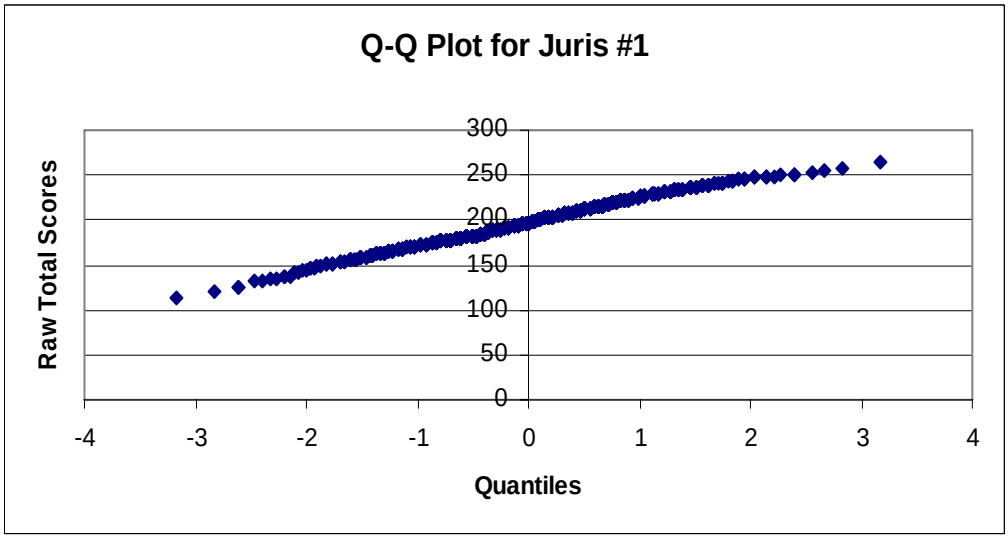


Figure 2: Q-Q Plots for the Jurisdictions



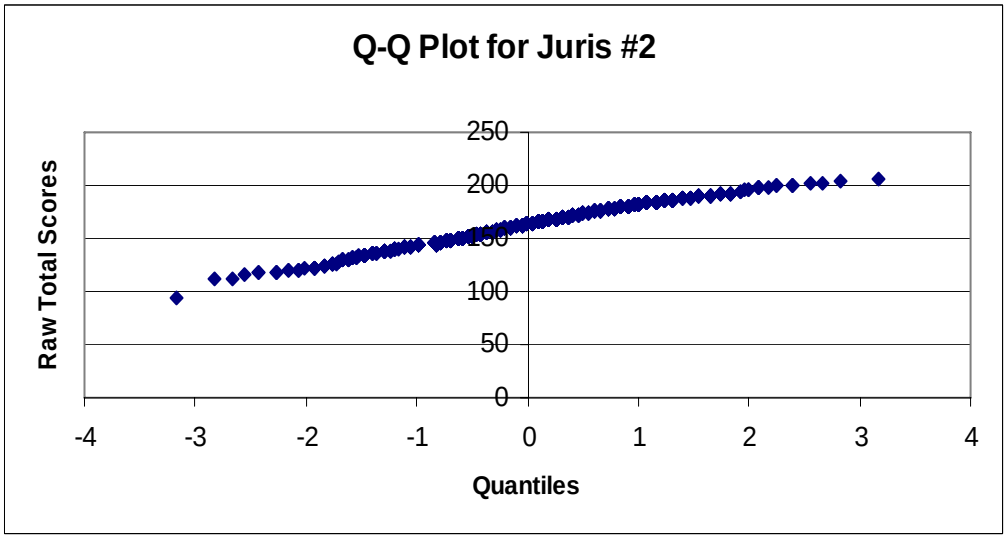
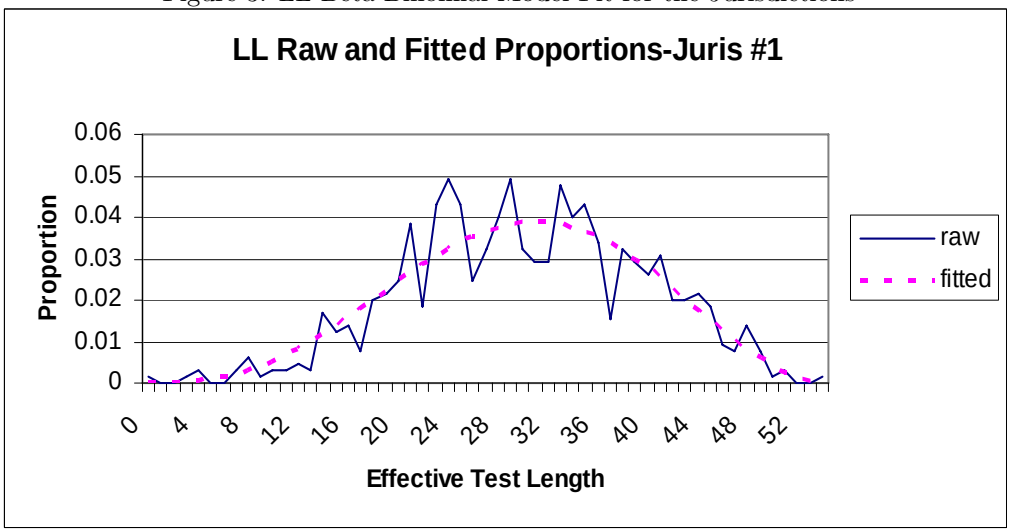
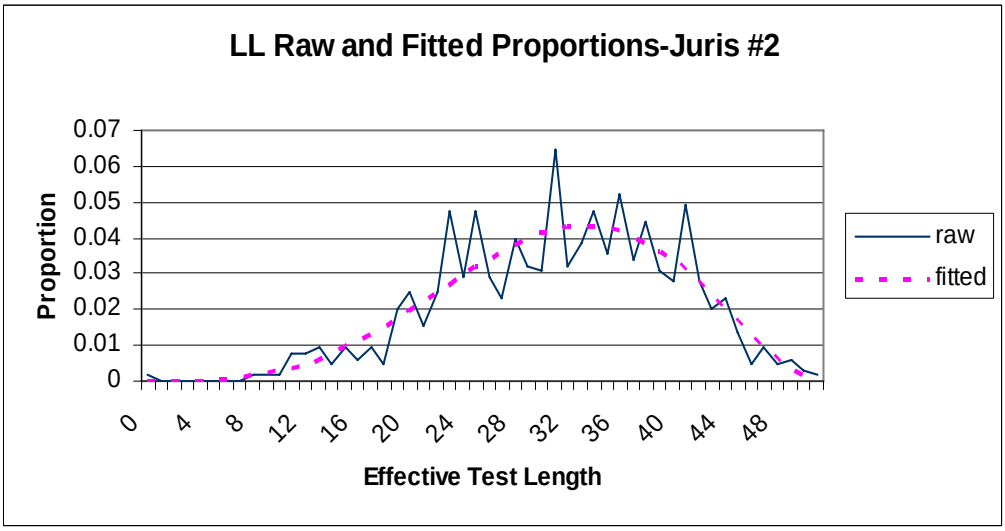


Figure 3: LL Beta Binomial Model Fit for the Jurisdictions





5 References

Algina, J., & Noe, M. (1978). A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. *Journal of Educational Measurement*, 15, 101-110.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L., & Wan, L. (2004). *Bootstrap procedures for estimating decision consistency for single-administration complex assessments* (CASMA Research Report No. 7). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Berk, R. (1980). A consumer's guide to criterion-referenced test reliability. *Journal of Educational Measurement*, 17, 323-346.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy, Version 1.1* (CASMA Research Report No. 9). Iowa City, IA: University of Iowa.

Brennan, R. L., & Lee, W. (2006). *Correcting for bias in single-administration decision consistency indexes* (CASMA Research Report No. 18). Iowa City, IA: University of Iowa.

Breyer, F. J., & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method* (ETS Research Report No. 94-39). Princeton, NJ: Educational Testing Service.

Case, S. (2005). Demystifying scaling to the MBE: How'd you do that? *The Bar Examiners*, 74, 45-46.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education and Macmillan.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and methods for criterion-referenced tests. *Journal of Educational Measurement*, 10, 159-170.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, 27, 345-359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13, 253-264.

Johnson, R., & Wichern, D. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

Klein, S. (1995). Options for combining MBE and Essay scores. *The Bar Examiners*, 64, 38-43.

Kline, R. (2004). *Principles and practice of structural equation modeling*. The Guilford Press, NY.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices* (2nd edition). New York: Springer.

Lee, W. (2005a). *A multinomial error model for tests with polytomous items* (CASMA Research Report No. 10). Iowa City, IA: University of Iowa.

Lee, W. (2005b). *Classification consistency under the compound multinomial model* (CASMA Research Report No. 13). Iowa City, IA: University of Iowa.

Lee, W. (2005c). *Manual for MULT-CLASS: For multinomial and compound-multinomial classification consistency*. Iowa City, IA: University of Iowa.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, 26, 412-432.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179-197.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: NJ. Lawrence Erlbaum Associates, Publishers.

Lukhele, R., Thissen, D., & Wainer, H. (1994). The relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, 31, 234-250.

Muraki, E. (1997). A generalized partial credit model. In W. J. Linden, & R. K. Hambleton (Ed.), *Handbook of modern item response theory*. Springer.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement*, 17, 359-368.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13, 265-276.

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement*, 15, 111-116.

Subkoviak, M. J. (1980). Decision-consistency approaches. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore MD: Johns Hopkins University Press.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25, 47-55.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision theoretic formulation. *Journal of Educational Measurement*, 11, 263-267.

Traub, R., & Rowley, G. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement*, 4, 517-545.

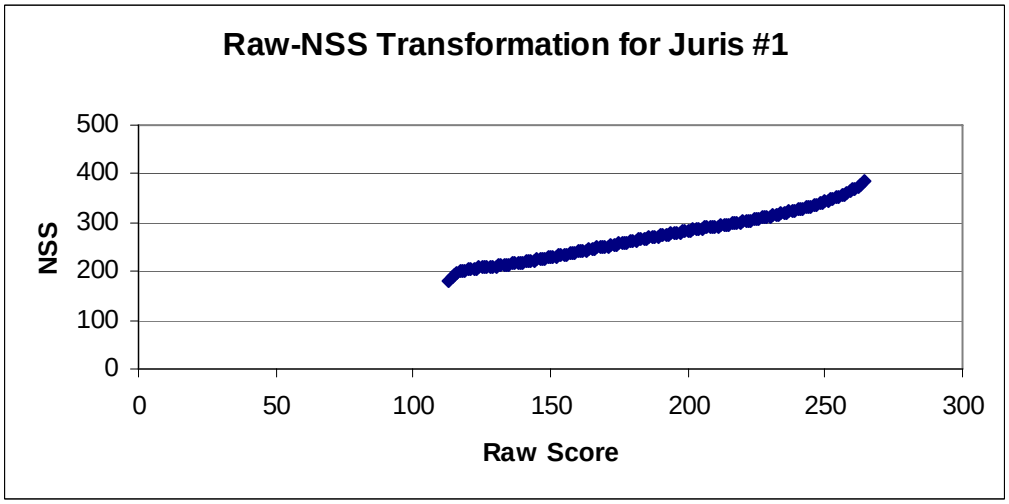
Wan, L. (2006). *Estimating classification consistency for single-administration complex assessments using non-IRT procedures*. Unpublished dissertation. The University of Iowa, Iowa City, IA.

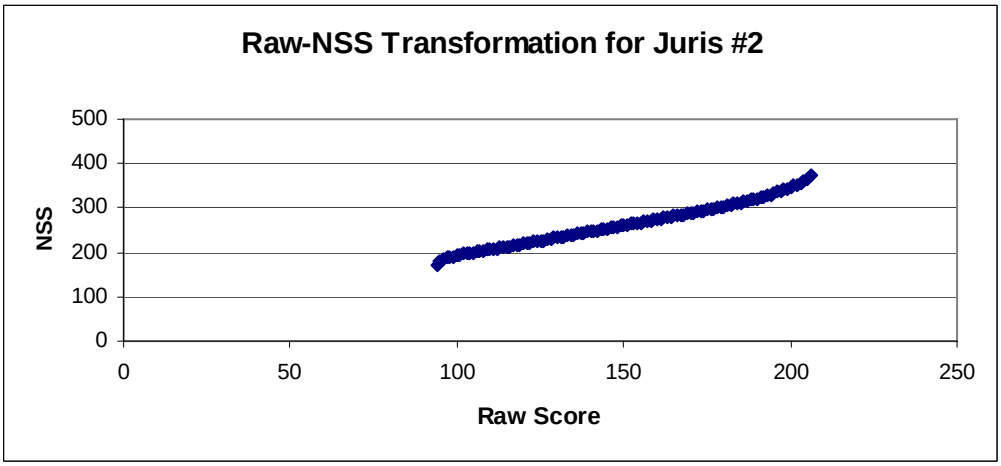
Wan, L., Lee, W., Brennan, R., & Chien, Y. (2006). *Comparison of procedures for estimating classification consistency and accuracy for complex assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco, CA.

Woodruff, D. J., & Sawyer, R. L. (1989). Estimating measures of pass-fail reliability from parallel half-tests. *Applied Psychological Measurement*, 13, 33-43.

A Raw to NSS Transformation Plots

Figure 1: Raw Score to NSS Transformation Plots





B Score Ranges and Cut Scores

Data Type	Score Range	Cut Score(s)
Juris #1 raw	113-264	208
Juris #1 CSS	191-351	267

Juris #1 NSS	182-387	291
Juris #2 raw	94-206	159
Juris #2 CSS	183-339	254
Juris #2 NSS	173-376	273
25/2 raw	0-33	17, 22, 27
25/5 raw	0-45	23, 30, 36
50/2 raw	0-58	29, 38, 47
50/5 raw	0-70	35, 46, 56
100/5 raw	0-120	60, 78, 96
100/10 raw	0-140	70, 91, 112
200/5 raw	0-220	110, 143, 176
200/10 raw	0-240	120, 156, 192

C Percentile Ranks with Different Test Length

Percentile Ranks of the Score Points for the 25/2 Type (r = 0.5)

R Marg	C Marg	R PRs	C PRs	prop	score
0.00000	0.00000	0.00	0.00	0.000	0
0.00000	0.00000	0.00	0.00	0.030	1
0.00000	0.00000	0.00	0.00	0.061	2
0.00000	0.00000	0.00	0.00	0.091	3
0.00001	0.00002	0.00	0.00	0.121	4
0.00005	0.00005	0.00	0.00	0.152	5
0.00015	0.00016	0.01	0.02	0.182	6
0.00038	0.00034	0.04	0.04	0.212	7
0.00087	0.00085	0.10	0.10	0.242	8
0.00177	0.00171	0.23	0.23	0.273	9
0.00336	0.00330	0.49	0.48	0.303	10
0.00600	0.00591	0.96	0.94	0.333	11
0.00963	0.00967	1.74	1.72	0.364	12
0.01543	0.01515	2.99	2.96	0.394	13
0.02256	0.02161	4.89	4.80	0.424	14
0.03122	0.03124	7.58	7.44	0.455	15
0.04182	0.04166	11.23	11.08	0.485	16
0.05322	0.05318	15.98	15.83	0.515	17
0.06488	0.06486	21.89	21.73	0.545	18
0.07637	0.07664	28.95	28.80	0.576	19
0.08568	0.08594	37.05	36.93	0.606	20
0.09281	0.09172	45.98	45.82	0.636	21
0.09374	0.09342	55.31	55.07	0.667	22
0.09072	0.09132	64.53	64.31	0.697	23
0.08264	0.08303	73.20	73.03	0.727	24
0.07072	0.07190	80.86	80.77	0.758	25

0.05680	0.05704	87.24	87.22	0.788	26
0.04147	0.04148	92.15	92.15	0.818	27
0.02771	0.02771	95.61	95.61	0.848	28
0.01650	0.01662	97.82	97.82	0.879	29
0.00866	0.00862	99.08	99.08	0.909	30
0.00361	0.00358	99.69	99.69	0.939	31
0.00109	0.00111	99.93	99.93	0.970	32
0.00017	0.00017	99.99	99.99	1.000	33

0.64377	0.64437	Mean
0.12573	0.12549	SD
-0.23938	-0.24171	Skewness
2.82279	2.82721	Kurtosis

Percentile Ranks of the Score Points for the 25/5 Type (r = 0.5)

R Marg	C Marg	R PRs	C PRs	prop	score
0.00000	0.00000	0.00	0.00	0.000	0
0.00000	0.00000	0.00	0.00	0.022	1
0.00000	0.00000	0.00	0.00	0.044	2
0.00000	0.00000	0.00	0.00	0.067	3
0.00000	0.00000	0.00	0.00	0.089	4
0.00001	0.00000	0.00	0.00	0.111	5
0.00003	0.00003	0.00	0.00	0.133	6
0.00007	0.00007	0.01	0.01	0.156	7
0.00014	0.00015	0.02	0.02	0.178	8
0.00032	0.00032	0.04	0.04	0.200	9
0.00064	0.00060	0.09	0.09	0.222	10
0.00118	0.00114	0.18	0.17	0.244	11
0.00201	0.00202	0.34	0.33	0.267	12
0.00324	0.00314	0.60	0.59	0.289	13
0.00507	0.00491	1.02	0.99	0.311	14
0.00728	0.00710	1.63	1.59	0.333	15
0.01016	0.01002	2.51	2.45	0.356	16
0.01364	0.01347	3.70	3.62	0.378	17
0.01797	0.01733	5.28	5.16	0.400	18
0.02260	0.02217	7.30	7.14	0.422	19
0.02778	0.02757	9.82	9.63	0.444	20
0.03383	0.03324	12.90	12.67	0.467	21
0.03998	0.03902	16.60	16.28	0.489	22
0.04589	0.04522	20.89	20.49	0.511	23
0.05196	0.05120	25.78	25.31	0.533	24
0.05756	0.05732	31.26	30.74	0.556	25
0.06259	0.06200	37.26	36.70	0.578	26

0.06628	0.06500	43.71	43.05	0.600	27
0.06848	0.06821	50.44	49.71	0.622	28
0.06805	0.06856	57.27	56.55	0.644	29
0.06613	0.06667	63.98	63.32	0.667	30
0.06273	0.06300	70.42	69.80	0.689	31
0.05722	0.05820	76.42	75.86	0.711	32
0.05012	0.05102	81.79	81.32	0.733	33
0.04269	0.04329	86.43	86.04	0.756	34
0.03451	0.03508	90.29	89.95	0.778	35
0.02674	0.02768	93.35	93.09	0.800	36
0.01967	0.02047	95.67	95.50	0.822	37
0.01380	0.01416	97.34	97.23	0.844	38
0.00904	0.00949	98.49	98.41	0.867	39
0.00537	0.00575	99.21	99.18	0.889	40
0.00301	0.00311	99.63	99.62	0.911	41
0.00142	0.00151	99.85	99.85	0.933	42
0.00059	0.00056	99.95	99.95	0.956	43
0.00019	0.00017	99.99	99.99	0.978	44
0.00002	0.00003	100.00	100.00	1.000	45

0.61523	0.61723	Mean
0.12647	0.12662	SD
-0.20803	-0.21616	Skewness
2.77612	2.77573	Kurtosis