Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 46

Interval Estimation Procedures for Weighted Composite Scores

Kyung Yong Kim[†] *and Won–Chan Lee*[††]

April 2016

[†]Kyung Yong Kim, 210G Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: kyungyong-kim@uiowa.edu)

[††]Won-Chan Lee is Associate Professor and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu).

Contents

List of Tables

List of Figures

Abstract

Assuming that errors of measurement follow a compound binomial distribution, five interval estimation procedures that can be used to construct intervals for weighted composite scores are introduced in this paper. The intervals that result from these five procedures are the compound normal approximation, Mee, Haldane, Jeffreys-Perks, and score intervals. These five intervals are compared through a simulation study under various study conditions. In general, the Mee, Haldane, Jeffreys-Perks, and score intervals returned coverage probabilities that are much closer to the nominal level than the compound normal approximation intervals, with the Jeffreys-Perks intervals achieving somewhat better results. In addition, the widths for these four intervals are, on average, shorter than the compound normal approximation intervals. Among the four intervals, the Haldane and Jeffreys-Perks intervals are more attractive than the Mee and score intervals in the sense that the former intervals both have closed-form expressions.

1 Introduction

In classical test theory (CTT), the true score for an examinee is the score that would be obtained by taking the average of the examinee's observed scores from an infinite number of test administrations. Because it is impossible to have an examinee take the same test an infinite number of times, true score can only exist as a hypothetical concept. For this reason, observed score is often used as a point estimate for true score. However, even though an examinee's observed score is an unbiased estimate of that examinee's true score, it does not provide information about the amount of uncertainty in the point estimate. An alternative way to estimate true score is through an interval estimate, which is likely to contain the true score.

Under the assumption that measurement errors are normally distributed, interval estimates have traditionally been constructed using a common standard error of measurement (SEM) for all examinees. However, for tests consisting of equally weighted dichotomous items, more satisfying results can be obtained by using the normal approximation interval with Lord's conditional SEM (Lord, 1955, 1957) or the Wilson score interval (Wilson, 1927), assuming that measurement errors follow a binomial distribution (Lee, Brennan, & Kolen, 2006).

Interval estimates for composite scores across multiple stratified domains have been an infrequent topic of discussion in the literature. Feldt (1984) proposed a formula for computing conditional SEMs for a test composed of stratified domains under the assumption that the error distribution for composite scores follows a compound binomial distribution (i.e., the error distribution for each domain is binomial). Brossman and Lee (2009) considered normal approximation intervals using Feldt's conditional SEMs; however, they did not provide a theoretical justification for using a normal distribution to approximate a compound binomial distribution. Another procedure that takes stratified domains into account is the tolerance interval estimation procedure proposed by Brossman and Lee (2009). Tolerance intervals, unlike other intervals that use observed scores, are based on Kelley's (1927) regressed score estimates. However, this procedure is difficult to extend to more than two stratified domains because of the complexity of the construction, and it tends to return intervals with inaccurate coverage probabilities for most of the true score points.

Most testing programs use test specifications when developing test forms, and it is usually the case that some domains (e.g., contents) are considered more important than other domains. The relative importance of various domains is typically addressed by allocating different number of items across domains or using differential weights when forming composite scores. One advantage of the interval estimation procedures using the compound binomial error model compared to those using the binomial error model is that the former carries test specifications into the interval estimation process by constructing intervals using domain scores. As a result, examinees with the same test score can be assigned different intervals depending on the pattern of domain scores. On the other hand, the interval estimation procedures using the binomial error model ignore domain stratification and always assign an identical interval to examinees

with the same composite score, regardless of their domain score configuration. In spite of this advantage, there is very little literature that describes interval estimation procedures using the compound binomial error model. Thus, the objectives of the present study are as follows:

1. Introduce and compare five interval estimation procedures for true weighted composite scores based on the compound binomial error model using a simulation study.
2. Compare the five compound-binomial procedures to the binomial-based normal approximation and Wilson score procedures using real data.

2 True Weighted Composite Score

Suppose a test form consists of k stratified domains. Let $\{X_i\}_{i=1}^k$ denote a set of random variables representing observed number-correct scores for the k domains. Further, let $\{\tau_i\}_{i=1}^k$ be a set of true number-correct scores for the k domains, where τ_i is defined as the expected value of X_i over repeated measurements. Then, for any set of weights $\{w_i\}_{i=1}^k$, the observed weighted composite score is $X = \sum_{i=1}^k w_i X_i$, and the true weighted composite score τ is defined as the expected value of X over repeated measurements; that is,

$$\tau = E(X) = E\left(\sum_{i=1}^k w_i X_i\right) = \sum_{i=1}^k w_i E(X_i) = \sum_{i=1}^k w_i \tau_i, \tag{1}$$

where E is the expectation operator. Under the compound binomial error model, the ith domain score X_i conditional on true proportion-correct score π_i is assumed to have a binomial distribution with the number of items n_i. Thus, $E(X_i) = \tau_i = n_i \pi_i$, and Equation 1 becomes $\tau = \sum_{i=1}^k w_i n_i \pi_i$.

3 Interval Estimation Procedures

This section provides detailed descriptions of five interval estimation procedures using the compound binomial error model. The intervals that result from these five procedures are the compound normal approximation interval, Mee interval, Haldane interval, Jeffreys–Perks interval, and score interval.

Construction of all five intervals except for the score interval is based on solving

$$(x - \tau)^2 \leq z_{\alpha/2}^2 \tilde{\vartheta}(X), \tag{2}$$

where x is a realized value of the weighted composite score X, $z_{\alpha/2}$ is the $(1 - \alpha/2)$–th quantile of the standard normal distribution, and $\tilde{\vartheta}(X)$ is any expression for the variance of X. The derivation of Equation 2 under the compound binomial error model relies on three properties of a normal distribution: (a) the log-likelihood function has a parabolic shape, (b) the maximum

likelihood estimator (MLE) for the mean is normally distributed, and (c) a linear combination of independent normal random variables is also normally distributed. Specifically, when the log-likelihood functions for all k domains are well-approximated by parabolas, $X_1, X_2, \cdots, X_k$, which are the MLEs for the means of the binomial distributions for the k domains, will be (approximately) normally distributed with mean $E(X_i) = n_i\pi_i$ and variance $\vartheta(X_i) = n_i\pi_i(1-\pi_i),\ i = 1, 2, \cdots, k$. Furthermore, because a linear combination of independent normal random variables has a normal distribution, the weighted composite score $X = \sum_{i=1}^{k} w_i X_i$ is also (approximately) normally distributed with mean $E(X) = \tau$ and variance $\vartheta(X) = \sum_{i=1}^{k} w_i^2 n_i \pi_i (1-\pi_i)$. Thus,

$$\begin{aligned} 1-\alpha &\approx P\left(-z_{\alpha/2} \leq \frac{X-\tau}{\sqrt{\vartheta(X)}} \leq z_{\alpha/2}\right) \\ &= P\left(X - z_{\alpha/2}\sqrt{\vartheta(X)} \leq \tau \leq X + z_{\alpha/2}\sqrt{\vartheta(X)}\right). \end{aligned} \tag{3}$$

After substituting a realized value $x = \sum_{i=1}^{k} w_i x_i$ in place of X and $\tilde{\vartheta}(X)$ in place of $\vartheta(X)$, we obtain Equation 2.

3.1 Compound Normal Approximation Interval

An interval can be obtained by using the expression $\tilde{\vartheta}(X) = \sum_{i=1}^{k} w_i^2 n_i \hat{\pi}_i (1 - \hat{\pi}_i)$, where $\hat{\pi}_i = x_i/n_i$ is the sample proportion for the ith stratified domain. With this expression, the $100(1-\alpha)\%$ interval estimate is given by

$$x \pm z_{\alpha/2}\sqrt{\sum_{i=1}^{k} w_i^2 n_i \hat{\pi}_i (1-\hat{\pi}_i)}. \tag{4}$$

For the purposes of this study, this interval will be referred to as the compound normal approximation interval because its construction is similar to the normal approximation interval under the binomial error model in the sense that both use sample proportions to estimate the variance of X. Note that the compound normal approximation interval is identical to the normal approximation interval using Feldt's conditional SEM (Brossman & Lee, 2009) when the domain weights are 1 and a bias correction factor $n_i/(n_i - 1)$ is multiplied to the variance term in Equation 4.

The compound normal approximation interval is attractive because it is easy to compute. However, it performs poorly when the number of items in each domain is small, in the sense that its coverage probability is far below the nominal level.

3.2 Mee Interval

For constructing intervals for the difference of two binomial proportions, i.e., $\Delta = \pi_2 - \pi_1$, Mee (1984) suggested using the MLEs for π_1 and π_2 under the

assumption that the value of Δ is known. Beal (1987) compared this procedure to the usual procedure that uses sample proportions and found that the former yields more satisfying results. Later, Decrouez and Robinson (2012) obtained similar results for the weighted sum of two binomial proportions. Even though previous studies only focused on two binomial proportions, the Mee procedure can be extended to more than two binomial proportions (i.e., stratified domains).

Denote $\hat{\pi}_i(\tau)$ as the MLE for π_i, $i = 1, 2, \cdots, k$, assuming that the true weighted composite score τ is known; that is,

$$\hat{\boldsymbol{\pi}}(\tau) = \underset{\boldsymbol{\pi}}{\operatorname{argmax}}\, \mathcal{L}(\tau, \boldsymbol{\pi}), \tag{5}$$

where $\hat{\boldsymbol{\pi}}(\tau) = (\hat{\pi}_1(\tau), \hat{\pi}_2(\tau), \cdots, \hat{\pi}_k(\tau))$, $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_k)$, $\mathcal{L}(\tau, \boldsymbol{\pi})$ is the log–likelihood function for the compound binomial distribution assuming that τ is known, and $\operatorname{argmax}_{\boldsymbol{\pi}} \mathcal{L}(\tau, \boldsymbol{\pi}) := \{\boldsymbol{\pi} \,|\, \forall \boldsymbol{\pi'} : \mathcal{L}(\tau, \boldsymbol{\pi'}) \leq \mathcal{L}(\tau, \boldsymbol{\pi})\}$. The Mee interval is constructed using $\tilde{\vartheta}(X) = \sum_{i=1}^{k} w_i^2 n_i \hat{\pi}_i(\tau)(1 - \hat{\pi}_i(\tau))$. It should be noted that the log–likelihood function needs to be maximized under the following constraints:

$$0 < \hat{\pi}_i(\tau) < 1, \quad i = 1, 2, \cdots, k, \tag{6}$$

and

$$\sum_{i=1}^{k} w_i n_i \hat{\pi}_i(\tau) = \tau. \tag{7}$$

One disadvantage of using the Mee interval is the lack of a closed-form solution. The endpoints need to be found with iterative computation using suitable starting values (e.g., endpoints of the compound normal approximation interval) and evaluation of $\hat{\boldsymbol{\pi}}(\tau)$ at each iteration.

3.3 Haldane and Jeffreys-Perks Intervals

The Haldane and Jeffreys-Perks interval estimation procedures were first discussed by Beal (1987) to construct intervals for the difference of two binomial proportions, and Decrouez and Robinson (2012) applied the two procedures to construct intervals for the weighted sum of two binomial proportions. Building upon earlier works on Haldane and Jeffreys-Perks intervals for two binomial proportions, more general versions of these two procedures are proposed here to obtain interval estimates for true weighted composite scores.

By defining a set of parameters $\{\psi_i\}_{i=2}^{k}$ as

$$\psi_i = \sum_{j=1,\, j\neq i}^{k} w_j n_j \pi_j - (k-1) w_i n_i \pi_i, \qquad i = 2, 3, \cdots, k, \tag{8}$$

domain true proportion-correct scores $\pi_1, \pi_2, \cdots, \pi_k$ can be expressed in terms of τ and $\{\psi_i\}_{i=2}^{k}$. This can be done by first subtracting each ψ_i from $\tau =$

$\sum_{i=1}^{k} w_i n_i \pi_i$ to obtain

$$\pi_i = \frac{\tau - \psi_i}{k w_i n_i}, \qquad i = 2, 3, \cdots, k, \tag{9}$$

and then substituting these expressions into τ to get

$$\pi_1 = \frac{\tau + \sum_{i=2}^{k} \psi_i}{k w_1 n_1}. \tag{10}$$

Using these expressions for $\pi_1, \pi_2, \cdots, \pi_k$, the variance of X can be expressed as a function of τ and $\{\psi\}_{i=2}^{k}$, say $\vartheta(\tau, \psi_2, \psi_3, \cdots, \psi_k) = a_2\tau^2 + a_1\tau + a_0$, where

$$a_2 = -\frac{1}{k^2} \sum_{i=1}^{k} \frac{1}{n_i} \tag{11}$$

$$a_1 = \frac{1}{k^2} \left[\frac{(k w_1 n_1 - 2\sum_{i=2}^{k} \psi_i)}{n_1} + \sum_{i=2}^{k} \frac{k w_i n_i + 2\psi_i}{n_i} \right] \tag{12}$$

$$a_0 = \frac{1}{k^2} \left[\frac{k w_1 n_1 \sum_{i=2}^{k} \psi_i - (\sum_{i=2}^{k} \psi_i)^2}{n_1} - \sum_{i=2}^{k} \frac{w_i n_i \psi_i + \psi_i^2}{n_i} \right]. \tag{13}$$

With $\tilde{\vartheta}(X) = \vartheta(\tau, \hat{\psi}_2, \hat{\psi}_3, \cdots, \hat{\psi}_k)$, where $\hat{\psi}_i = \sum_{j=1,\, j\neq i}^{k} w_j n_j \hat{\pi}_j - (k-1) w_i n_i \hat{\pi}_i$, Equation 2 becomes quadratic in τ, and it can be shown that the endpoints of the interval are

$$\frac{(2x + a_1 z^2) \pm \sqrt{(2x + a_1 z^2)^2 - 4(1 - a_2 z^2)(x^2 - a_0 z^2)}}{2(1 - a_2 z^2)}, \tag{14}$$

where $z = z_{\alpha/2}$. This interval is similar in construction to the Wilson score interval under the binomial error model. However, it should be noted that Equation 14 is not a score interval because its construction is not based on a score test.

Extending Beal's (1987) reasoning to k domains, a family of intervals that are closely related to the one in Equation 14 can be obtained by considering the posterior means of $\psi_i, i = 2, 3, \cdots, k$, using a prior density proportional to $\left[\prod_{i=1}^{k} \pi_i(1 - \pi_i)\right]^{\gamma}$ on $(\pi_1, \pi_2, \cdots, \pi_k)$ as alternatives to $\psi_i, i = 2, 3, \cdots, k$. Because the likelihood of a compound binomial distribution is the product of k binomial distributions with different true proportion-correct scores, the joint posterior distribution of $\pi_i, i = 1, 2, \cdots, k$, can be computed as follows:

$$\begin{aligned} h(\boldsymbol{\pi}|\mathbf{x}) &\propto f(\mathbf{x}|\boldsymbol{\pi}) h(\boldsymbol{\pi}) \\ &= \prod_{i=1}^{k} \binom{n_i}{x_i} \pi_i^{x_i} (1 - \pi_i)^{n_i - x_i} \left[\prod_{i=1}^{k} \pi_i (1 - \pi_i)\right]^{\gamma} \\ &= \prod_{i=1}^{k} \binom{n_i}{x_i} \pi_i^{x_i + \gamma} (1 - \pi_i)^{n_i - x_i + \gamma}, \end{aligned} \tag{15}$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_k)$ and $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_k)$. After somewhat tedious algebra, it can be shown that the posterior mean of ψ_i for each $i = 2, 3, \cdots, k$ is

$$E(\psi_i|\mathbf{x}) = \sum_{j=1,\, j\neq i}^{k} w_j n_j \frac{x_j + \gamma + 1}{n_j + 2\gamma + 2} - (k-1) w_i n_i \frac{x_i + \gamma + 1}{n_i + 2\gamma + 2}.$$

Various intervals can be obtained with different values of γ. In the present study, two intervals are considered. The first one is the interval obtained with $\gamma = -1$ and is referred to as the Haldane interval because $\left[\prod_{i=1}^{k} \pi_i(1-\pi_i)\right]^{-1}$ is the product of Haldane priors. Note that as $E(\psi_i|\mathbf{x}) = \psi_i$ when $\gamma = -1$, the interval obtained in Equation 14 is in fact the Haldane interval. Another interval of interest is when $\gamma = -0.5$. This interval is referred to as the Jeffreys-Perks interval because the prior with $\gamma = -0.5$ arises naturally from the Jeffreys-Perks invariance theory (Jeffreys, 1946; Perks, 1947).

3.4 Score Interval

The score interval can be constructed by inverting a score test:

$$\frac{S_\tau(\tau, \hat{\boldsymbol{\pi}}(\tau))^2}{\vartheta(S_\tau(\tau, \hat{\boldsymbol{\pi}}(\tau)))}, \tag{16}$$

where $S_\tau(\tau, \hat{\boldsymbol{\pi}}(\tau))$ is the score function for τ (i.e., the first derivative of the log–likelihood function with respect to τ) and the other symbols are the same as previously defined. It is known that the distribution of $S_\tau(\tau, \hat{\boldsymbol{\pi}}(\tau))$ is asymptotically normal with variance (Davison, 2003, p. 132)

$$\vartheta(S_\tau(\tau, \hat{\boldsymbol{\pi}}(\tau))) = I_{\tau\tau} - I_{\tau\boldsymbol{\pi}}[I_{\boldsymbol{\pi}\boldsymbol{\pi}}]^{-1} I_{\tau\boldsymbol{\pi}}^T, \tag{17}$$

where

$$I_{\tau\tau} = E\left(-\frac{\partial^2 \mathcal{L}(\tau, \hat{\boldsymbol{\pi}}(\tau))}{\partial \tau^2}\right), \tag{18}$$

$$I_{\tau\boldsymbol{\pi}} = E\left(-\frac{\partial^2 \mathcal{L}(\tau, \hat{\boldsymbol{\pi}}(\tau))}{\partial \tau \partial \boldsymbol{\pi}}\right), \tag{19}$$

$$I_{\boldsymbol{\pi}\boldsymbol{\pi}} = E\left(-\frac{\partial^2 \mathcal{L}(\theta, \hat{\boldsymbol{\pi}}(\tau))}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}^T}\right), \tag{20}$$

and A^T denotes the transpose of A. Thus, the lower and upper bounds of the score interval are the solutions to

$$\frac{S_\tau(\tau, \hat{\boldsymbol{\pi}}(\tau))^2}{\vartheta(S_\tau(\tau, \hat{\boldsymbol{\pi}}(\tau)))} = z_{\alpha/2}^2. \tag{21}$$

Because τ is unknown as the Mee interval, the score interval also has to be solved numerically.

4 Method

A simulation study was conducted to compare the compound normal approximation, Mee, Haldane, and Jeffreys-Perks intervals under the compound binomial error model. A total of 12 conditions were studied by fully crossing five factors, which are summarized in Table 1. Note that the score interval is not included in the simulation study. Decrouez and Robinson (2012) have shown algebraically that the score interval is identical to the Mee interval for the weighted sum of two binomial proportions. It turns out that this is also true for the weighted composite score of three stratified domains. The proof is provided in the Appendix.

Table 1: Summary of the study factors and conditions used for the simulation

Factor	Condition
Domain Size	3
Domain Weight	(1, 1, 1), (2, 2, 1)
Number of Items	(10, 10, 10), (10, 10, 20), (20, 20, 20)
Nominal Level	0.95
Correlation Between π_i's	0.7, 0.9

Correlated domain true proportion-correct scores for 1,000 simulees were generated from a negatively skewed beta distribution with shape parameters 3.4 and 1.9 (Lee et al., 2006) using the procedure presented by Cario and Nelson (1997):

1. A random sample of size 1,000 was generated from a three-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{0}$ and correlation matrix $\boldsymbol{\Sigma}$ whose off-diagonal elements were either $\rho = 0.7$ or 0.9. That is,

$$\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_{1000}} \sim N_3(\mathbf{0}, \boldsymbol{\Sigma}). \tag{22}$$

2. $F^{-1}[\Phi(y_{hi})]$ was computed for each element y_{hi} of $\mathbf{y_h}$, where h and i are indices for simulee and domain, respectively, Φ is the univariate standard normal cumulative distribution function, and F^{-1} is the inverse cumulative distribution function of the beta distribution with shape parameters 3.4 and 1.9.

3. The ith domain true proportion-correct score for simulee h was set to $F^{-1}[\Phi(y_{hi})]$.

Note that Lee et al. (2006) used a beta distribution with shape parameters 3.4 and 1.9 to generate overall true proportion-correct scores, whereas the same distribution was used here to generate true proportion-correct scores for each of the k domains. Once the domain true proportion-correct scores were generated, for each simulee, observed scores for the k domains were obtained by first randomly sampling $n = \sum_{i=1}^{k} n_i$ deviates from a uniform distribution between

0 and 1, and comparing these values to the simulee's domain true proportion-correct scores. If $\pi_i > u_{ij}$, where u_{ij} is the jth uniform deviate for the ith domain, then the response to that item was coded as one, and zero otherwise. Finally, the 0/1 responses across all the items within each domain were added to obtain the k observed scores. After generating data, the four interval estimation procedures considered in the present study were applied to the simulated data. This process was repeated 1,000 times for each simulee.

Five evaluation criteria based on Agresti and Coull (1998) and Decrouez and Robinson (2012) were used to compare the four intervals: (a) coverage probability, (b) interval width, (c) absolute mean distance from the nominal level, (d) proportion of coverage probabilities within 0.02 of the nominal level, and (e) proportion of coverage probabilities below 0.90. For simulee h, the coverage probability (C_h) and interval width (W_h) are defined as

$$C_h = \frac{1}{R}\sum_{r=1}^{R}\mathbf{1}_{(L_r(\mathbf{x}_h),\,U_r(\mathbf{x}_h))}(\tau_h), \tag{23}$$

$$W_h = \frac{1}{R}\sum_{r=1}^{R}(U_r(\mathbf{x}_h) - L_r(\mathbf{x}_h)), \tag{24}$$

where R $(= 1,000)$ is the number of replications, τ_h is the true weighted composite score for simulee h, $\mathbf{x}_h$ is the domain score vector of length k for simulee h, $U_r(\mathbf{x}_h)$ and $L_r(\mathbf{x}_h)$ are, respectively, the upper and lower bounds of the rth interval, and $\mathbf{1}_A(\tau)$ is an indicator variable that is one if $\tau \in A$ and zero otherwise. The absolute mean distance, proportion of coverage probabilities within 0.02 of 0.95, and proportion of coverages below 0.90 are given by

$$D = \frac{1}{N}\sum_{h=1}^{N}|C_h - 0.95|, \tag{25}$$

$$P = \frac{1}{N}\sum_{h=1}^{N}\mathbf{1}_{(0.93,0.97)}(C_h), \tag{26}$$

$$B = \frac{1}{N}\sum_{h=1}^{N}\mathbf{1}_{(0,0.90)}(C_h), \tag{27}$$

where N $(= 1,000)$ is the number of simulees.

5 Results

Tables 2 and 3 summarize the average coverage probabilities and interval widths averaged over 1,000 simulees, absolute mean distances from the nominal level, proportions of coverage probabilities within 0.02 of the nominal level, and proportion of coverage probabilities below 0.90 for raw-score (i.e., weighted composite score) intervals when the domain weights are $(w_1, w_2, w_3) = (1, 1, 1)$ and

$(2, 2, 1)$, respectively. In addition, coverage probabilities for all $1,000$ raw-score intervals (one for each simulee) for two sets of conditions are shown in Figures 1 and 2. (Note that coverage probability plots for the other conditions that are not provided here showed similar patterns to those in Figures 1 and 2.)

In general, as the domain size increased, the performance of all four interval estimation procedures improved. This result was somewhat expected because all four procedures are based on normal approximation to the binomial distribution (see Section 3), which is known to perform better with larger sample sizes. On the other hand, correlation between true domain proportion-correct scores slightly lowered coverage probabilities for the compound normal approximation intervals but had little, if any, impact on the performance of the other three procedures.

The compound normal approximation interval estimation procedure performed poorly and returned average coverage probabilities that were far below the nominal level of 0.95 for all 12 study conditions. It can be seen from Figures 1 and 2 that coverage probabilities fall below the nominal level throughout the entire score scale, especially at the extreme score points. The significant drop in the coverage probabilities near both ends of the score scale is a consequence of using the observed weighted composite score x as the midpoint for a highly skewed compound binomial distribution. As a result, the proportions of coverage probabilities between 0.93 and 0.97 were much lower than the other procedures. It should be noted that the poor results of the compound normal approximation intervals were not due to short interval widths.

Intervals obtained with the Haldane and Jeffreys-Perks procedures had average coverage probabilities that were close to the nominal level, with mean distances no larger than 0.01 for all twelve simulation conditions. In addition, all intervals had coverages greater than 0.90. As depicted in Figures 1 and 2, coverage probabilities were scattered near the nominal level 0.95 for most of the score points but tended to be slightly conservative at the extremes. Comparing the two, when the number of items for each domain was small, the Haldane intervals performed slightly worse than the Jeffreys-Perks intervals in terms of coverage probabilities and proportions of coverages between 0.93 and 0.97, but the performance improved and became comparable to the Jeffery-Perks intervals as the number of items increased. It is worth noting that, although the Haldane and Jeffreys-Perks intervals showed better results than the compound normal approximation intervals, the interval widths were shorter.

The Mee interval estimation procedure also showed good performance based on the five evaluation criteria. One exception was when each domain contained 10 items and had a weight of one for which the proportions of coverage probabilities within 0.02 of the nominal level were lower than 0.90. Also, coverage probabilities were a bit too large at the right end of the score scale. In terms of interval widths, the Mee procedure returned intervals that were slightly longer than the Haldane intervals but shorter than the Jeffreys-Perks intervals.

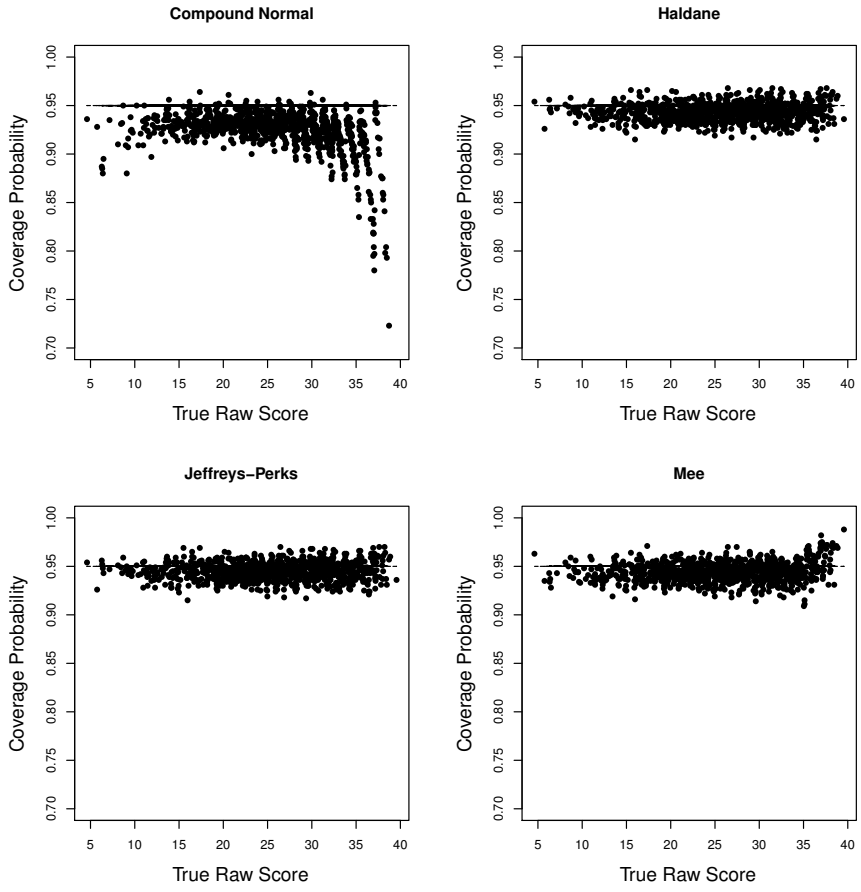
Table 2: Summary of the five evaluation criteria for $(w_1, w_2, w_3) = (1, 1, 1)$

Criteria	Compound Normal	Haldane	Jeffreys–Perks	Mee
$(n_1, n_2, n_3) = (10, 10, 10)$				
$\rho = 0.7$				
Coverage Probability	0.913	0.943	0.946	0.942
Interval Width	8.754	8.459	8.544	8.522
Mean Distance from 0.95	0.037	0.010	0.008	0.012
Coverage $\in (0.93, 0.97)$	0.235	0.907	0.943	0.867
Below 0.90	0.161	0.000	0.000	0.000
$\rho = 0.9$				
Coverage Probability	0.910	0.943	0.945	0.942
Interval Width	8.710	8.451	8.515	8.494
Mean Distance from 0.95	0.040	0.010	0.009	0.011
Coverage $\in (0.93, 0.97)$	0.258	0.912	0.926	0.843
Below 0.90	0.184	0.001	0.001	0.000
$(n_1, n_2, n_3) = (10, 10, 20)$				
$\rho = 0.7$				
Coverage Probability	0.923	0.943	0.945	0.944
Interval Width	10.263	9.939	10.007	10.037
Mean Distance from 0.95	0.027	0.009	0.008	0.009
Coverage $\in (0.93, 0.97)$	0.467	0.942	0.954	0.924
Below 0.90	0.086	0.000	0.000	0.000
$\rho = 0.9$				
Coverage Probability	0.920	0.943	0.945	0.944
Interval Width	10.225	9.931	9.979	10.009
Mean Distance from 0.95	0.030	0.009	0.009	0.009
Coverage $\in (0.93, 0.97)$	0.422	0.942	0.952	0.920
Below 0.90	0.105	0.000	0.000	0.000
$(n_1, n_2, n_3) = (20, 20, 20)$				
$\rho = 0.7$				
Coverage Probability	0.932	0.946	0.947	0.946
Interval Width	12.785	12.541	12.585	12.579
Mean Distance from 0.95	0.018	0.007	0.007	0.007
Coverage $\in (0.93, 0.97)$	0.692	0.975	0.976	0.966
Below 0.90	0.033	0.000	0.000	0.000
$\rho = 0.9$				
Coverage Probability	0.930	0.947	0.947	0.946
Interval Width	12.733	12.514	12.542	12.533
Mean Distance from 0.95	0.020	0.008	0.007	0.008
Coverage $\in (0.93, 0.97)$	0.647	0.963	0.965	0.949
Below 0.90	0.048	0.000	0.000	0.001

Table 3: Summary of the five evaluation criteria for $(w_1, w_2, w_3) = (2, 2, 1)$

Criteria	Compound Normal	Haldane	Jeffreys–Perks	Mee
$(n_1, n_2, n_3) = (10, 10, 10)$				
$\rho = 0.7$				
Coverage Probability	0.908	0.941	0.945	0.945
Interval Width	15.136	14.616	14.801	14.736
Mean Distance from 0.95	0.042	0.010	0.008	0.009
Coverage $\in (0.93, 0.97)$	0.109	0.915	0.961	0.951
Below 0.90	0.182	0.000	0.000	0.000
$\rho = 0.9$				
Coverage Probability	0.904	0.942	0.945	0.945
Interval Width	15.056	14.596	14.750	14.681
Mean Distance from 0.95	0.046	0.010	0.009	0.009
Coverage $\in (0.93, 0.97)$	0.106	0.916	0.946	0.945
Below 0.90	0.206	0.000	0.000	0.000
$(n_1, n_2, n_3) = (10, 10, 20)$				
$\rho = 0.7$				
Coverage Probability	0.915	0.943	0.946	0.944
Interval Width	16.085	15.628	15.794	15.694
Mean Distance from 0.95	0.035	0.009	0.007	0.009
Coverage $\in (0.93, 0.97)$	0.216	0.942	0.968	0.940
Below 0.90	0.122	0.000	0.000	0.000
$\rho = 0.9$				
Coverage Probability	0.912	0.944	0.947	0.944
Interval Width	16.008	15.605	15.740	15.640
Mean Distance from 0.95	0.038	0.008	0.007	0.009
Coverage $\in (0.93, 0.97)$	0.190	0.964	0.975	0.937
Below 0.90	0.162	0.001	0.000	0.000
$(n_1, n_2, n_3) = (20, 20, 20)$				
$\rho = 0.7$				
Coverage Probability	0.929	0.946	0.947	0.947
Interval Width	22.133	21.699	21.802	21.723
Mean Distance from 0.95	0.021	0.007	0.007	0.007
Coverage $\in (0.93, 0.97)$	0.620	0.978	0.981	0.983
Below 0.90	0.038	0.000	0.000	0.000
$\rho = 0.9$				
Coverage Probability	0.927	0.946	0.948	0.947
Interval Width	22.039	21.645	21.725	21.643
Mean Distance from 0.95	0.023	0.007	0.007	0.006
Coverage $\in (0.93, 0.97)$	0.594	0.974	0.979	0.983
Below 0.90	0.056	0.000	0.000	0.000

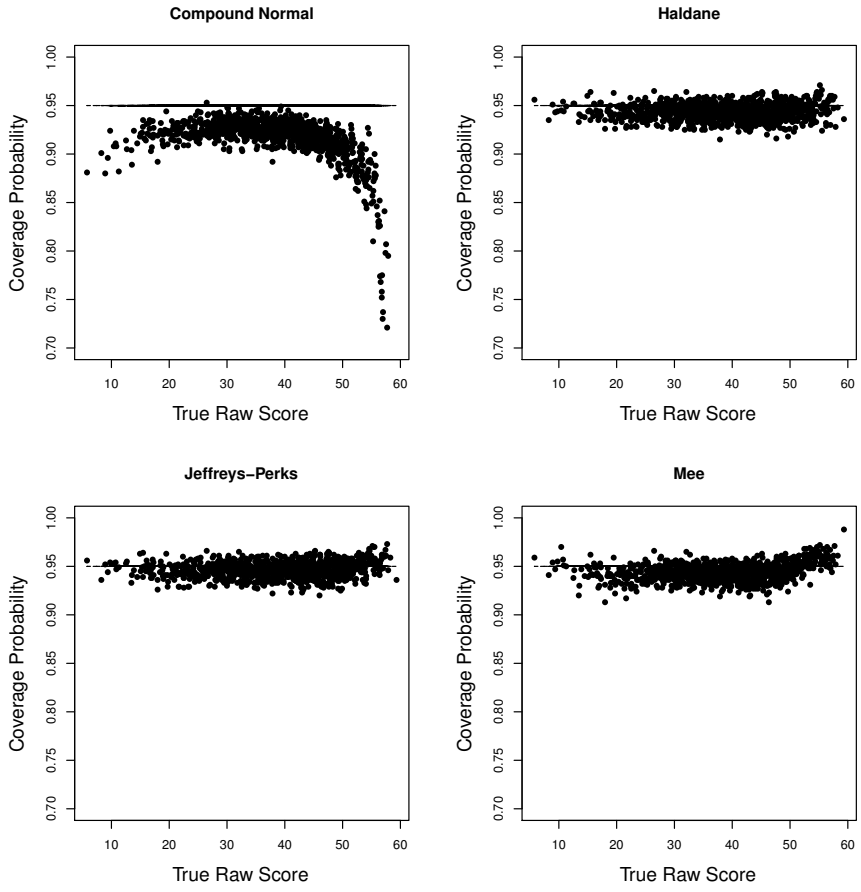
Figure 1: Coverage probabilities for true raw-score intervals when $\rho = 0.7$, $(n_1, n_2, n_3) = (10, 10, 20)$, and $(w_1, w_2, w_3) = (1, 1, 1)$



5.1 Real Data Application

A large-scale mathematics assessment with 54 multiple-choice items across three stratified domains was used to compare both raw-score (i.e., weighted composite score) and scale-score intervals obtained using the five aforementioned interval estimation procedures. In addition, for comparison purposes, the normal approximation and Wilson score intervals under the binomial error model were constructed by ignoring domain stratification and applying the two procedures directly to the weighted composite scores. The three domains contained 20, 18, and 16 items, respectively, and the domain weights were all fixed at 1. A raw-to-scale score conversion table with scale scores ranging from 20 to 80 with a one-point increment was used to obtain the scale-score intervals. More specif-

Figure 2: Coverage probabilities for true raw-score intervals when $\rho = 0.7$, $(n_1, n_2, n_3) = (10, 10, 20)$, and $(w_1, w_2, w_3) = (2, 2, 1)$



ically, scale score intervals were constructed by computing the scale scores that corresponded to the two endpoints of the raw-score intervals using linear interpolation. Two examinees who had the same weighted composite score but different domain scores were considered.

As can be seen from Table 4, interval estimation procedures using the compound binomial error model produced different intervals for the two examinees. Comparing the same procedure across the two examinees, longer intervals were returned for Examinee A than for Examinee B. This is mainly attributable to the distribution of errors associated with each examinee's true score. Note that the compound binomial error model assumes that errors of measurement for each stratified domain follow a binomial distribution. Because measurement

errors are the greatest when the true proportion-correct score is 0.5 under the binomial error model, Examinee A, whose domain proportion-correct scores are approximately 0.5 for all three domains, will have large measurement errors compared to Examinee B, whose domain proportion-correct scores are more towards the extremes. Taking domain information into account, interval estimation procedures using the compound binomial error model returned longer intervals for Examinee A than Examinee B. In contrast, the normal approximation and Wilson score procedures, which ignore domain stratification, assigned the same intervals to the two examinees.

It is worth noting that intervals constructed under the compound binomial error model and the binomial error model are more similar for Examinee A than Examinee B. This is because the magnitudes of measurement errors under the two models for Examinee A are expected to be more similar than Examinee B, as the domain and composite proportion-correct scores are both near 0.5. However, this is not the case for Examinee B whose composite score is near the middle, but whose domain scores are located near the ends.

6 Discussion

The present article introduced the compound normal approximation, Mee, Haldane, Jeffreys-Perks, and score interval estimation procedures using the compound binomial error model, and compared the five procedures through a simulation study. In addition, these five procedures were compared to the normal approximation and Wilson score procedures using the binomial error model with a real data set.

Overall, the Mee, Haldane, and Jeffreys-Perks intervals returned coverage probabilities that were close to the nominal level, with the Jeffreys-Perks intervals achieving better coverages. It is important to note that these good coverage probabilities were obtained with interval widths that were, on average, shorter than the compound normal approximation intervals. Furthermore, the proportions of coverage probabilities between 0.93 and 0.97 for these three intervals were consistently higher than those for the compound normal approximation intervals. Another attractive aspect of the Haldane and Jeffreys-Perks intervals is that they both have closed-form expressions. By constrast, there is no closed-form expression for the Mee interval and, therefore, the interval needs to be solved numerically. This is also the case for the score interval estimation procedure. Moreover, the complexity of computation for the Mee and score intervals increases with the increase of the number of stratified domains, which also makes these two procedure less practical than the Haldane and Jeffreys-Perks procedures. Comparing the Haldane and Jeffreys-Perks procedures, the latter returned intervals with better coverage probabilities and larger proportions of coverage probabilities near the nominal level.

Although the results are not reported in this article, for unequal domain weights, the normal approximation and Wilson score interval estimation procedures returned coverage probabilities that were far below the nominal level. The

Table 4: 95% interval estimates for a large–scale mathematics assessment

Interval	Raw Score	Scale Score	
		Unrounded	Rounded
Examinee A: 30 (10 / 10 / 10)			
Compound Binomial			
Compound Normal			
Interval	(22.881, 37.119)	(47.931, 59.091)	(48, 59)
Width	14.238	11.160	11
Haldane			
Interval	(22.904, 36.665)	(47.949, 58.725)	(48, 59)
Width	13.761	10.776	11
Jeffreys–Perks			
Interval	(22.901, 36.669)	(47.947, 58.729)	(48, 59)
Width	13.768	10.782	11
Mee			
Interval	(22.915, 36.695)	(47.957, 58.749)	(48, 59)
Width	13.780	10.792	11
Binomial			
Normal			
Interval	(22.843, 37.157)	(47.901, 59.121)	(48, 59)
Width	14.314	11.220	11
Wilson Score			
Interval	(22.883, 36.719)	(47.932, 58.769)	(48, 59)
Width	13.836	10.837	11
Examinee B: 30 (18 / 6 / 6)			
Compound Binomial			
Compound Normal			
Interval	(23.943, 36.057)	(48.766, 58.238)	(49, 58)
Width	12.114	9.472	9
Haldane			
Interval	(24.018, 35.723)	(48.826, 57.971)	(49, 58)
Width	11.705	9.145	9
Jeffreys–Perks			
Interval	(23.903, 35.831)	(48.735, 58.058)	(49, 58)
Width	11.928	9.323	9
Mee			
Interval	(23.993, 36.093)	(48.806, 58.268)	(49, 58)
Width	12.100	9.462	9
Binomial			
Normal			
Interval	(22.843, 37.157)	(47.901, 59.121)	(48, 59)
Width	14.314	11.220	11
Wilson Score			
Interval	(22.883, 36.719)	(47.932, 58.769)	(48, 59)
Width	13.836	10.837	11

main reason for the low converges is because these procedures use the weighted composite scores ignoring stratification, and unequal domain weights increase the difference between true and observed weighted composite scores when one of the domain observed scores is low compared to the domain true score. Suppose there is a 30–item test with three stratified domains, each having 10 items, and an examinee whose true domain proportion-correct scores are 0.6, 0.7, and 0.7. If the domain weights are 1 and this examinee scores 6, 3, and 7 on the three domains, the examinee's true and observed weighted composite scores are 20 and 16, respectively. In this case, an interval constructed under the binomial error model is likely to contain the examinee's true score because of the moderate difference between the true and observed scores (the two endpoints of the normal approximation interval is 9.54 and 22.46). However, when domain weights are 2, 2, and 1, the same examinee's true and observed weighted composite scores become 33 and 25, respectively, and because of the large difference between the two scores, it is less likely that the normal approximation interval will contain the true score (the two endpoints are 18.07 and 31.93). As mentioned previously, his large difference between the true and observed weighted composite scores is the consequence of only answering correctly to three items in domain 2 in spite of the high true domain proportion-correct score 0.7. Note that the opposite case also occurs but much less frequently, and as a result, the overall coverage probability decreases when the domain weights are not all 1.

Finally, even though the interval estimation procedures introduced in the present paper were only applied to three stratified domains, they can in theory be applied to any number of domains. When applying Haldane and Jeffreys-Perks interval estimation procedures to four stratified domains with 10 items each and equal domain weights, the coverage probabilities were 0.941 and 0.944, respectively, which are very similar to the results observed for the three stratified domain case.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126.

Beal, S. L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics*, *43*, 941-950.

Brossman, B. G., & Lee, W. (2009). *A comparison of confidence intervals and tolerance intervals for stratified domains under the compound binomial model.* Paper presented at the International Meeting of the Psychometric Society, Durham, N.H.

Cario, M. C., & Nelson, B. L. (1997). *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix* (Tech. Rep.). Citeseer.

Davison, A. C. (2003). *Statistical models* (Vol. 11). Cambridge University Press.

Decrouez, G., & Robinson, A. P. (2012). Confidence intervals for the weighted sum of two independent binomial proportions. *Australian & New Zealand Journal of Statistics*, *54*, 281-299.

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement*, *44*, 883-891.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* (pp. 453–461).

Kelley, T. L. (1927). *The interpretation of educational measurements.* New York: World Book.

Lee, W., Brennan, R. L., & Kolen, M. J. (2006). Interval estimation for true raw and scale scores under the binomial error model. *Journal of Educational and Behavioral Statistics*, *31*, 261-281.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*, 325-336.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, *17*, 510-521.

Mee, R. W. (1984). Confidence bounds for the difference between two probabilities. *Biometrics*, *40*, 1175-1176.

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries (1886-1994)*, *73*, 285–334.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the Americal Statistical Association*, *22*, 209-212.

Appendix

The proof that the Mee and score intervals are identical for the three stratified domain case is provided in this Appendix. This can be done by showing that

$$\frac{S_\tau(\tau, \hat{\boldsymbol{\pi}}(\tau))^2}{\vartheta(S_\tau(\tau, \hat{\boldsymbol{\pi}}(\tau)))} = \frac{(x-\tau)^2}{\sum_{i=1}^3 w_i n_i \hat{\pi}_i(\tau)}. \tag{28}$$

In order to make the computation somewhat simpler, let $\theta = \tau/(w_3 n_3)$, $a = (w_1 n_1)/(w_3 n_3)$, and $b = (w_2 n_2)/(w_3 n_3)$. Then $\tau = \sum_i^3 w_i n_i \pi_i$ becomes $\theta = a\pi_1 + b\pi_2 + \pi_3$, and Equation 28 can be written as

$$\frac{S_\theta(\theta, \hat{\boldsymbol{\pi}}(\theta))^2}{\vartheta(S_\theta(\theta, \hat{\boldsymbol{\pi}}(\theta)))} = \frac{(\hat{\theta}-\theta)^2}{a^2\hat{\pi}_1(\theta) + b^2\hat{\pi}_2(\theta) + \hat{\pi}_3(\theta)}, \tag{29}$$

where $\hat{\theta} = a(x_1/n_1) + b(x_2/n_2) + (x_3/n_3)$.

The log-likelihood function (ignoring the constant term) for the compound binomial distribution can be expressed in terms of θ, π_1, and π_2 as follows:

$$\begin{aligned} \mathcal{L}(\theta, \boldsymbol{\pi}) &= x_1 \log(\pi_1) + (n_1 - x_1)\log(1-\pi_1) \\ &+ x_2 \log(\pi_2) + (n_2 - x_2)\log(1-\pi_2) \\ &+ x_3 \log(\theta - a\pi_1 - b\pi_2) + (n_3 - x_3)\log(1 - \theta + a\pi_1 + b\pi_2), \end{aligned} \tag{30}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2)$. The score functions for θ, π_1, and π_2 are, respectively,

$$S_\theta(\theta, \boldsymbol{\pi}) = \frac{\partial \mathcal{L}(\theta, \boldsymbol{\pi})}{\partial \theta} = \frac{x_3 - n_3\pi_3}{\pi_3(1-\pi_3)}, \tag{31}$$

$$S_{\pi_1}(\theta, \boldsymbol{\pi}) = \frac{\partial \mathcal{L}(\theta, \boldsymbol{\pi})}{\partial \pi_1} = \frac{x_1 - n_1\pi_1}{\pi_1(1-\pi_1)} - a\frac{x_3 - n_3\pi_3}{\pi_3(1-\pi_3)}, \tag{32}$$

$$S_{\pi_2}(\theta, \boldsymbol{\pi}) = \frac{\partial \mathcal{L}(\theta, \boldsymbol{\pi})}{\partial \pi_2} = \frac{x_2 - n_2\pi_2}{\pi_2(1-\pi_2)} - b\frac{x_3 - n_3\pi_3}{\pi_3(1-\pi_3)}, \tag{33}$$

and the variances and covariances of the score functions are, respectively,

$$E\left(-\frac{\partial^2 \mathcal{L}(\theta, \boldsymbol{\pi})}{\partial \theta^2}\right) = \frac{n_3}{\pi_3(1-\pi_3)}, \tag{34}$$

$$E\left(-\frac{\partial^2 \mathcal{L}(\theta, \boldsymbol{\pi})}{\partial \theta \partial \boldsymbol{\pi}}\right) = \left(-a\frac{n_3}{\pi_3(1-\pi_3)}, -b\frac{n_3}{\pi_3(1-\pi_3)}\right), \tag{35}$$

$$E\left(-\frac{\partial^2 \mathcal{L}(\theta, \boldsymbol{\pi})}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}'}\right) = \begin{pmatrix} \frac{n_1}{\pi_1(1-\pi_1)} + a^2\frac{n_3}{\pi_3(1-\pi_3)} & ab\frac{n_3}{\pi_3(1-\pi_3)} \\ ab\frac{n_3}{\pi_3(1-\pi_3)} & \frac{n_2}{\pi_2(1-\pi_2)} + b^2\frac{n_3}{\pi_3(1-\pi_3)} \end{pmatrix}. \tag{36}$$

After somewhat tedious algebra, it can be shown that the left-hand side of

Equation 29 is

$$\begin{aligned}\frac{S_\theta(\theta,\hat{\boldsymbol{\pi}}(\theta))^2}{\vartheta(S_\theta(\theta,\hat{\boldsymbol{\pi}}(\theta)))} &= \left[\frac{x_3 - n_3\hat{\pi}_3(\theta)}{\hat{\pi}_3(\theta)(1-\hat{\pi}_3(\theta))}\right]^2 \times \left[a^2\frac{\hat{\pi}_1(\theta)(1-\hat{\pi}_1(\theta))}{n_1}\right. \\ &+ \left. b^2\frac{\hat{\pi}_2(\theta)(1-\hat{\pi}_2(\theta))}{n_2} + \frac{\hat{\pi}_3(\theta)(1-\hat{\pi}_3(\theta))}{n_3}\right]. \qquad (37)\end{aligned}$$

Because $\hat{\pi}_1(\theta)$ and $\hat{\pi}_2(\theta)$ are the solutions for Equations 32 and 33, respectively, we get (after some algebra)

$$a\frac{x_1}{n_1} - a\hat{\pi}_1(\theta) = a^2\frac{\hat{\pi}_1(\theta)(1-\hat{\pi}_1(\theta))(x_3 - n_3\hat{\pi}_3(\theta))}{n_1\hat{\pi}_3(\theta)(1-\hat{\pi}_3(\theta))}, \qquad (38)$$

$$b\frac{x_2}{n_2} - b\hat{\pi}_2(\theta) = b^2\frac{\hat{\pi}_2(\theta)(1-\hat{\pi}_2(\theta))(x_3 - n_3\hat{\pi}_3(\theta))}{n_2\hat{\pi}_3(\theta)(1-\hat{\pi}_3(\theta))}, \qquad (39)$$

and adding Equations 38 and 39 results in

$$\begin{aligned}a\frac{x_1}{n_1} + b\frac{x_2}{n_2} - \theta &= -\hat{\pi}_3(\theta) + a^2\frac{\hat{\pi}_1(\theta)(1-\hat{\pi}_1(\theta))(x_3 - n_3\hat{\pi}_3(\theta))}{n_1\hat{\pi}_3(\theta)(1-\hat{\pi}_3(\theta))} \\ &+ b^2\frac{\hat{\pi}_2(\theta)(1-\hat{\pi}_2(\theta))(x_3 - n_3\hat{\pi}_3(\theta))}{n_2\hat{\pi}_3(\theta)(1-\hat{\pi}_3(\theta))} \qquad (40)\end{aligned}$$

(note that $a\hat{\pi}_1(\theta) + b\hat{\pi}_2(\theta) = \theta - \hat{\pi}_3(\theta)$). Finally, substituting Equation 40 into the numerator of the right-hand side of Equation 29 and simplifying the expression yields

$$\begin{aligned}\frac{(\hat{\theta}-\theta)^2}{a^2\hat{\pi}_1(\tau) + b^2\hat{\pi}_2(\tau) + \hat{\pi}_3(\tau)} &= \left[\frac{x_3 - n_3\hat{\pi}_3(\theta)}{\hat{\pi}_3(\theta)(1-\hat{\pi}_3(\theta))}\right]^2 \times \left[a^2\frac{\hat{\pi}_1(\theta)(1-\hat{\pi}_1(\theta))}{n_1}\right. \\ &+ \left. b^2\frac{\hat{\pi}_2(\theta)(1-\hat{\pi}_2(\theta))}{n_2} + \frac{\hat{\pi}_3(\theta)(1-\hat{\pi}_3(\theta))}{n_3}\right], \qquad (41)\end{aligned}$$

which is identical to Equation 37.