Research Report

Improved Reliability Estimates for Small Samples Using Empirical Bayes Techniques

Hyeonjoo J. Oh

Hongwen Guo

Michael E. Walker

December 2009

ETS RR-09-46

Improved Reliability Estimates for Small Samples Using Empirical Bayes Techniques

Hyeonjoo J. Oh, Hongwen Guo, and Michael E. Walker
ETS, Princeton, New Jersey

December 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Issues of equity and fairness across subgroups of the population (e.g., gender or ethnicity) must be seriously considered in any standardized testing program. For this reason, many testing programs require some means for assessing test characteristics, such as reliability, for subgroups of the population. However, often only small sample sizes are available for the subgroups of interest. Traditionally used reliability estimates (e.g., Cronbach's alpha) can have low precision for small samples. This study investigated whether an empirical Bayes (EB) technique could produce more precise reliability estimates than traditional methods in the presence of small samples. Several Bayesian estimates were compared to estimates obtained by other methods (e.g., the traditionally and currently used Cronbach's alpha coefficient), in terms of both bias and variance. A secondary purpose of this study was to compare the various EB approaches across different sample sizes. This paper also discusses EB estimates of standard error of measurement (SEM), their accuracy and precision, and how they compare with SEM estimates derived from the alpha.

Key words: Empirical Bayes, reliability, standard error of measurement, prior distribution

Table of Contents

List of Tables

List of Figures

Page

Overview and Background

In an educational environment increasingly influenced by standardized testing, issues of equity and fairness across subgroups of the population (e.g., defined by ethnicity or disability status) must be seriously considered. An adequate assessment of the characteristics of a test in these subgroups becomes essential. Failure to assess the reliability of a test for a subgroup might result in the use of an unreliable instrument in high-stakes decision making (e.g., program admission or class placement). Alternatively, imprecise measurement of reliability for subgroups might result in the exclusion of a beneficial test from consideration as an assessment tool. Either one of these possibilities could have highly detrimental effects (e.g., increased social barriers for already disadvantaged groups). Many areas of this important issue remain to be addressed.

Although estimation of subgroup reliability might seem to be a straightforward or even trivial matter, it presents several problems. Most notable is that the subsamples from the population may be quite small. This small sample size could necessarily have an effect on the quality of the reliability estimates. Walker and Zhang (2004) suggested a minimum sample size of 125 to 150 for calculating reliability, with at least as many people in the sample as items on the test. Another problem involves score range restrictions as a result of subsampling. Reliability of a test can be affected by changes in group heterogeneity or by systematic selection of scores (e.g., subsampling) because observed variances can be different in the selected subgroup and total group (Allen & Yen, 1979). These score range restrictions can attenuate the estimate of reliability, possibly leading to erroneous conclusions about the adequacy of the test for the subpopulation in question.

Accordingly, research on estimating reliability with small samples is necessary. In the current study, empirical Bayes (EB) techniques were applied to the estimation procedure, integrating collateral information (e.g., reliability information on the same test from different subpopulations) to improve the accuracy of reliability estimates. This study investigated whether the EB-based reliability estimates improved the precision for estimating reliability of subgroups of a population, even for very small subsamples. It also compared the various EB approaches across different sample sizes.

No studies to our knowledge directly employ EB techniques to estimate reliability, although the use of EB techniques to improve the accuracy of estimates has a long history. Several researchers have investigated the use of EB techniques. Braun and Jones (1984) studied

the validity of academic predictors of graduate school performance using EB methods and found that the EB coefficients provided a useful way of combining information across subsamples. The researchers reported that the EB models not only yielded better predictions of first-year grade averages than cluster analysis estimates, but also facilitated the accurate assessment of the quality of these predictions. Moreover, the prediction equations were quite stable and rarely displayed implausible features such as negative weights (Braun & Jones). As Braun and Jones combined information across subsamples, the current study also used reliability information across subsamples to estimate the target group reliability.

Edwards and Vevea (2006) examined a subscore augmentation procedure using EB adjustments to improve the "overall accuracy" of measurement when information is limited. In a situation where tests originally designed for one purpose (e.g., producing a reliable overall score to rank examinees) are frequently being pressed into service for other purposes (e.g., providing subscores specific to a narrow content area as diagnostic information), an overall score on the test may be reliable; however, such a test may not provide reliable diagnostic information (or subscores) because the subscores are based on less information than the overall score. Edwards and Vevea investigated the feasibility of increasing the reliability of diagnostic subscores by incorporating information from the rest of the test. They found that the subscores produced by the EB augmentation procedure represented an overall improvement over nonaugmented subscores. The magnitude of the improvement gained was a function of the correlation among subscales and subscale length (reliability). A main focus of the study by Edwards and Vevea was to investigate reliability of diagnostic subscores (not the total test reliability), while for the current study the main focus was to estimate total test reliability and standard error of measurement (SEM) using subgroup information.

Bayesian techniques have been applied to differential item functioning (DIF) analysis. Zwick and Thayer (2002) applied EB methods to DIF analysis and found the EB estimate of DIF to be an improvement over the traditionally used Mantel-Haenszel delta-DIF (MH D-DIF) statistic. Sinharay, Dorans, Grant, Blew, and Knorr (2006) investigated a full Bayesian (FB) approach to small-sample DIF estimation using the 10 least recent administrations of the Praxis°I®: Pre-Professional Skills Test (PPST®). They reported that the FB approach performed better than the existing MH D-DIF for small samples, but the gain was not substantial.

Empirical Bayes techniques have also been proposed to estimate equating relationships with small numbers of test takers to improve the stability and accuracy" of equated scores in the target population (Livingston & Lewis, 2009, p. 1). The study proposed to estimate the equated scores separately at each score point, incorporating relevant prior information into the estimation process. This approach is very innovative and has some advantages, but it also has two limitations. First, it is not symmetric with respect to the new form and reference form. Second, in some situations the proposed procedure can produce a result that is less accurate than that provided by the current equating when the difficulty of the current form to be equated is significantly different from the difficulties of the forms used in prior equatings.

One advantage of the EB statistical methods over traditional frequentist methods is that the Bayesian methods can incorporate existing collateral (or prior) information into the inference problem and lead to improved estimation, especially for small samples. In this context, EB methodology involves the simultaneous estimation of parameters in several samples. The combined information from other samples is used to adjust the parameter estimate for a given sample in order to make it more precise. Empirical Bayes methodology adjusts the estimate more or less depending upon the precision of the estimates obtained from the other samples. The success of this procedure depends, among other things, on the strength of the relationships among the various samples. The primary purpose of this study is to investigate whether the EB-based reliability estimates improve the precision for estimating reliability of subgroups of a population, even from very small subsamples ranging from 25 to 250, by comparing the reliability estimates to Cronbach's alpha coefficient. The secondary purpose of this study is to compare the various EB approaches across different sample sizes. This paper also discusses EB estimates of the SEM, their accuracy and precision, and how they compare with traditional SEM estimates.

The following section briefly describes the Bayes and EB approaches.

The Bayes Approach

In general, the Bayesian analysis is a methodology to model and simulate the behavior of discrete events under uncertainty using past experience, data, or convenient assumptions in the form of a prior distribution (Brandel, 2004). Unlike the frequentist approach, probability in Bayesian statistics is not defined as the frequency of the occurrence of an event but as the

plausibility that a statement is true given the information (Botje, 2006). The basic assumption of Bayes methodology is that a state of variables to be modeled and simulated can be represented by probability functions (discrete variables) or probability density functions (continuous variables).

In the case of the Bayes theorem for *discrete* variables, consider a set of observed data $\mathbf{y} = (y_1, \ldots, y_n)$ with some probability distribution $\mathrm{f}(\mathbf{y} \mid \theta)$ and an associated vector θ of unknown parameters. Suppose that θ is also a random vector, having a prior distribution $\mathrm{g}(\theta \mid \eta)$, where η is a vector of hyperparameters. Here θ and η are assumed to be discrete variables. The Bayes formula is used to compute the posterior distribution $\mathrm{p}(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\eta})$ for discrete parameter θ:

$$\mathrm{p}(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\eta}) = \frac{\mathrm{f}(\mathbf{y} \mid \boldsymbol{\theta}) \mathrm{g}(\boldsymbol{\theta} \mid \boldsymbol{\eta})}{\sum_{\theta} \mathrm{f}(\mathbf{y} \mid \boldsymbol{\theta}) \mathrm{g}(\boldsymbol{\theta} \mid \boldsymbol{\eta})}. \tag{1}$$

In the case of the Bayes theorem for *continuous* variables, let $\mathbf{y} = (y_1, \ldots, y_n)$ denote a sample from a probability density function by a continuous parameter θ, with prior density distribution $\mathrm{g}(\theta \mid \eta)$. Then the posterior density distribution for θ is given by

$$\mathrm{p}(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\eta}) = \frac{\mathrm{f}(\mathbf{y} \mid \boldsymbol{\theta}) \mathrm{g}(\boldsymbol{\theta} \mid \boldsymbol{\eta})}{\int \mathrm{f}(\mathbf{y} \mid \boldsymbol{\theta}) \mathrm{g}(\boldsymbol{\theta} \mid \boldsymbol{\eta}) d\boldsymbol{\theta}}. \tag{2}$$

If one is unsure of the value of η, then a proper Bayesian approach would adopt a hyperprior distribution $\mathrm{h}(\boldsymbol{\eta})$. Then the posterior distribution for θ is obtained by integrating the conditional density function with respect to η as well:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int f(\mathbf{y} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta} \mid \boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta}}{\iint f(\mathbf{y} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta} \mid \boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta} d\boldsymbol{\theta}}. \quad (3)$$

The Empirical Bayes Approach

As an alternative to (3), simply replace η with an estimate $\hat{\eta}$ that maximizes the marginal distribution $m(\mathbf{y} \mid \boldsymbol{\eta})$:

$$m(\mathbf{y} \mid \boldsymbol{\eta}) = \int f(\mathbf{y} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta} \mid \boldsymbol{\eta}) d\boldsymbol{\theta}. \quad (4)$$

This estimate $\hat{\eta}$ is then used as a known quantity in (2). Consider, for example, the case in which $f(\mathbf{y} \mid \theta)$ is a normal distribution with mean θ and known variance σ^2. Let $g(\theta \mid \eta)$ also be a normal distribution with hyperparameters $\boldsymbol{\eta} = N(\mu, \tau^2)$. When the hyperparameters are known, or when estimates $\hat{\eta}$ are obtained from the data, then derivation of the posterior distribution for θ is as follows:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)}$$

$$= \frac{f(y \mid \theta) g(\theta)}{p(y)}, \quad (5)$$

when solving the above equation based on the density function of the normal distribution,[1]

$$p(\theta \mid y) = \frac{\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\theta)^2}{2\sigma^2}}\right)\left(\frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}\right)}{p(y)},$$

$$= \frac{\frac{1}{2\pi\sigma\tau} e^{\left\{-\frac{(\sigma^2+\tau^2)}{2\sigma^2\tau^2}\cdot\left(\theta-\frac{\tau^2\mu+\sigma^2 y}{\sigma^2+\tau^2}\right)^2\right\}}}{p(y)}.$$

(6)

Therefore, $p(\theta \mid y)$ is a normal distribution with variance of $\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$ and mean of $\frac{\sigma^2\mu+\tau^2 y}{\sigma^2+\tau^2}$ derived from equation (6).

That is, the posterior distribution for θ is

$$p(\theta \mid y) = N\left(\theta \mid \frac{\sigma^2\mu+\tau^2 y}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right).$$

(7)

Let $B = \frac{\sigma^2}{\sigma^2+\tau^2}$. Then the posterior distribution has mean $B\mu + (1-B)y$ and variance $B\tau^2 = (1-B)\sigma^2$. In this sense, *B* is a weighting factor that is positively proportional to σ^2 but inversely proportional to τ^2.

Methods

Data

Data were selected from the Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT®) program with a yearly population size of over three million. The

PSAT/NMSQT measures developed critical reading skills, math problem-solving skills, and writing skills, which are related to successful performance in college. The PSAT/NMSQT critical reading, math, and writing skills sections are shortened versions of the College Board SAT Reasoning Test™ and measure the same abilities. The scores for all three measures are reported on a two-digit 20-to-80 scale. To meet the purposes of this study, 19 states and Washington, DC,[2] were selected as subgroups by varying population size (e.g., small to large) and population reliability (e.g., low to high). Four sample sizes (i.e., N = 25, 50, 125, and 250) were examined, and these samples were randomly selected from the population for each selected state.

Two hundred replications (or resampling) of these samples of 25, 50, 125, and 250 examinees were conducted for each state. The smallest sample was fairly small relative to the number of items on the test (48 items for critical reading, 38 items for math, and 39 items for writing), whereas the largest sample size of 250 represented a sufficient sample size for reliability estimation according to previous research (Walker & Zhang, 2004). Descriptive statistics for 19 states and Washington, DC, are displayed in Table 1 for each measure (i.e., critical reading, math, and writing).

Research Design

Test scores, from which reliability can be computed, were available for each person. To aid in estimation, we assumed that the reliability coefficient for each state was randomly sampled from some distribution with unknown parameter θ of states' reliability coefficients, while θ had a distribution with parameter η. We estimated the parameters of this distribution using the data from the states in the study. Using Bayes' formula and taking this distribution as the collateral distribution, we obtained the posterior distribution (3) of the reliability coefficient for any given subgroup (i.e., state). The mean of this posterior distribution became the EB reliability estimate for the subgroup of interest. In practice, the EB estimate can be seen as the subgroup estimate that has shrunk so that it is closer to the estimate of the common parameter across all subgroups. The amount of shrinkage depends upon the relative precision of the subgroup and common mean estimates.

Table 1

Comparison of Population Reliabilities and Standard Error of Measurement for 20 States

		Critical reading		Math		Writing	
State	*N*	State reliability	State SEM	State reliability	State SEM	State reliability	State SEM
AL	24,117	0.883	3.500	0.901	2.869	0.872	3.243
AZ	19,415	0.883	3.529	0.899	2.905	0.871	3.297
AK	11,821	0.881	3.497	0.892	2.883	0.868	3.258
CT	**36,180**	**0.897**	**3.506**	**0.913**	**2.900**	**0.889**	**3.283**
DE	17,877	0.894	3.511	0.904	2.927	0.884	3.283
W. D.C.	11,387	0.911	3.509	0.915	2.868	0.901	3.220
ID	4,262	0.875	3.462	0.884	2.841	0.853	3.264
IA	9,954	0.865	3.443	0.880	2.803	0.837	3.225
KT	20,292	0.879	3.477	0.896	2.859	0.867	3.257
LA	**15,695**	**0.876**	**3.496**	**0.890**	**2.866**	**0.865**	**3.227**
ME	26,375	0.879	3.509	0.893	2.919	0.869	3.322
MI	29,945	0.878	3.499	0.895	2.884	0.861	3.282
MO	5,891	0.867	3.503	0.881	2.873	0.846	3.292
NE	6,838	0.860	3.508	0.879	2.871	0.839	3.293
ND	2,537	0.852	3.466	0.877	2.815	0.833	3.270
SD	3,217	0.853	3.471	0.863	2.835	0.823	3.258
TN	34,121	0.900	3.482	0.913	2.861	0.889	3.230
WA	30,878	0.898	3.467	0.904	2.894	0.882	3.274
WV	6,160	0.866	3.515	0.885	2.907	0.853	3.248
WY	**2,226**	**0.864**	**3.506**	**0.876**	**2.897**	**0.841**	**3.287**
Mean		0.878	3.493	0.892	2.874	0.862	3.266
SD		0.016	0.022	0.014	0.033	0.021	0.028
Min	2,226	0.852	3.443	0.863	2.803	0.823	3.220
Max	36,180	0.911	3.529	0.915	2.927	0.901	3.322

Note. Statistics in bold are the three states with large, medium, and small population size (i.e., Connecticut, Louisiana, and Wyoming) to illustrate how population size affects reliability and the standard error of measurement (SEM).

Criterion for Comparison

The population reliability coefficient and population SEM for each state were used as criteria in this study.

Empirical Bayes Reliability Estimators

A total of four EB-based reliability coefficients were estimated in (9) through (13). An uncorrected (UC) reliability estimate was determined from Cronbach's alpha coefficient calculated directly from each sample. A mean of the 200 replications of the reliability coefficients using EB methods were computed. Then, the mean of each of the four EB reliability estimates and the one UC reliability estimate were compared to the population reliability (criterion) for each state.

Reliability estimates were obtained using UC, and EB methods. All estimates were based on Cronbach's alpha coefficient:

$$\alpha = \frac{k}{k-1}\left[\frac{\sum_{i=1}^{k}\sigma_i^2}{\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma_{ij}}\right], \tag{8}$$

where *k* is the number of items in the test form, and σ_i^2 and σ_{ij} are the variance and covariance of the item scores, respectively. The estimates were compared to the actual test reliability of the population for each state. Because the procedure was repeated on multiple samples (i.e., 200 replications), standard errors and bias estimates were obtained. The average squared bias, average variance, and root mean squared error (RMSE) for the EB and conventional methods were compared across sample sizes. Four different EB approaches were tried.

Empirical Bayes Reliability Approach 1. For this approach, EB reliability was estimated using the following equation, appropriate for normally distributed estimators:

$$\alpha_{EB1}(R_iS_j) = \frac{\frac{1}{\sigma^2_{R_iS}}\mu_{R_iS} + \frac{1}{\sigma^2_{RS_j}}\alpha_{R_iS_j}}{\frac{1}{\sigma^2_{R_iS}} + \frac{1}{\sigma^2_{RS_j}}}, \quad (9)$$

where $\sigma^2_{R_iS}$ and μ_{R_iS} are the variance and mean of the reliability estimates for all states in the study for the *i*th replication (R_i); $\sigma^2_{RS_j}$ is the variance of State j for all 200 replications; and $\alpha_{R_iS_j}(=\alpha_{rs})$ is the reliability estimate for State *j* (S_j) at the *i*th replication (R_i). See Table 2 for an overview of the design.

Empirical Bayes Reliability Approach 2. The reliability estimates α_i are likely not normally distributed. It may, however, be possible to transform them to approximate normality. One could, for example, apply Fisher's z-transformation through $Z'(\alpha_{rs}) = \frac{1}{2}\ln\left(\frac{1+\alpha_{rs}}{1-\alpha_{rs}}\right)$ (C. Lewis, personal communication, September 12, 2006). Once the transformation was carried out, (9) can be applied and the results translated back to the original metric via

$$\alpha_{EB2_{rs}} = \frac{\exp\{2z'(\alpha_{rs})\}-1}{\exp\{2z'(\alpha_{rs})\}+1} \quad (10)$$

Empirical Bayes (EB) Reliability Approach *3.* This approach is very similar to EB Reliability Approach 2, except for the treatment of α_{rs} in the estimation of Z'. If a reliability coefficient is considered to be a squared correlation between observed and true score, then the square roots of the reliabilities should be found and then Fisher's z-transformation should be used:

$$Z'(\alpha_{rs}) = \frac{1}{2}\ln\left(\frac{1+\sqrt{\alpha_{rs}}}{1-\sqrt{\alpha_{rs}}}\right). \quad (11)$$

Once the transformation was carried out, (9) was applied and the results were translated back to the original metric via

$$\alpha_{EB3_{rs}} = \left(\frac{\exp\left\{2z'\left(\alpha_{rs}\right)\right\} - 1}{\exp\left\{2z'\left(\alpha_{rs}\right)\right\} + 1} \right)^2 . \quad (12)$$

Empirical Bayes (EB) Reliability Approach 4. In EB Reliability Approach 4, the reliability was estimated using the EB SEM formula,

$$\alpha_{EB4_{rs}} = 1 - \frac{SEM^2_{EB1_{rs}}}{SD^2_{rs}} , \quad (13)$$

where $SEM_{EB1_{rs}}$ is based on (15) in the next section, and $SD_{rs} (= SD_{R_i S_j})$ is the standard deviation of the score at the *i*th replication (R_i) for State j (S_j).

Table 2

Layout of the Empirical Bayes Approaches

Replications	States										
	1	2	3	.	.	.	.	19	20	Mean	Variance
1	$\alpha_{R_1S_1}$	$\alpha_{R_1S_2}$	$\alpha_{R_1S_3}$	.	.	.	.	$\alpha_{R_1S_{19}}$	$\alpha_{R_1S_{20}}$	$\mu_{\alpha_{R_1S}}$	$\sigma^2_{\alpha_{R_1S}}$
2	$\alpha_{R_2S_1}$										
3	$\alpha_{R_3S_1}$										
.	.										
.	.										
.	.										
.	.										
199	$\alpha_{R_{199}S_1}$										
200	$\alpha_{R_{200}S_1}$	.	.	.	.	.	.	.	$\alpha_{R_{200}S_{20}}$	$\mu_{\alpha_{R_iS_j}}$	$\sigma^2_{\alpha_{R_iS_j}}$
Mean	$\mu_{\alpha_{R_iS_1}}$	.	.	.	.	.	.	.	$\mu_{\alpha_{R_iS_{20}}}$		
Variance	$\sigma^2_{\alpha_{R_iS_1}}$	.	.	.	.	.	.	.	$\sigma^2_{\alpha_{R_iS_{20}}}$		

Empirical Bayes Standard Error of Measurement

Similarly, a total of four EB-based SEM estimates were estimated in (15) through (18) and a mean of the 200 replications were computed. Then, the mean of each of the four EB reliability estimates and one UC reliability estimate were compared to the population SEM (criterion) for each state.

The current study evaluated empirically the accuracy of reliability estimates computed by EB approaches, as well as the traditionally used Cronbach's alpha coefficient (UC reliability estimate). While estimating reliability based on the EB approaches is an attractive approach with small sample groups, a potential danger exists when one group is large and the others small. In this case, the estimates for the small groups will be regressed toward the estimate for the total group, which is dominated by the largest group. The resulting values may give the misleading impression that the groups are similar because their EB estimates are similar. Another concern is that reliability is sensitive to the degree of heterogeneity within a group. In general, the more heterogeneous the group, the higher the reliability (Nunnally & Bernstein, 1994, p. 261).[3] Thus it may not be reasonable to try to achieve similar reliability from group to group. Instead, a quantity less sensitive to heterogeneity/homogeneity, such as the SEM, may be a more appropriate parameter to estimate using EB methods (Charles Lewis, personal communication, October 6, 2006; Lord & Novick, 1968). Therefore, four different EB approaches for SEM were also tried.

Empirical Bayes standard error of measurement (EB SEM) Approach 1. This approach estimated SEM and EB SEM as follows:

$$SEM = SD_{R_iS_j}\sqrt{1-\alpha_{R_iS_j}}\ , \tag{14}$$

where $SD_{R_iS_j}$ is the standard deviation of the score for State j (s_j) at the *i*th replication (R_i), and $\alpha_{R_iS_j}$ is the corresponding reliability:

$$SEM_{EB1}(R_iS_j) = \frac{\frac{1}{\sigma^2_{R_iS}}\mu_{R_iS} + \frac{1}{\sigma^2_{RS_j}}SEM_{R_iS_j}}{\frac{1}{\sigma^2_{R_iS}} + \frac{1}{\sigma^2_{RS_j}}}\ , \tag{15}$$

where $\sigma^2_{R_iS}$ and μ_{R_iS} are the variance and mean of SEM for the states at the *i*th replication (R_i) respectively; $\sigma^2_{RS_j}$ is the variance of SEM for State j for all 200 replications; and $SEM_{R_iS_j}$ is SEM for State j (S_j) at the *i*th replication (R_i).

Empirical Bayes standard error of measurement (EB SEM) Approach 2. This approach was estimated using the EB-based reliability described in EB Reliability Approach 1. The formula for EB SEM Approach 2 is defined as follows:

$$SEM_{EB2_{rs}} = SD_{rs}\sqrt{1-\alpha_{EB1_{rs}}}\ , \tag{16}$$

Where $\alpha_{EB1_{rs}}$ is based on (9).

Empirical Bayes standard error of measurement (EB SEM) Approach 3. This approach estimated SEM by employing the EB reliability with z-transformation used for the EB Reliability Approach 2:

$$SEM_{EB3_{rs}} = SD_{rs}\sqrt{1-\alpha_{EB2_{rs}}}\ , \tag{17}$$

where $\alpha_{EB2_{rs}}$ is based on (10).

Empirical Bayes Standard Error of Measurement (EB SEM) Approach 4. This approach 4 estimated SEM by employing the reliability from EB Reliability Approach 3:

$$SEM_{EB4_{rs}} = SD_{rs}\sqrt{1-\alpha_{EB3_{rs}}}\ , \tag{18}$$

where $\alpha_{EB3_{rs}}$ is based on (12).

Evaluation Indexes

The current study compared the estimated EB reliabilities and EB SEMs with each state's population reliabilities and SEMs using average squared bias, average variance, and RMSE. Average squared bias for reliability is equivalent to the sum of the average squared mean differences between the EB-based reliabilities (or uncorrected reliabilities) and population reliabilities for the states, divided by the number of states (i.e., 20). That is,

$$Average\ Squared\ Bias = \frac{\sum_{j=1}^{20}\left[\sum_{i=1}^{200}\left[\hat{\rho}_{i(S_j)} - \rho_{i(S_j)}\right]/I\right]^2}{J}, \quad (19)$$

where $\hat{\rho}_{i(S_j)}$ represents the EB-based reliability estimates for the *j*th state (J = 20) and the *i*th replication (I = 200) and $\rho_{i(S_j)}$ represent the state population reliability for the *j*th state which is regarded as a criterion reliability. Accordingly, a total of 200 resamplings were replicated for the states for the analysis. We regard the population reliability as truth for the evaluation.

Average variance is defined as the sum of the variances of the 200 replications for the states, divided by 20 (the number of states):

$$Average\,Variance = \frac{\sum_{j=1}^{20}\left[\sigma(\hat{\rho}_{i(S_j)} - \rho_{i(S_j)})\right]^2}{J}. \quad (20)$$

RMSE is defined as follows:

$$RMSE = \sqrt{Average\ Squared\ Bias + Average\ Variance}. \quad (21)$$

Average squared bias for SEM is equivalent to the sum of the average squared mean differences between the EB-based SEMs (or UC SEMs) and the population SEM for the states, divided by the number of states (i.e., 20). The formulas for the average squared bias, average variance, and RMSE for SEM are the same as (19), (20), and (21), replacing the reliability estimators with the SEM estimators.

Procedure

The current study was conducted using the following steps:

1. Selected 20 states out of 50 states by varying population size and reliability from small to large.
2. Calculated population reliability and SEM for the 20 states to use as a criterion.
3. Performed random sampling of four sample sizes of 25, 50, 125, and 250 from the population for each of the states.

4. Conducted resamplings 200 times for each sample size of 25, 50, 125, and 250 for each of the states.
5. Calculated UC reliability and UC SEM for each sample.
6. Calculated four EB reliabilities as in (9) to (13) and EB SEMs as in (15) to (18) for each sample.
7. Compared the UC and EB reliabilities and SEMs to the population reliability and SEM across replications for each sample size and state using the evaluation indexes (i.e., average squared bias, average variance, and RMSE).

Results

Empirical Bayes Reliability Estimators

Table 1 displays population reliability and SEM for the three measures (critical reading, math, and writing) for all 20 states. The mean of the 20 states' reliabilities for critical reading is 0.88 and ranges from 0.85 to 0.91; for math, the mean is 0.89 and ranges from 0.86 to 0.92; and for writing, the mean is 0.86 and ranges from 0.82 to 0.90. The mean SEMs for all 20 states are 3.49, 2.87, and 3.27 for critical reading, math, and writing, respectively. The SEM for reading ranges from 3.44 to 3.53; for math, from 2.80 to 2.93; and for writing, from 3.22 to 3.32. The reliability coefficient for each state for each measure is moderately high; the SEMs also look reasonably small. Overall, math reliabilities are slightly higher and SEMs are slightly lower than those of critical reading and writing. Statistics in bold font in the Table 1 are for the three states with large, medium, and small population size (i.e., Connecticut, Louisiana, and Wyoming) to illustrate how population size affects reliability and SEM. As shown, the reliability of the state with large population size tends to be higher than that of the state with small population size. Table 2 displays a layout of the empirical Bayes approaches for 200 replications across 20 states.

Table 3 presents a comparison of average squared bias, average variance, and RMSE for uncorrected (UC) and empirical Bayes (EB) reliability estimators.[4] The shaded areas in the table indicate the smallest indexes within each test and each sample size. As expected, as sample sizes are increased from 25 to 250, the extent of average squared bias, average variance, and RMSE are reduced.

For the smallest sample size of 25, EB Reliability Approach 2 (using *z*-transformation)

Table 3

Comparison of Statistics for Uncorrected and Empirical Bayes Reliability Estimators

		(Avg. squared bias) ×100			(Avg. variance) ×100			RMSE		
	Reliability coefficient	Reading	Math	Writing	Reading	Math	Writing	Reading	Math	Writing
N = 25	Uncorrected	0.0106	0.0071	0.0101	0.1664	0.1251	0.1824	0.0421	0.0363	0.0439
	Empirical Bayes	0.0134	0.0090	0.0154	0.0596	0.0454	0.0691	0.0270	0.0233	0.0291
	EB with *z*-transform	0.0089	0.0155	0.0096	0.0529	0.0386	0.0532	0.0249	0.0233	0.0251
	EB with *z*-transform (sqrt)	0.0089	0.0155	0.0097	0.0531	0.0387	0.0533	0.0249	0.0233	0.0251
	Based on EB SEM	0.0103	0.0072	0.0096	0.1569	0.1184	0.1698	0.0409	0.0354	0.0424
N = 50	Uncorrected	0.0021	0.0015	0.0022	0.0635	0.0473	0.0751	0.0256	0.0221	0.0278
	Empirical Bayes	0.0046	0.0037	0.0066	0.0263	0.0192	0.0335	0.0176	0.0151	0.0200
	EB with *z*-transform	0.0041	0.0077	0.0085	0.0257	0.0193	0.0271	0.0172	0.0165	0.0189
	EB with *z*-transform (sqrt)	0.0041	0.0077	0.0086	0.0257	0.0193	0.0271	0.0172	0.0165	0.0189
	Based on EB SEM	0.0021	0.0018	0.0021	0.0596	0.0448	0.0701	0.0249	0.0216	0.0269
N = 125	Uncorrected	0.0005	0.0003	0.0007	0.0229	0.0174	0.0284	0.0153	0.0133	0.0171
	Empirical Bayes	0.0023	0.0022	0.0034	0.0118	0.0088	0.0161	0.0119	0.0105	0.0140
	EB with *z*-transform	0.0023	0.0042	0.0063	0.0119	0.0094	0.0136	0.0119	0.0117	0.0141
	EB with *z*-transform (sqrt)	0.0023	0.0042	0.0063	0.0119	0.0094	0.0136	0.0119	0.0117	0.0141
	Based on EB SEM	0.0005	0.0004	0.0007	0.0216	0.0165	0.0265	0.0149	0.0130	0.0165
N = 250	Uncorrected	0.0001	0.0001	0.0001	0.0109	0.0082	0.0131	0.0105	0.0091	0.0115
	Empirical Bayes	0.0012	0.0010	0.0015	0.0069	0.0052	0.0091	0.0090	0.0078	0.0103
	EB with *z*-transform	0.0012	0.0017	0.0036	0.0070	0.0055	0.0080	0.0091	0.0084	0.0108
	EB with *z*-transform (sqrt)	0.0012	0.0017	0.0036	0.0070	0.0055	0.0080	0.0091	0.0084	0.0108
	Based on EB SEM	0.0002	0.0001	0.0002	0.0104	0.0079	0.0124	0.0103	0.0090	0.0112

Note. Shaded areas in the table indicate the smallest indexes within each test and each sample size. EB SEM = empirical Bayes standard of error measurement, RMSE = root mean squared error, sqrt = square root.

produced the smallest bias for reading and writing, and the UC reliability estimate yielded the smallest bias for math. For the sample size 50, the UC reliability estimates produced the smallest average squared bias for critical reading and math, and EB Reliability Approach 4 (using EB SEM) yielded the smallest bias for writing. For the sample sizes of 125 and 250, the UC reliability estimates produced the smallest average squared bias for all three measures. The results showed that when sample size is very small (i.e., $N = 25$), EB-based reliability estimates produced smaller bias than the bias of the UC reliability estimates for critical reading and writing. In terms of average variance and RMSE, EB reliability estimates (either EB Reliability Approach 1 or Approach 2) produced relatively small average variance and RMSE for all measures. It suggests that EB reliability estimates performed better than the traditional UC reliability estimates in terms of stability. As shown in Table 3, EB estimates with *z*-transformation (EB Reliability Approach 2) and EB estimates with *z*-transformation sqrt (EB Reliability Approach 3) are identical.

Figures 1 through 6 depict distributions of the UC and EB reliability estimates of 200 replications for critical reading, math, and writing, respectively. The first top panel of Figure 1 includes the reliability estimates distributions of two population sizes (large and small) among the states for sample size 25. Connecticut ($N = 36{,}180$) and Wyoming ($N = 2{,}226$) were selected as large and small population sizes of states, respectively, for illustration purposes. The next panel in Figure 1 is for sample size 50, and in Figure 2, it is for sample sizes 125 and 250. In each plot, four distributions are depicted: UC reliability (solid line), EB reliability (dashed line), EB reliability with *z*-transformation (dashed-dotted line), and EB reliability using EB SEM (dotted line). Since EB estimates in Approach 2 and Approach 3 are nearly identical, Approach 3 was dropped from the figures. The criterion line (i.e., population reliability estimates) is depicted with a vertical solid line. The horizontal axis of the plot shows the reliability and the vertical axis shows relative frequency in percent of the 200 replications whose estimates fell at a given level. As shown in Figures 1 and 2, the sample size used to estimate the reliability has an impact on each distribution of reliability estimates. As sample sizes are increased from 25 to 250, shapes of distributions of the UC and EB reliability estimates become closer to each other; and their means approach the criterion (dotted vertical line), which is the population reliability value. Distribution shapes of UC reliability (solid line) and EB reliability estimates using the EB SEM approach (dotted line) are very similar compared to other EB reliability estimates. Their means are also

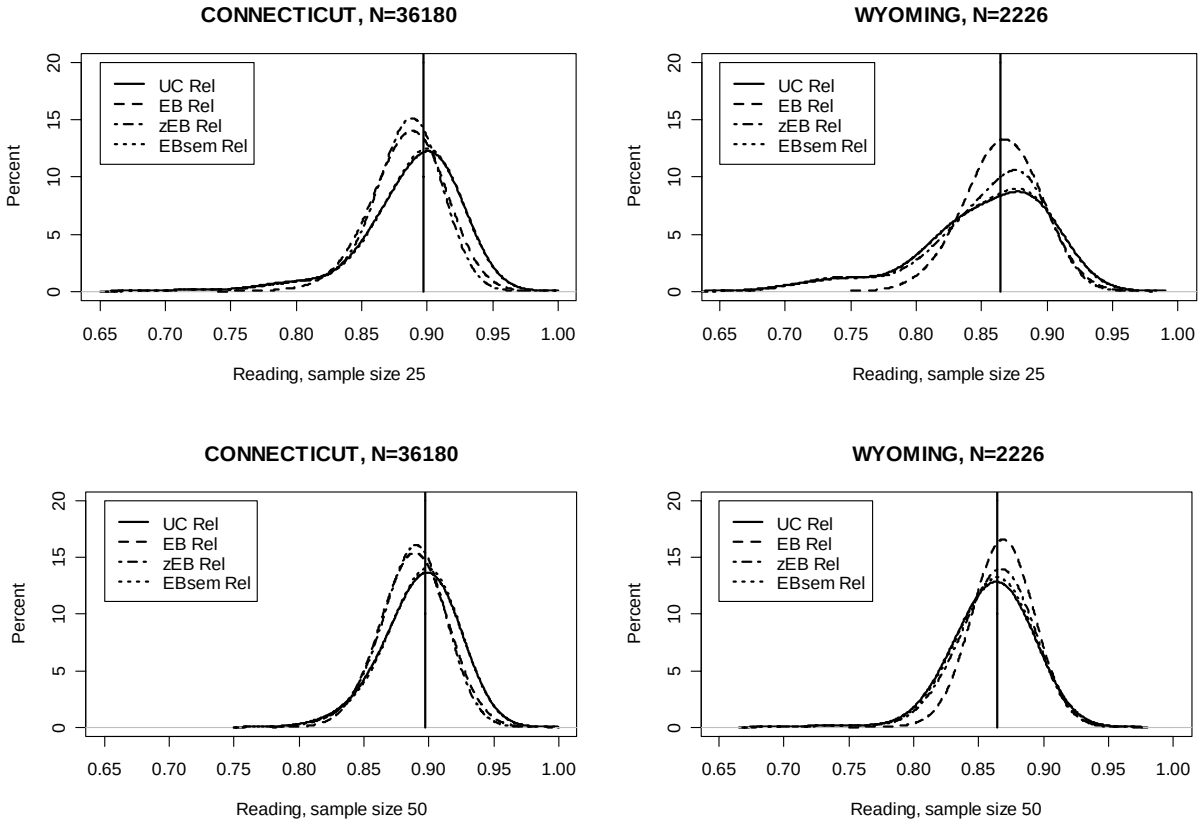

***Figure 1*. Frequency distribution of uncorrected and empirical Bayes reliability estimators from the resampling (*N* = 200) for reading, sample sizes 25 and 50.**

Note. UC Rel = uncorrected reliability, EB Rel = empirical Bayes reliability, *z*EB Rel = empirical Bayes reliability with *z*-transformation, EBsem Rel = empirical Bayes reliability using standard error of measurement.

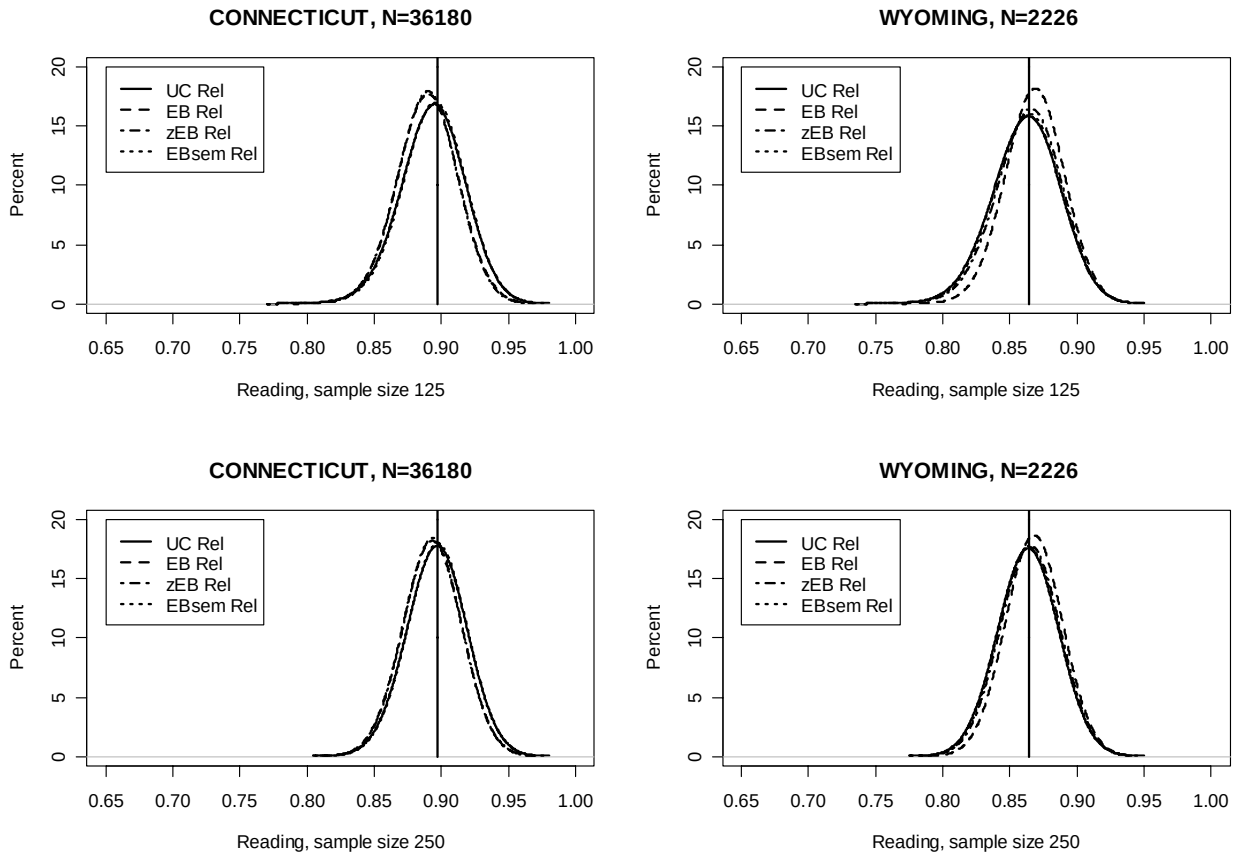

***Figure 2*. Frequency distribution of uncorrected and empirical Bayes reliability estimators from the resampling (*N* = 200) for reading, sample sizes 125 and 250.**

Note. UC Rel = uncorrected reliability, EB Rel = empirical Bayes reliability, *z*EB Rel = empirical Bayes reliability with *z*-transformation, EBsem Rel = empirical Bayes reliability using standard error of measurement.

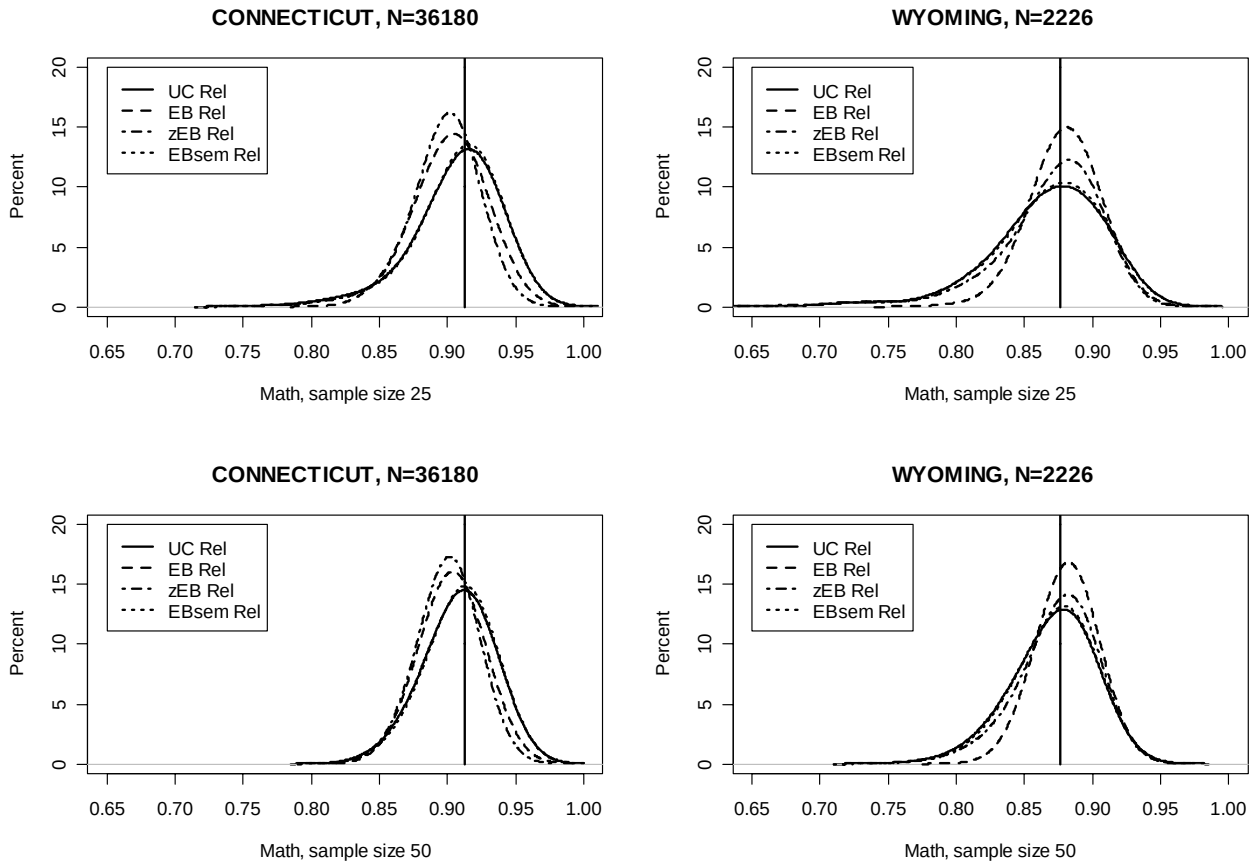

***Figure 3*. Frequency distribution of uncorrected and empirical Bayes reliability estimators from the resampling (*N* = 200) for mathematics, sample sizes 25 and 50.**

Note. UC Rel = uncorrected reliability, EB Rel = empirical Bayes reliability, *z*EB Rel = empirical Bayes reliability with *z*-transformation, EBsem Rel = empirical Bayes reliability using standard error of measurement.

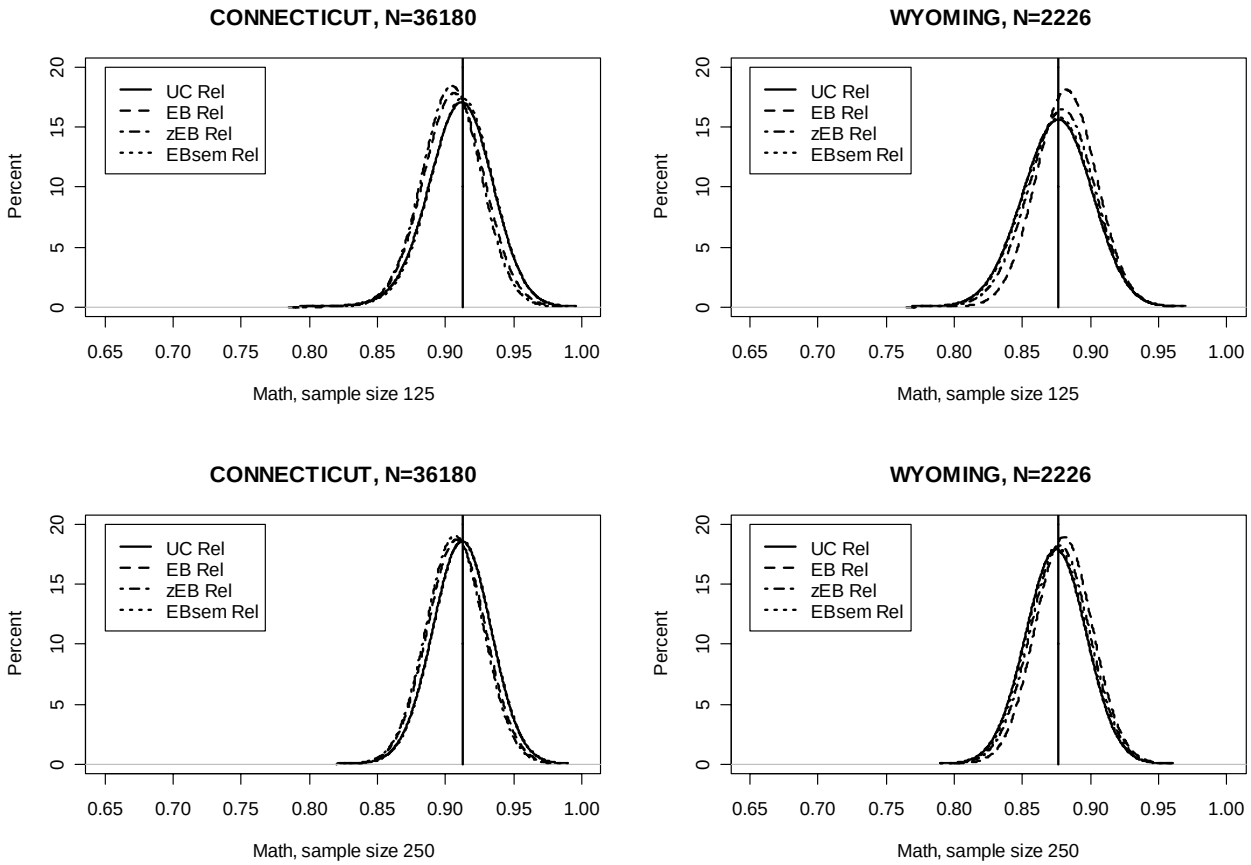

Figure 4. **Frequency distribution of uncorrected and empirical Bayes reliability estimators from the resampling (*N* = 200) for mathematics, sample sizes 125 and 250.**

Note. UC Rel = uncorrected reliability, EB Rel = empirical Bayes reliability, *z*EB Rel = empirical Bayes reliability with *z*-transformation, EBsem Rel = empirical Bayes reliability using standard error of measurement.

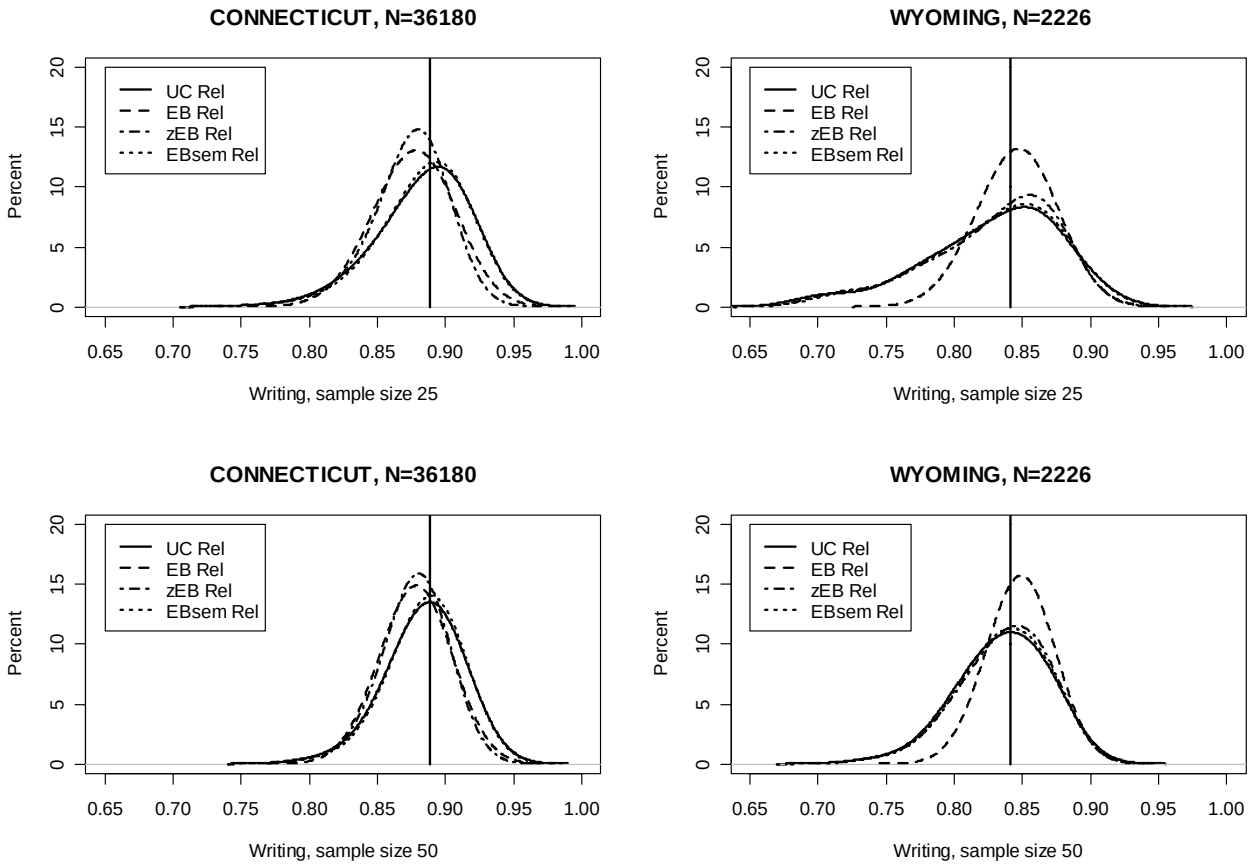

***Figure 5*. Frequency distribution of uncorrected and empirical Bayes reliability estimators from the resampling (*N* = 200) for writing, sample sizes 25 and 50.**

Note. UC Rel = uncorrected reliability, EB Rel = empirical Bayes reliability, *z*EB Rel = empirical Bayes reliability with *z*-transformation, EBsem Rel = empirical Bayes reliability using standard error of measurement.

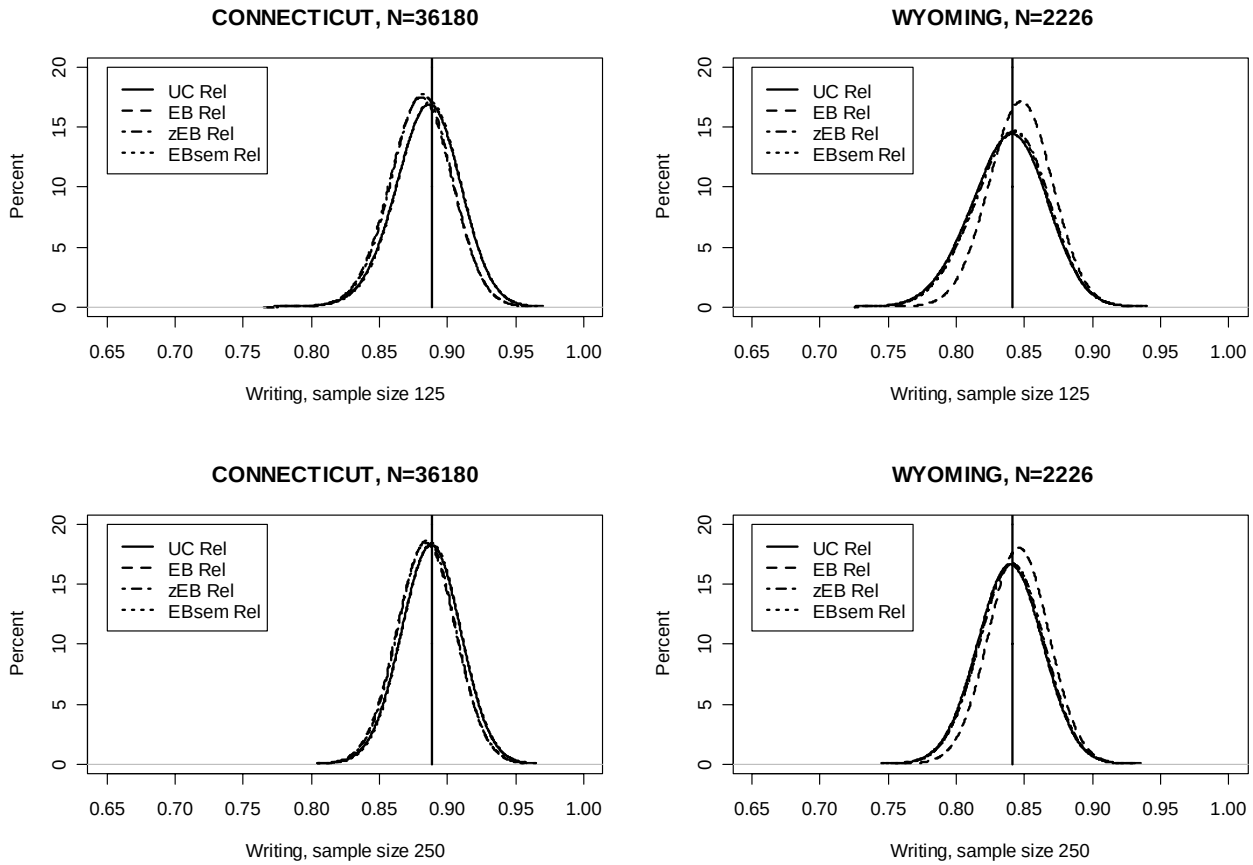

***Figure 6*. Frequency distribution of uncorrected and empirical Bayes reliability estimators from the resampling (*N* = 200) for writing, sample sizes 125 and 250.**

Note. UC Rel = uncorrected reliability, EB Rel = empirical Bayes reliability, *z*EB Rel = empirical Bayes reliability with *z*-transformation, EBsem Rel = empirical Bayes reliability using standard error of measurement.

close to the population reliability for all sample sizes with a few exceptions (e.g., sample size 25 for Wyoming), indicating low bias. It appears that the population size of each state affects the distribution shape of the UC and EB reliability estimates using the SEM approach appreciably only when the sample sizes are very small. The distribution shapes of UC reliability (solid line) and EB reliability estimate using the EB SEM approach (dotted line) are more skewed than other EB estimates for the states with small population size (Wyoming), particularly for a sample size of 25. Overall, the EB Reliability Approach 1 (dashed line) performs better than the UC reliability and other EB reliability estimates in terms of accuracy and stability, particularly when both the sample and population sizes are very small (i.e., sample size = 25 for Wyoming).

Figures 3 and 4 show the comparison of UC and EB reliability distributions for math. The distribution shapes and patterns are similar to the findings from the reading plots in Figures 1 and 2. As sample sizes increase, the shapes of distributions of the UC and EB reliability estimates converge and their means approach the population reliability. Distribution shapes of UC reliability (solid line) and EB reliability estimates using the EB SEM approach (dotted line) are very similar compared to the other EB reliability estimates. Also, their means are close to the population reliability estimates for almost all sample sizes. When the sample and population sizes are small, the EB Reliability Approach 1 (dashed line) performs better than the UC reliability and other EB reliability estimates in terms of accuracy and stability (i.e., sample size = 25 for Wyoming).

Figures 5 and 6 display a comparison of UC and EB reliability distributions for writing. The distribution shapes and patterns are similar to the finding from critical reading and math. While shapes of distributions of the UC and EB reliability estimates become more similar to each other and their means are closer to the population reliability as sample sizes increase, when both sample and population sizes are small, EB Reliability Estimate Approach 1 (dashed line) performs better than UC and other EB reliability estimates.

In summary, EB-based approaches (either EB Reliability Approach 1 or EB Reliability Approach 2) produced relatively small average variances and RMSEs compared to the UC reliability estimates for all three measures. In terms of average bias, EB reliability estimates produce relatively smaller average squared bias than UC reliability estimates for reading and writing, particularly when sample size is very small, while UC reliability estimates produce

relatively smaller average squared bias than EB reliability estimates for all three measures, particularly when sample size is greater than 25, except for a few exceptions in writing.

Among the four EB reliability approaches, Reliability Approach 2 and Reliability Approach 3 are nearly identical for average squared bias, average variance, and RMSE for all three measures across different sample sizes (see Table 3). Interestingly, EB Reliability Approach 4 seems to be more similar to UC approach than other EB reliability approaches in terms of average squared bias, average variance, and RMSE. As shown in the Figures 1 to 6, when both the sample and population sizes are very small (N = 25 for Wyoming), EB Reliability Approach 1 (dashed line) performs better than UC and other EB estimates.

Empirical Bayes Standard Error of Measurement Estimators

Table 4 displays a comparison of average squared bias, average variance, and RMSE for UC versus EB SEM estimators.[5] As observed in the reliability results, as sample sizes increased, average squared bias, average variance, and RMSE of the SEM indexes decreased. Analysis results for SEM are very consistent across different sample sizes and measures. Uncorrected SEM estimates produce the smallest average squared bias for all three measures across different sample sizes. As observed in the reliability results, a sample size of 250 yields the smallest bias. For average variance and RMSE, EB SEM Approach 1 (EB methodology applied to the SEM computed from the uncorrected reliability index) produces the smallest average variance and RMSE for all three measures across different sample sizes.

Table 4 shows that the estimates produced by EB SEM Approach 3 (EB reliability with *z*-transformation) and EB SEM Approach 4 (EB reliability with *z*-transformation sqrt) are virtually identical for average squared bias, average variance, and RMSE for all three measures across different sample sizes. The difference between EB SEM Approach 3 and EB SEM Approach 4 is that Approach 3 computed SEM using the EB reliability with Fisher's *z*-transformation while Approach 4 employed square roots of the EB reliabilities and then used the *z*-transformation.

Figures 7 through 12 illustrate distributions of the UC and EB SEM estimates across 200 replications for critical reading, math, and writing, respectively. The top panel of Figure 7 includes the SEM distributions for states with large and small population sizes for sample size 25. The next panel is for sample size 50, and so on. As shown in the figures, sample sizes used

Table 4

Comparison of Statistics for Uncorrected and Empirical Bayes Standard Error of Measurement Estimators

		(Avg. squared bias) ×100			(Avg. variance) ×100			RMSE		
	SEM coefficient	Reading	Math	Writing	Reading	Math	Writing	Reading	Math	Writing
N = 25	Uncorrected	0.0032	0.0019	0.0022	0.6226	0.5434	0.6223	0.0791	0.0738	0.0790
	Empirical Bayes SEM	0.0125	0.0247	0.0164	0.2006	0.1845	0.2065	0.0462	0.0457	0.0472
	Based on EB reliability	1.4567	0.8733	1.4441	4.5296	3.0007	3.0202	0.2446	0.1968	0.2113
	Based on EB reliability with *z*-transform	1.1286	1.6020	1.6965	3.9668	3.1806	2.8176	0.2257	0.2187	0.2125
	Based on EB reliability with *z*-transform (sqrt)	1.1289	1.6026	1.6971	3.9645	3.1766	2.8186	0.2257	0.2186	0.2125
N = 50	Uncorrected	0.0019	0.0014	0.0010	0.3071	0.2564	0.2912	0.0556	0.0508	0.0541
	Empirical Bayes SEM	0.0113	0.0201	0.0132	0.1037	0.1000	0.1086	0.0339	0.0347	0.0349
	Based on EB reliability	1.0821	0.6296	0.9791	1.4869	0.9860	0.9630	0.1603	0.1271	0.1394
	Based on EB reliability with *z*-transform	1.0753	1.1816	1.4711	1.3699	1.0790	1.0823	0.1564	0.1504	0.1598
	Based on EB reliability with *z*-transform (sqrt)	1.0751	1.1809	1.4744	1.3696	1.0781	1.0830	0.1564	0.1503	0.1599
N = 125	Uncorrected	0.0009	0.0005	0.0006	0.1142	0.1081	0.1209	0.0339	0.0330	0.0349
	Empirical Bayes SEM	0.0086	0.0124	0.0115	0.0453	0.0535	0.0522	0.0232	0.0257	0.0252
	Based on EB reliability	0.5704	0.3865	0.5006	0.3245	0.2358	0.2035	0.0946	0.0789	0.0839
	Based on EB reliability with *z*-transform	0.6104	0.7562	0.9399	0.3060	0.2534	0.2537	0.0957	0.1005	0.1093
	Based on EB reliability with *z*-transform (sqrt)	0.6102	0.7554	0.9424	0.3060	0.2532	0.2540	0.0957	0.1004	0.1094
N = 250	Uncorrected	0.0003	0.0001	0.0002	0.0582	0.0512	0.0578	0.0242	0.0227	0.0241
	Empirical Bayes SEM	0.0071	0.0070	0.0070	0.0263	0.0313	0.0306	0.0183	0.0196	0.0194
	Based on EB reliability	0.2692	0.1708	0.2019	0.0926	0.0677	0.0569	0.0602	0.0488	0.0509
	Based on EB reliability with *z*-transform	0.2917	0.3183	0.4632	0.0882	0.0733	0.0731	0.0616	0.0626	0.0732
	Based on EB reliability with *z*-transform (sqrt)	0.2916	0.3178	0.4648	0.0882	0.0733	0.0731	0.0616	0.0625	0.0733

Note. Shaded areas in the table indicate the smallest indexes within each test and each sample size. EB SEM = empirical Bayes standard of error measurement, RMSE = root mean squared error, sqrt = square root.

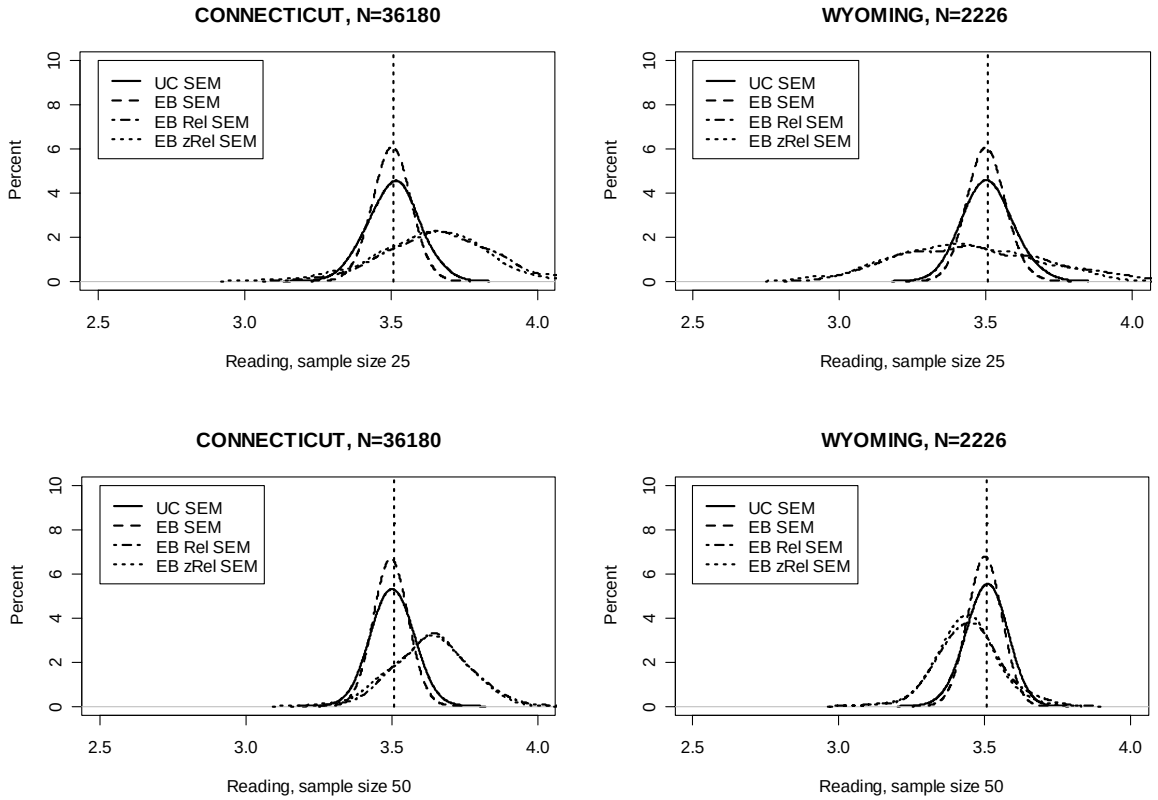

***Figure 7*. Frequency distribution of uncorrected and empirical Bayes standard error of measurement estimators from the resampling (*N* = 200) for reading, sample sizes 25 and 50.**

Note. UC SEM = uncorrected standard of error measurement, EB SEM = empirical Bayes standard of error measurement, EBsem Rel = empirical Bayes reliability standard error of measurement EB *z*Rel SEM = empirical Bayes reliability with *z*-transformation using standard of error measurement.

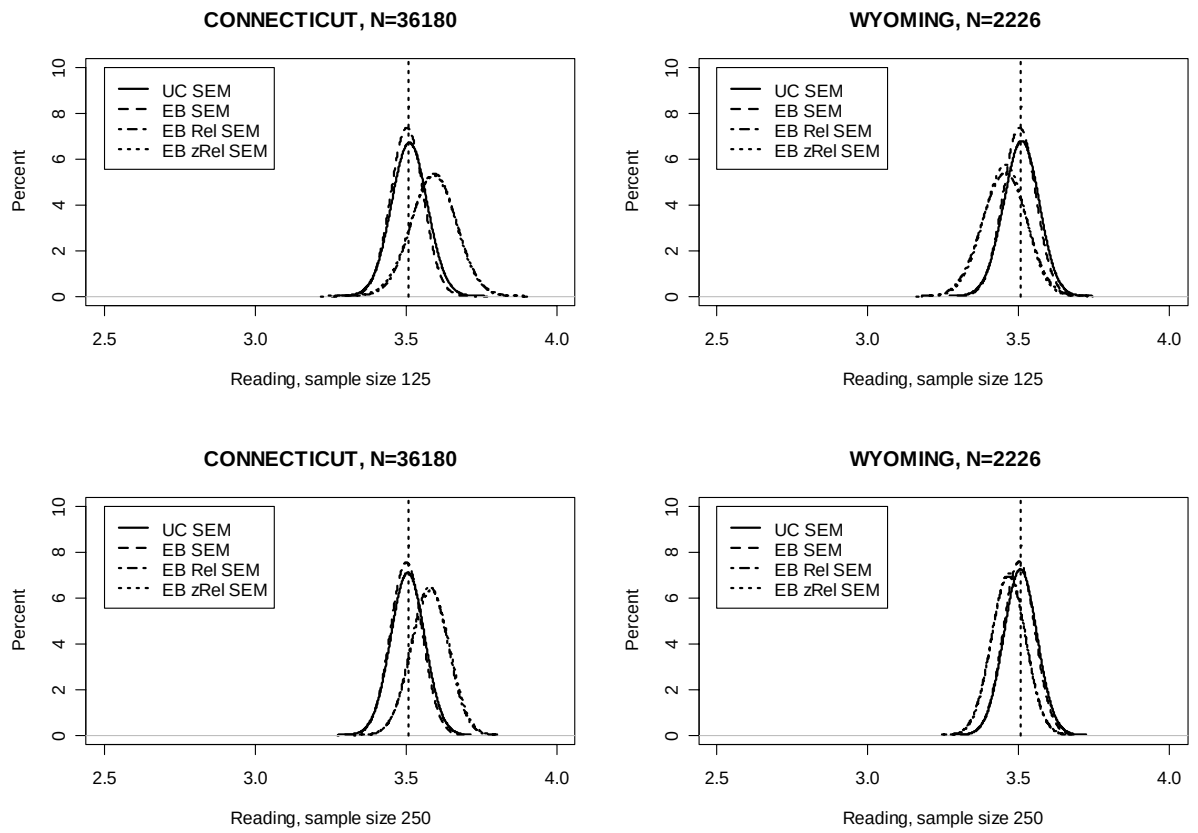

***Figure 8*. Frequency distribution of uncorrected and empirical Bayes standard error of measurement estimators from the resampling (*N* = 200) for reading, sample sizes 125 and 250.**

Note. UC SEM = uncorrected standard of error measurement, EB SEM = empirical Bayes ustandard of error measurement, EBsem Rel = empirical Bayes reliability using standard error of measurement EB *z*Rel SEM = empirical Bayes reliability with *z*-transformation using standard of error measurement.

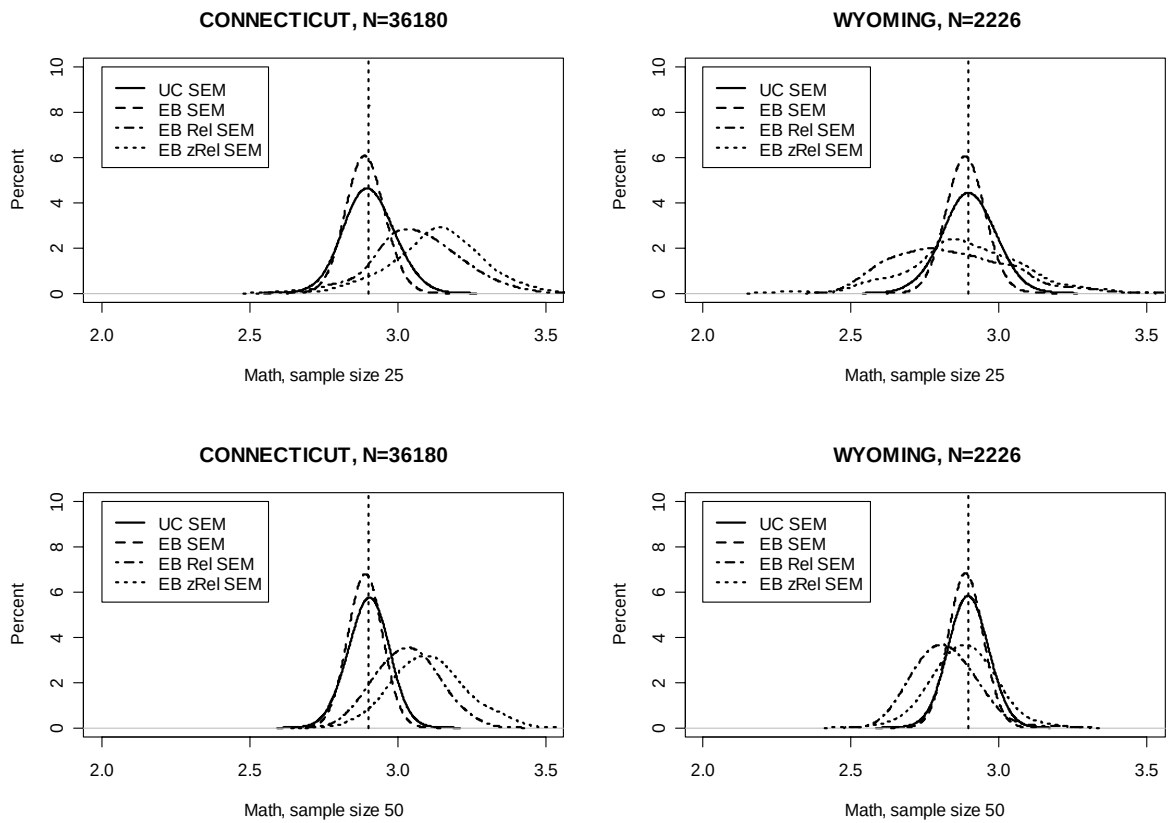

***Figure 9*. Frequency distribution of uncorrected and empirical Bayes standard error of measurement estimators from the resampling (*N* = 200) for mathematics, sample sizes 25 and 50.**

Note. UC SEM = uncorrected standard of error measurement, EB SEM = empirical Bayes standard of error measurement, EBsem Rel = empirical Bayes reliability standard error of measurement EB *z*Rel SEM = empirical Bayes reliability with *z*-transformation using standard of error measurement.

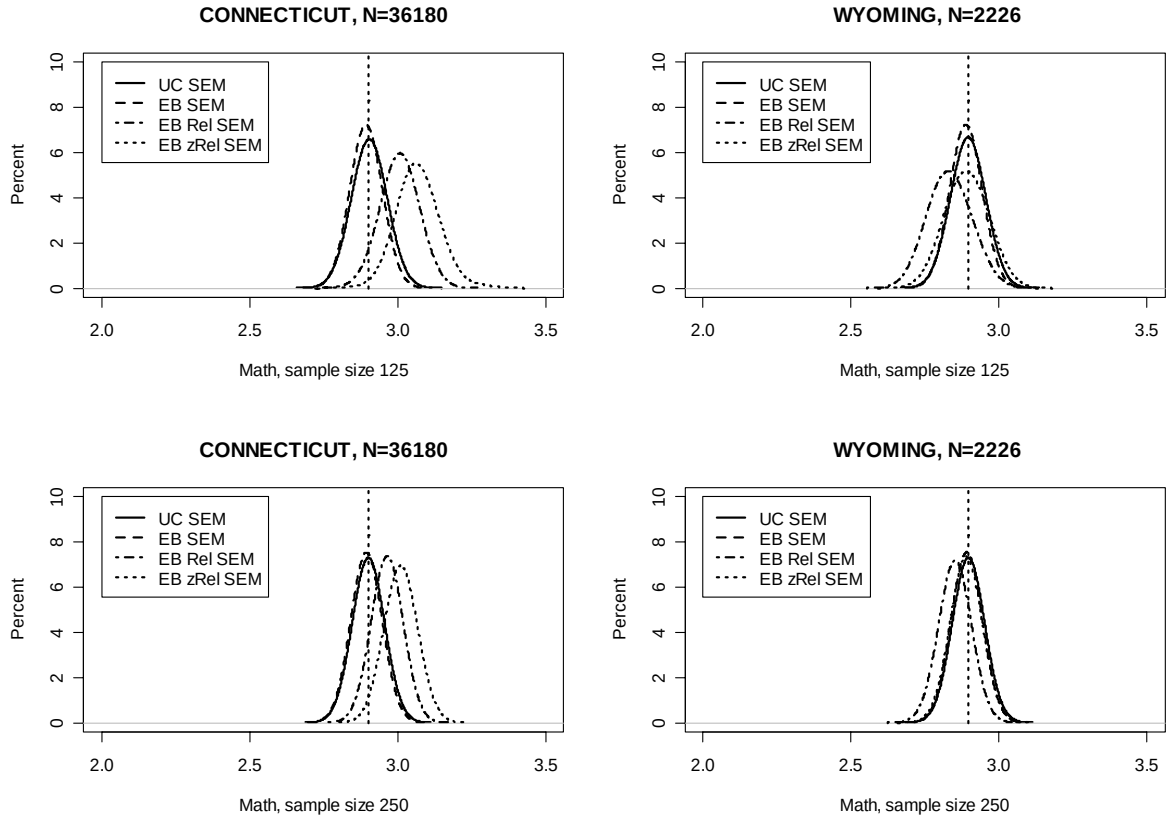

***Figure 10*. Frequency distribution of uncorrected and empirical Bayes standard error of measurement estimators from the resampling (*N* = 200) for mathematics, sample sizes 125 and 250.**

Note. UC SEM = uncorrected standard of error measurement, EB SEM = empirical Bayes standard of error measurement, EBsem Rel = empirical Bayes reliability standard error of measurement EB *z*Rel SEM = empirical Bayes reliability with *z*-transformation using standard of error measurement.

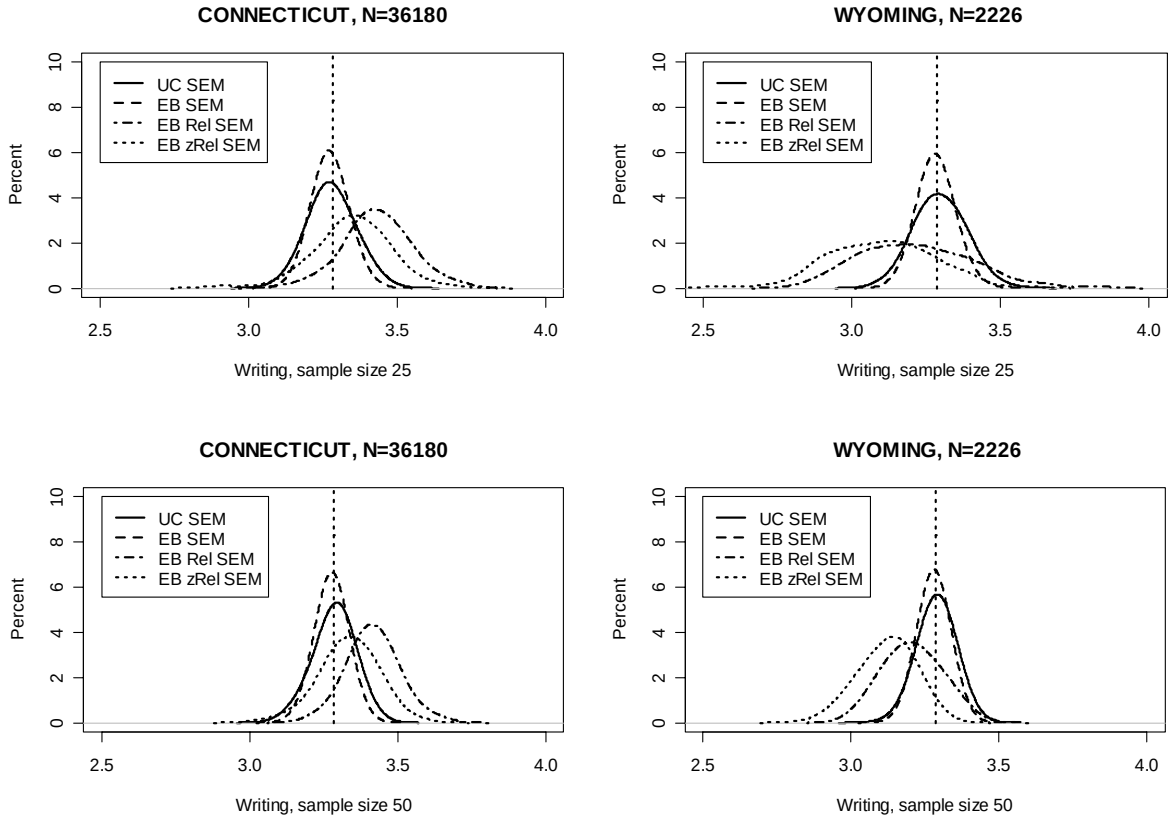

Figure 11. **Frequency distribution of uncorrected and empirical Bayes standard error of measurement estimators from the resampling (*N* = 200) for writing, sample sizes 25 and 50.**

Note. UC Rel = uncorrected reliability, EB Rel = empirical Bayes reliability, zEB Rel = empirical Bayes reliability with *z*-transformation, EBsem Rel = empirical Bayes reliability using standard error of measurement.

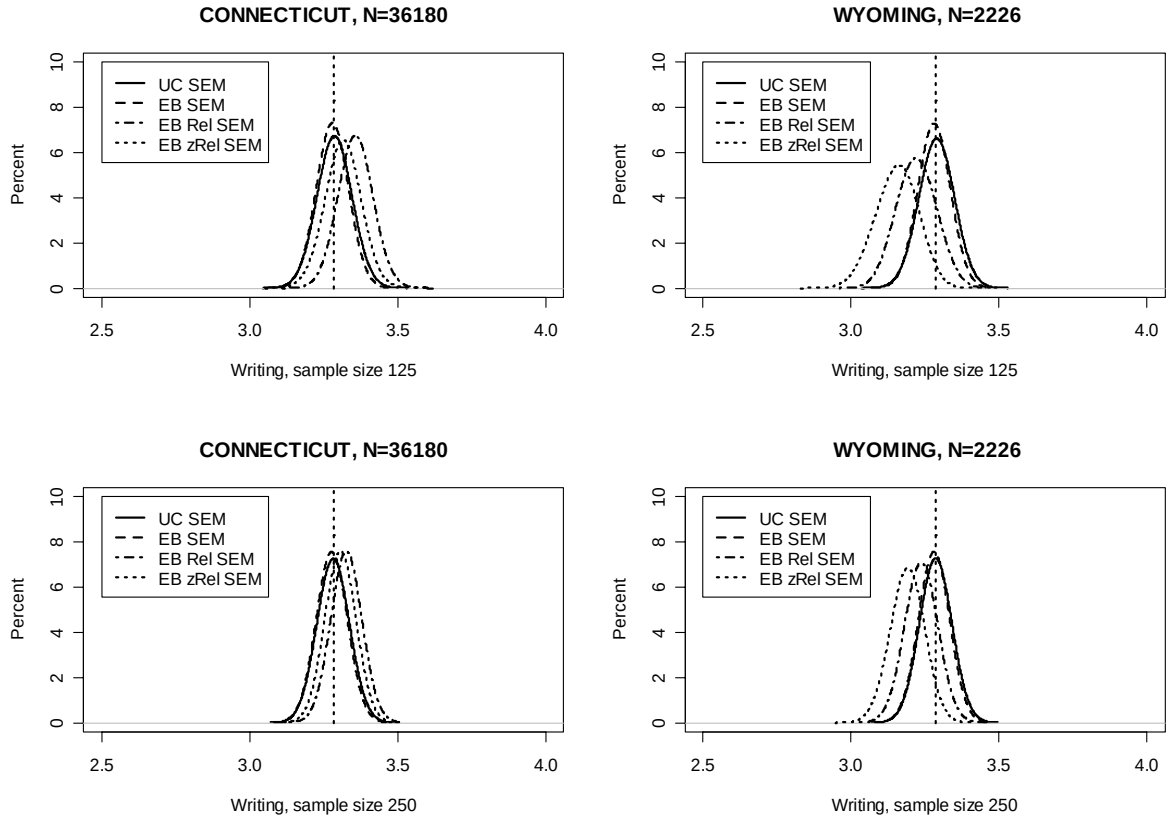

***Figure 12*. Frequency distribution of uncorrected and empirical Bayes standard error of measurement estimators from the resampling (*N* = 200) for writing, sample sizes 125 and 250.**

Note. UC SEM = uncorrected standard of error measurement, EB SEM = empirical Bayes standard of error measurement, EBsem Rel = empirical Bayes reliability standard error of measurement EB zRel SEM = empirical Bayes reliability with *z*-transformation using standard of error measurement

to estimate the SEM impact the SEM distributions. As sample sizes increase from 25 to 250, shapes of the distributions of the UC and EB SEM estimates become more similar to each other and their means approach the criterion (the population SEM, indicated by the dotted vertical line). The shapes of distributions of the UC SEM and EB SEM Approach 1 are most similar when compared to other EB SEM distributions. The means of the UC SEM and EB SEM Approach 1 are also close to the population SEM for all sample sizes. Population size does not appear to affect appreciably the distribution shape of the UC and EB SEM estimates for large sample sizes, but it seems to affect the distributions of the EB SEM approaches in the small sample size of 25, particularly for EB SEM Approach 2 (dotted-dashed line) and EB SEM Approach 3 (dotted line) for all three measures.

In summary, UC SEM estimates produce smaller average squared bias than EB SEM estimates, while EB-based approaches (either EB SEM Approach 1) produce relatively small average variances and RMSEs compared to the UC SEM estimates for all three measures. For all three measures across different sample sizes, EB SEM Approach 3 and EB SEM Approach 4 are nearly identical for bias, error, and RMSE (see Table 3). As displayed in Figures 7 through 12, when sample sizes are small (N = 25 or N = 50), EB SEM Approach 1 performs better than the UC and other EB SEM approaches.

Discussion

In the current study, an EB procedure was evaluated for estimating reliability of subgroups of a population, even for very small subgroups; this evaluation was to improve the precision and accuracy of reliability estimation by integrating collateral information from the reliability of other subgroups (i.e., states). The Bayesian estimates were compared to the traditionally and currently used Cronbach's alpha coefficient, in terms both of average squared bias, average variance, and RMSE.

The general findings for both reliability and SEM estimates from the current study are that the EB-based approach produced greater bias but less error, with a few exceptions. Sample size seems to have the a sizable impact on both EB and UC analyses results in terms of bias and error, and population size does have some impact on the distribution of EB and UC estimates only when the sample size is small.

More specifically, as sample sizes increase, sizes of average squared bias, average variance, and RMSE decrease, the shapes of distributions of the UC and EB estimates became more similar to each other and the UC and EB estimates' means are close to the criterion.

In regard to population size, it appears that population size of each state does not affect appreciably the distribution shape of reliability and SEM estimates, particularly in the large sample sizes, but it seems to affect the distributions of the reliability and SEM estimates only in small sample size of 25. Particularly, UC reliability and EB Reliability Approach 4, as well as EB SEM Approach 2 and EB SEM Approach 3, are distant from a normal distribution when both population size and sample size are small.

For comparison of UC and EB reliability estimates, the EB reliability estimates usually produce relatively greater average squared bias than UC reliability estimates except for a few exceptions in writing. However, absolute differences in the average squared bias (×100) between UC and EB reliability estimates are very small even in the small sample size of 25. Differences ranged from 0.0002 to 0.0028 for reading, from 0.0016 to 0.0049 for math, and from 0.0010 to 0.0048 for writing. Particularly, the absolute differences between UC reliability and EB Reliability Approach 4 (based on EB SEM) are almost negligible. In addition, EB Reliability Approach 1 and EB Reliability Approach 2 produce relatively small average variances and RMSEs compared to the UC reliability for all three measures across different sample sizes. These results indicate that EB-based reliability estimation seems promising with even small sample sizes, particularly in terms of average variance and RMSE.

Although estimating reliability based on the EB approaches is attractive for small sample groups, the EB estimates for the small groups will be regressed toward the estimate for the large group when one group is large and the others small. Another concern is that the reliability is sensitive to the degree of heterogeneity within a group. Therefore, we calculated the SEM estimate, which is a quantity less sensitive to heterogeneity/homogeneity.

For comparison of the UC and EB SEM estimates, the pattern of results is similar to the results that we found from the reliability analysis results, but the SEM results are more consistent across different sample sizes and measures than the results from the reliability estimates. The EB SEM estimates produce relatively larger average squared bias than UC reliability estimates for all three measures across four sample sizes. The EB SEM Approach 1 produced small average variances and RMSEs for all three measures across sample sizes. The results of EB SEM

Approach 3 and EB SEM Approach 4 were nearly identical for average squared bias, average variance, and RMSE for all three tests across different sample sizes.

As shown in the Table 1, the data used for the current study were reasonably stable and did not include any extreme small or large values in population size, reliability coefficients, and SEMs. From a practical point of view, the results of the current study provide some support for EB-based reliability estimation using collateral information with even very small sample sizes of 25 in terms of average squared bias, average variance, and RMSE.

The current study leaves a number of issues to be considered even though the EB-based approaches would seem to have some benefits that traditional reliability estimation methods do not have. As commonly recognized, EB estimation works better when more groups are used, because the between-groups variability must be estimated from the empirical group information, so the current study used 20 states. In practice, however, some situations may not have enough subgroups (e.g., gender group) and, consequently, not enough collateral information. Or in some situations, one group is large (e.g., the White group in the ethnic subgroup) but the other group is small (e.g., the American Indian group). In such cases, EB-based reliability estimation could not be expected to function very well because the estimates for small groups will be regressed toward the estimates for the large group. That is, the effectiveness of the EB-based approaches greatly depends on the characteristics of the collateral information. Therefore, it requires prudent judgment to select adequate collateral information to obtain precise and accurate reliability estimates.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

Botje, M. (2006). *Bayesian inference.* Retrieved April 3, 2009, from http://www.nikhef.nl/~h24/bayes/ Bayestopical3.ppt

Brandel, J. (2004). *Empirical Bayes methods for missing data analysis* (Project Rep. No. 2004:11). Uppsala, Sweden: Uppsala University, Department of Mathematics.

Braun, H. I., & Jones, D. H. (1984). *Use of empirical Bayes methods in the study of the validity of academic predictors of graduate school performance* (ETS Research Rep. No. RR-84-34). Princeton, NJ: ETS.

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*(3), 241–259.

Livingston, S. A., & Lewis, C. (2009). *Small-sample equating with prior information* (ETS Research Rep. No. RR-09-25). Princeton, NJ: ETS.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Sinharay, S., Dorans, N. J., Grant, M. C., Blew, E. O., & Knorr, C. M. (2006). *Using past data to enhance small-sample DIF estimation: A Bayesian approach* (ETS Research Rep. No. RR-06-09). Princeton, NJ: ETS.

Walker, M. E., & Zhang, L. Y. (2004, April). *Estimating internal consistency reliability of tests for ethnic and gender subgroups within a population.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel DIF analysis to a computerized adaptive test. *Applied Psychological Measurement, 26*(1), 57–76.

Notes

[1] The continuous probability density function of the normal distribution is defined as

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

[2] In the interests of conciseness, this paper will refer to both the 19 states in the study and Washington, DC, generally as states.

[3] Nunnally and Bernstein (1994) made the argument by referring to the formula from classical test theory for the reliability coefficient: $\rho_{xx'} = 1 - \sigma_e^2 \Big/ \sigma_x^2$. If the error variance remains fairly constant across groups (an assumption of many derivations from classical test theory), then the size of the reliability coefficient is completely determined by the variance of *X*.

[4] Because average squared bias and average variance of UC and EB estimates were very small in numerical value (e.g., 0.0001065 for reading average squared bias in sample size 25), those indexes were multiplied by 100.

[5] Because average squared bias and average variance of UC and EB estimates were very small in numerical value (e.g., 0 000032 for reading average squared bias in sample size 25), multiplied by 100 to those indexes.