A Simulation Study for Comparing Three Lower Bounds to Reliability

Wei Tang, CRAME, Department of Educational Psychology
University of Alberta, CANADA
Ying Cui, CRAME, Department of Educational Psychology
University of Alberta, CANADA

Paper presented on April 17, 2012 at the AERA Division D: Measurement and Research Methodology, Section 1: Educational Measurement, Psychometrics, and Assessment. (Paper titled in AERA program: *Comparing Four Lower Bound Coefficients of Reliability Under Violation of the Assumption of Unidimensionality*).

Abstract

Reliability is the quantification of the consistency of results from a measurement procedure across replications. By far, Guttman’s Lambda 3 or coefficient alpha, as a lower bound to the reliability, is the most frequently used and reported index of reliability because of its simplicity in computation and accessibility in statistical programs. However, coefficient alpha tends to be negatively biased when the internal structure of the measurement is heterogeneous. This study is designed to use simulated data to evaluate and compare two alternatives for coefficient alpha: the greatest lower bound and Lambda 2 with coefficient alpha. In light of our simulation results, lambda 2 is recommended because it shows the least amount of bias under most of the simulation conditions.

Keywords: Reliability; Lower bounds to reliability; Coefficient alpha; The greatest lower bound; Lambda 2; Sample bias.

Introduction

Reliability is the quantification of the consistency of results from a measurement procedure across replications. The reliability of measurement results is initially defined by Spearman (1904) as the correlation between the total scores on two independent administrations or two repeated or parallel forms of the test. However, data from repeated testing or parallel forms are rarely available in practice. Thus, reliability is usually estimated based on scores from a single test administration, and defined by the ratio of the true score variance to the observed score variance. The reliability coefficients based on a single administration of one test form is traditionally called lower bounds to reliability, named after Guttman (1945)'s influential six lower bounds to reliability (i.e., Lambda 1 to 6). Guttman (1945) proposed the idea of bounding the estimation of the true test score variance or error test score variance to derive the estimates of reliability. That is, specific constraints/inequalities are used to minimize the true score variance or to maximize error score variance so that the estimate of reliability can be always lower than or at most equal to the true reliability, thereby producing a lower bound to reliability.

By far, Guttman's Lambda 3 or coefficient alpha (Cronbach, 1951; Guttman, 1945; Kuder & Richardson, 1937), as a lower bound to the reliability, is the most frequently used and reported index of reliability because of its simplicity in computation and accessibility in statistical programs. However, many studies (e.g., Cortina, 1993; Green & Hershberg, 2000; Green & Yang, 2009; Miller, 1995; Raykov, 1998; Sijtsma, 2009) have suggested that coefficient alpha may not be the best estimate of reliability. These studies revealed that coefficient alpha tends to be negatively biased when the internal structure of the measurement is heterogeneous in the sense that items measure one or more constructs/factors and have different item weights/loadings on the factor(s). For example, Green and Yang (2009) demonstrated how the bias in the population coefficient alpha varies as the internal structure systematically changes. They showed that the bias in population coefficient alpha even reached as high as -.114, 18.9% lower than the true reliability.

To promote better reliability estimation practice, alternative reliability coefficients have been studied and compared with coefficient alpha in order to find the most appropriate replacement for researchers and practitioners. A review of the past comparison studies on reliability coefficients shows that these studies have examined the bias of different coefficients at either the *population* level or the *sample* level (e.g., Callender & Osburn, 1979; Green & Hershberg, 2000; Osburn, 2000; Shapiro & Ten Berge, 2000; Ten Berge & Sočan, 2004; Zinbarg, Revelle, Yovel, & Li, 2005).

Studies at the population level examined the bias of the population values of different coefficients from the true reliability of the test under various conditions. For example, Osburn's study (2000) examined up to ten coefficients by varying the degree of heterogeneity of test items. All the coefficients were calculated based on the population correlation matrices and then compared to the true reliability of the test. On the other hand, studies at the sample level examined the bias of sample values of different coefficients from their corresponding population values. For example, Ten Berge and Sočan (2004) simulated 500 samples of size 100, 250, 500, and 1000 based on the correlation matrix of a real data with 119 subjects and 6 items. The authors compared coefficient alpha, the maximized Lambda 4 (Guttman, 1945) and the greatest lower bound (the glb; Woodhouse & Jackson, 1977; Bentler & Woodward, 1980; Ten Berge, Snijders & Zegers 1981) in terms of their sample bias from the corresponding population coefficients.

Studies at the sample level are useful in illustrating the degree to which the sample estimates of different reliability coefficients is consistent with their corresponding population values. More importantly, however, one needs to know the degree of the bias of a sample estimate relative to the true reliability of the test. That is, the population level and sample level bias must be considered together so as to provide a complete picture in helping researchers and practitioners identify the best sample estimate(s) of reliability in real settings. However, no studies have been conducted to consider the reliability coefficient bias at the population and sample levels simultaneously so as to compare different sample reliability estimates relative to the true reliability of the test. In addition, we found that the effect of dimensionality of the test on the sample bias of reliability coefficients had not been studied in the literature.

To fill this gap in the literature, the present study is designed to use simulated data to compare two selected candidate alternatives, Lambda 2 and the glb, with coefficient alpha at both the population level and sample level, and specifically analyze the bias of their sample estimates with respect to the true reliability. The reason for selecting Lambda 2 and the glb as alternatives for coefficient alpha is due to their theoretical appealingness, which will be discussed in the next section. This study is intended to compare the accuracy of the sample estimates of the two coefficients in reflecting the true reliability when coefficient alpha is expected to be severely biased. Therefore, this study specifically focuses on simulation conditions under which the absolute value of alpha's bias is greater than 10% of true reliability as reported by Green and Yang (2009).

In section 2, we review the two coefficients and present the rationale for choosing them as the possible candidates for replacing coefficient alpha. In section 3, we present the methods, results and discussion of our simulation study. In section 4, we summarize and discuss the findings from the study. Recommendations for practitioners are provided.

Review of Lambda 2 and the greatest lower bound

Lambda 2

Guttman (1945) developed six lower bounds measures to reliability, namely, Lambda 1 to 6. Although Guttman initially recommended the use of Lambda 3 (i.e., coefficient alpha) and Lambda 4 among the six proposed measures, this recommendation was mainly due to the consideration of the relative ease of computation of the two coefficients. As proved by Guttman (1945), Lambda 2 is always equal to or greater than coefficient alpha (which is always greater than Lambda 1), indicating that Lambda 2 is closer to the true reliability. Therefore, one may consider Lambda 2 as the better lower bound to reliability than coefficient alpha and Lambda 1. Furthermore, Lambda 5 and Lambda 6 are generally lower than Lambda 2, meaning that they tend to be further away from the true reliability. In addition, Ten Berge and Zegers (1978) demonstrated an infinite series of successive improvement to Guttman's Lambda 1 where coefficient alpha and Lambda 2 are the first two in the series. As concluded by the authors, the series does not improve much after Lambda2.

The value of Lambda 4 depends on how the test is split into halves, and thus is somewhat subjective. Although one can find the maximized Lambda 4 (i.e., the maximum of the values of all the possible splits of a test), this is not recommended for two reasons. First, it is not always feasible to compute Lambda 4 on all possible splits of the test. For example, a 50-item test can be divided into two 25-item half tests in over 63 trillion ways (Osburn, 2000). Thus, the practical value of the maximized Lambda 4 is diminished when the number of items becomes large. Second, Ten Berge and Sočan (2004) suggested that the maximized Lambda 4 may grossly overestimate the population values when computed in small samples. And Woodhouse and Jackson (1977) demonstrated the maximized Lambda 4 is always lower than or equal to the glb. Hence, Lambda 4 was not considered in this study.

A general equation for lower bounds of reliability can be expressed as

$$\rho = \frac{\sum_{j=1}^{n}\sum_{i=1\ (i\neq j)}^{n}\sigma_{ij}^{2}+\mathrm{MinTr}\sum_{\mathrm{T}}}{\mathbf{1}'\sum_{X}\mathbf{1}} = 1-\frac{\mathrm{MaxTr}\sum_{\mathrm{E}}}{\mathbf{1}'\sum_{X}\mathbf{1}} \tag{1}$$

where $\sum_{\mathrm{T}}$, $\sum_{\mathrm{E}}$ and $\sum_{\mathrm{X}}$ are the true score, error score and observed score variance covariance matrices, respectively, σ_{ij}^{2} is the covariance between item i and j, n is the number of items in the test, $\mathrm{MinTr}\sum_{\mathrm{T}}$ is the minimal trace of the true score variance, $\mathrm{MaxTr}\sum_{\mathrm{E}}$ is the maximal trace of the true score variance, and $\mathbf{1}'\sum_{\mathrm{X}}\mathbf{1}$ is the total score variance. By this general equation, different lower bounds of reliability can be derived with the different constraints for estimating $\text{MinTr}\sum_{\text{T}}$ or $\text{MaxTr}\sum_{\text{E}}$.

The constraint for deriving Lambda 2 is that the determinant of the true score covariance matrix must be non-negative, which leads to

$$t_i t_j \geq \sigma_{ij}^2 \; (i \neq j), \tag{2}$$

where t_i and t_j are the trace elements of $\sum_{\text{T}}$. Based on the inequality (2), $\text{MaxTr}\sum_{\text{E}}$ is estimated by the following inequality,

$$\sum_{j=1}^{n} \theta_j \leq \sum_{j=1}^{n} \sigma_j^2 - \sqrt{\frac{n}{n-1}\left(\sum_{j=1}^{n} \sum_{i=1\,(i \neq j)}^{n} \sigma_{ij}^2\right)} \;\; , \tag{3}$$

where θ_j represents the jth trace element of $\sum_{\text{E}}$, and n is the number of items. Then from equation (1), we obtain Lambda 2 as

$$\lambda_2 = 1 - \frac{\sum_{j=1}^{n} \sigma_j^2 - \sqrt{\frac{n}{n-1}\left(\sum_{j=1}^{n} \sum_{i=1\,(i \neq j)}^{n} \sigma_{ij}^2\right)}}{\mathbf{1}'\sum_{X}\mathbf{1}}. \tag{4}$$

The greatest lower bound

The greatest lower bound (glb; Woodhouse & Jackson, 1977; Bentler & Woodward, 1980; Ten Berge et al., 1981) for the reliability of the total score on a test uses the constraints that both $\sum_{\text{T}}$ and $\sum_{\text{E}}$ are non-negative definite matrix for estimating $\text{MinTr}\sum_{\text{T}}$ or $\text{MaxTr}\sum_{\text{E}}$. The greatest here means this estimate is theoretically the largest compared with other lower bound estimates, which is also empirically evidenced by the population-level comparison studies related to the glb (Woodhouse & Jackson, 1977; Sijtsma, 2009). Therefore, some researchers (Green & Yang, 2009; Sijtsma, 2009; Ten Berge &Sočan, 2004) recommended the use of the glb to replace coefficient alpha.

Different from the traditional reliability coefficients derived from a single formula, the greatest lower bound is obtained by some algorithm under the constraints that both $\sum_{\mathrm{T}}$ and $\sum_{\mathrm{E}}$ are non-negative definite matrix. These constraints suggest that the quadratic forms $\mathbf{Y}'\sum_{\mathrm{E}}\mathbf{Y}$ and $\mathbf{Y}'\sum_{\mathrm{T}}\mathbf{Y}$ are all equal or larger than zero, where $\mathbf{Y}' = (y_1, y_2, \dots, y_n)'$ is a vector of coordinates in an n-dimensional Cartesian space. Bentler (1972), Woodhouse and Jackson (1977), Bentler and Woodward (1980), and Ten Berge et al. (1981) all contributed to computational algorithms for obtaining $\mathrm{MaxTr}\sum_{\mathrm{E}}$ to get the glb.

Although the calculation of the glb is more complicated because of the algorithm involved, it is worth studying as it is theoretically the optimal lower bound estimate of reliability. A few simulation studies (Cronbach, 1988; Shapiro & Ten Berge, 2000; Ten Berge & Sočan, 2004) have been conducted to examine the sample bias of the glb. Cronbach (1988)'s study was based on one random sample of 100 cases by using the preset factor scores for a 20-item test with 5 strata, and the absolute sample bias of the glb was found to be .087, 1.6% higher than the population value.

Shapiro and Ten Berge (2000) sampled respectively 100, 500, 2000 cases from the multivariate normal distribution. And the given population covariance matrix was obtained based on 5 and 10 items from a scholastic achievement test. The results suggested that the sample bias in the glb increases as the number of test items increases and decreases as the sample size increases. The sample bias was found to range from 1.8% to 13.5% higher than the population values, and the largest bias occurred under the condition of 10 items and 100 cases.

Ten Berge and Sočan (2004) simulated 500 samples of size 100, 250, 500, and 1000 based on the correlation matrix of a real data set with 119 subjects and 6 items.

However, the largest sample mean bias in the glb is .008, only .9% higher than its corresponding population value, under the condition of sample size of 100.

Given mixed results on the sample bias of the glb and no studies that have been conducted to compare the sample performance of the glb and Lambda 2, the present study is intended to further examine the bias of the glb and compare it with that of Lambda 2 and coefficient alpha. We take both the sample level and the population level bias into account by comparing the sample estimates of different reliability estimates from the true reliability.

In this study, we did not include the coefficients derived from a specific factor analytic model or structure equation model, such as McDonald's ω_h and ω_t (Bollen, 1989; Jorekog, 1971; McDonald, 1978, 1999; Raykov & Shrout, 2002) or ρ_{SEM}(Green & Hershberger, 2000; Green & Yang, 2009). The accuracy of this type of coefficients relies on whether the correct model is hypothesized and tested. Due to the subjectivity of choosing a specific model, reliability coefficients based on factor analysis or structure equation modeling are not considered in this study.

Method

The focus of the study is to evaluate and compare the sample bias estimates of coefficient alpha, Lambda 2 and the glb with respect to the true reliability under different simulation conditions. The study examined how the bias in the three lower bound estimates of reliability varies by manipulating three factors: 1) the dimensionality of a test with three levels (one-, two-, three-dimension); 2) the number of test items with two levels (6 and 12); and 3) sample size with five levels (50, 100, 500, 1000, and 2000). The internal structure of each simulated test was determined based on Green and Yang's study (2009). We selected the conditions in which coefficient alpha has the most severe bias under each level of dimensionality. Only the conditions associated with large population bias in alpha were examined because the purpose of the study is to find the alternatives for alpha when it has severe bias. According to findings from Green and Yang (2009), for tests with one dimension, 1/6 items were assigned with loading .8 and 5/6 items with loading .2. For tests with two dimensions, all items were assumed to load on dimension one with loading .3, and only 1/3 items on dimension 2 with loading .6. For tests with three dimensions, similarly, all items were assumed to load on dimension one with loading .3, 1/3 items on dimension 2 with loading .6 and other 1/3 items on dimension 3 with loading .6.

Without the loss of generality, the observed score variance for each item was assigned as one, and an item's true score variance was calculated as the sum of squared item factor loadings. Then the item error variance was equal to 1 minus the true score variance of the item. The sum of the cross product of the factor loadings for any item pair was calculated and equal to the covariance for the two items. In this way, the true score, error score, and observed score variance covariance matrices were formed. And the true reliability was calculated by the definition of reliability as the ratio of the true score variance to the observed score variance. That is,

$$\rho = \frac{\mathbf{1}'\sum_T \mathbf{1}}{\mathbf{1}'\sum_X \mathbf{1}}. \tag{5}$$

The sample observed score matrix was generated by the common factor analytic model as

$$\mathbf{Z}_{\mathrm{sn}} = \mathbf{F}_{\mathrm{sf}}\mathbf{P}_{\mathrm{fn}}{}' + \mathbf{U}_{\mathrm{sn}}\mathbf{D}_{\mathrm{nn}}{}', \tag{6}$$

where $\mathbf{Z}_{\mathrm{sn}}$ is the student-by-item observed score matrix, $\mathbf{F}_{\mathrm{sf}}$ is the student-by-factor common factor score matrix, $\mathbf{P}_{\mathrm{fn}}{}'$ is the factor-by-item common factor loading matrix, $\mathbf{U}_{\mathrm{sn}}$ is the student-by-item unique score matrix, and $\mathbf{D}_{\mathrm{nn}}{}'$ is the item-by-item diagonal matrix of unique factor loadings.

Elements of $\mathbf{F}_{\mathrm{sf}}$ are students' common factor scores, which were randomly generated from the normal distribution with mean of 0 and standard deviation of 1. Elements of $\mathbf{P}_{\mathrm{fn}}{}'$ are the item factor loadings, which were manipulated in the simulation. Elements of $\mathbf{U}_{\mathrm{sn}}$ are student unique scores on each item, which were generated randomly again from the standardized normal distribution. Elements of $\mathbf{D}_{\mathrm{nn}}{}'$ are item unique factor loadings, which were obtained by the square root of one minus the sum of squared common factor loadings of each item. In this study, the unique scores were assumed to be measurement errors, although this may not always hold in practice.

After sample observed score matrices were generated, coefficient alpha and Lambda 2 were computed by their corresponding formulas. Although there is an existing program, called MRFA2 (Ten Berge & Kiers, 2003), which can be used to compute the glb, the program cannot run in batch model thereby prohibiting its use in simulation studies. To calculate the glb for each simulated data set, the computational algorithm described by Ten Berge and Kiers (1991) was programmed with Mathematica7.0 (Wolfram Research). To check the accuracy of our programming codes, results from more than 50 random examples were compared with those calculated by MRFA2, and no discrepancy was found. With Mathematica 7.0, each simulation condition was replicated 2000 times. The means and standard deviations of the estimates of reliability, as well as the bias in each coefficient with respect to the true reliability were reported for each condition as follows.

Results

Table 1 shows the mean and standard deviation of sample estimates of each of the three reliability coefficients for the one-dimension tests manipulated in the simulation, with respect to test length and sample size. The population values of each coefficient are also presented in Table 1. According to the manipulated factorial structure, the true reliability of tests with 6 and 12 items are .386 and .557, respectively. The population values of the glb are equal to the true reliabilities for both the 6- and 12-item tests, while those of Lambda 2 and coefficient alpha are lower than the true reliabilities. The reason that the population values of the glb are equal to the true reliability of the test is that student unique scores (i.e., elements of $\mathbf{U}_{\mathrm{sn}}$) are assumed to be entirely measurement errors. Otherwise, the population values of the glb are expected to be smaller than the true reliabilities. At the sample level, the sample estimates of the glb are higher than the true reliabilities (i.e., positive bias), and the bias becomes smaller as sample size increases. On the other hand, the sample bias of Lambda 2 and coefficient alpha are all negative, meaning they are lower than the corresponding true reliabilities.

TABLE 1

Reliability estimates for tests with one dimension (1/6 items with loading .8, 5/6 items .2)

	Population Value	Sample Estimates				
		n=50	n=100	n=500	n=1000	n=2000
#items=6	True reliability=.386					
Alpha	.343	.315 (.163)	.332 (.103)	.340 (.046)	.342 (.032)	.342 (.022)
Lambda2	.356	.385 (.127)	.375 (.088)	.360 (.043)	.358 (.030)	.357 (.021)
glb	.386	.553 (.106)	.502 (.079)	.433 (.040)	.418 (.028)	.407 (.020)
#items=12	True reliability=.557					
Alpha	.529	.502 (.109)	.521 (.072)	.527 (.031)	.527 (.021)	.529 (.015)
Lambda2	.548	.559 (.090)	.559 (.063)	.550 (.029)	.549 (.020)	.549 (.014)
glb	.557	.784 (.051)	.728 (.045)	.637 (.025)	.613 (.019)	.598 (.014)

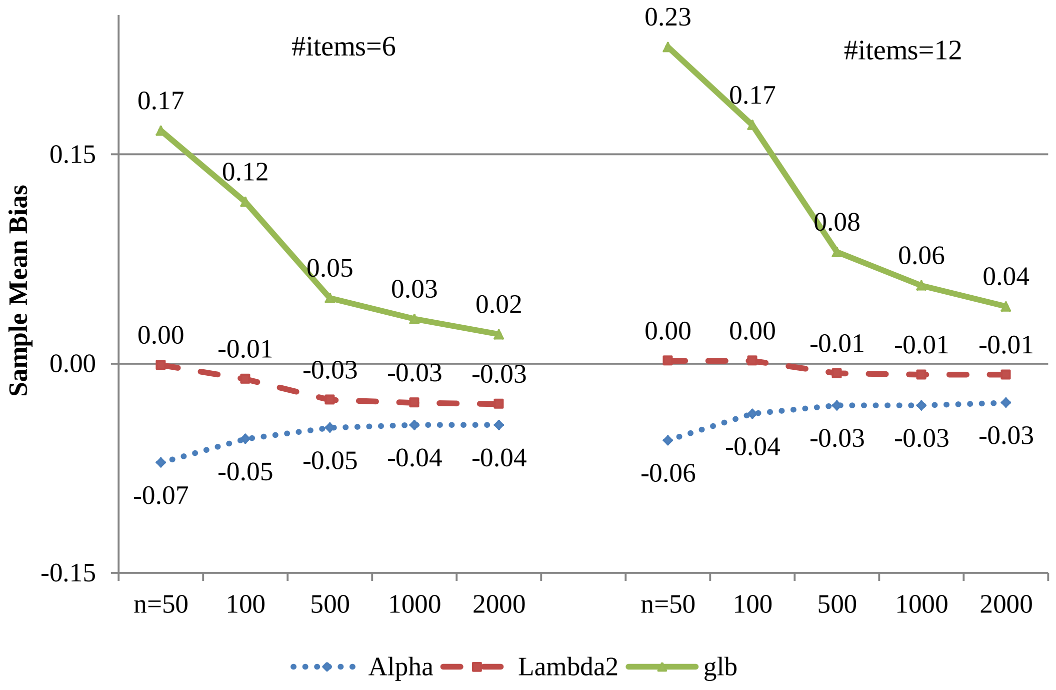

FIGURE 1. *The comparison of sample biases in the three reliability coefficients under the condition of one dimension.*

Figure 1 compares the sample bias of the three reliability coefficients relative to the corresponding true reliabilities of the manipulated tests. The plot on the left panel is for the 6-item test, while the plot on the right panel is for the 12-item test. The glb is severely biased under the sample size of 100 or less, and the bias gradually decreases as the sample size increases. The impact of sample size on the bias of coefficient alpha is also found to be negative in the sense that the bias becomes less severe as sample size increases. However, as sample size increases, the sample bias of Lambda 2 tends to be larger, indicating a positive effect. Although this finding appears to be surprising, it is understandable because the sample bias of Lambda 2 is positive with respect to its population value and the population bias of Lambda 2 is negative in the sense that the population Lambda 2 underestimates the true reliability. When the sample estimate of Lambda 2 is compared with the true reliability, the negative sample bias tends to cancel out the positive population bias. When sample size increases, the sample bias of Lambda 2 relative to the population value becomes smaller and closer to zero, and therefore the total bias (i.e., the bias of sample estimates relative to the true reliability) becomes larger. In addition, the bias of coefficient alpha and Lambda 2 becomes smaller as the number of items increases from 6 to 12. However, the opposite trend is found for the glb.

Across different simulation conditions, Lambda 2 appears to be associated with the smallest absolute bias among the three coefficients, with one exception that the glb is the least biased when the sample size is 2000 and the number of items is six. When the number of items is 12, however, the sample bias in the glb is the largest among the three coefficients even when the sample size goes up to 2000.

Table 2 presents the mean and standard deviation of sample estimates of each of the three reliability estimates for the two-dimension tests manipulated in the simulation, with respect to test length and sample size. The population values of each coefficient are also presented. The true reliability of tests with 6 and 12 items are .497 and .664, respectively. Again, the population values of the glb are equal to the true reliabilities for both the 6- and 12-item tests, while the population values of Lambda 2 and coefficient alpha are lower than the true reliabilities.

TABLE 2

Reliability estimates for tests with two dimensions (Dimension One: all items with loading .3, Dimension Two:1/3 items with loading .6, 2/3 items with loadings 0)

	Population Value	Sample Estimates				
		n=50	n=100	n=500	n=1000	n=2000
#items=6	True reliability=.497					
Alpha	.436	.410 (.132)	.428 (.088)	.434 (.039)	.435 (.028)	.435 (.020)
Lambda2	.455	.470 (.105)	.467 (.076)	.458 (.035)	.457 (.025)	.456 (.018)
glb	.497	.627 (.089)	.590 (.070)	.537 (.035)	.524 (.026)	.516 (.019)
#items=12	True reliability=.664					
Alpha	.627	.612 (.086)	.621 (.056)	.626 (.025)	.625 (.017)	.626 (.012)
Lambda2	.643	.653 (.071)	.649 (.049)	.645 (.023)	.643 (.016)	.643 (.011)
glb	.664	.833 (.040)	.788 (.034)	.721 (.020)	.704 (.015)	.692 (.011)

The comparison of sample biases of the three coefficients is shown in Figure 2. The effects of sample size and number of items on the patterns of sample bias of the three reliability coefficients are consistent with those found from Table 1, which will not be repeated here. For the six-item test, when sample size is 500 or less, Lambda 2 is associated with the least amount of bias. However, when sample size is 1000 or larger, the glb slightly outperforms Lambda 2. For the 12-item test, however, Lambda 2, among the three coefficients, appears to be the best choice consistently across different sample sizes.

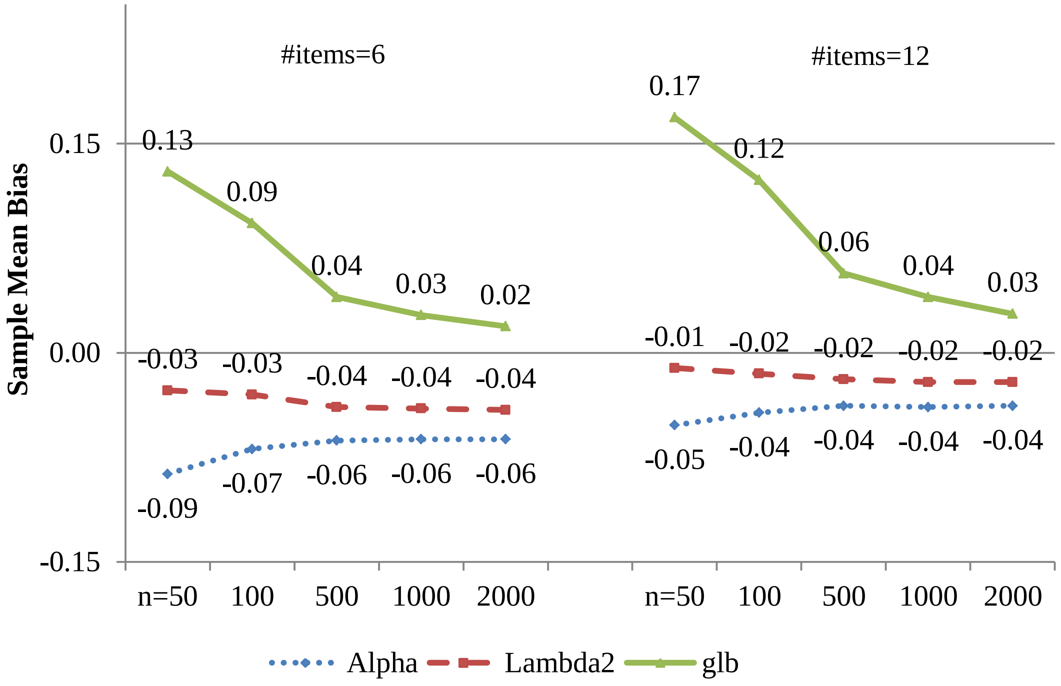

FIGURE 2. *The comparison of sample biases in the three reliability coefficients under the condition of two dimensions*

Table 3 and Figure 3 present the results of different reliability coefficients for tests with three dimensions. The true reliability is equal to .604 and .753 for the 6- and 12-item test respectively. Again, the population values of the glb are equal to the true reliabilities for both tests, while the population values of Lambda 2 and coefficient alpha are lower than the true reliabilities. For the 6-item test, when sample size is 50 or 100, Lambda 2 has the least of amount of bias. However, when sample size is 500 or larger, the glb becomes the least biased coefficient. For the 12-item test, the glb is the least biased only when sample size is 2000, and the absolute bias of the glb is only slightly lower than that of Lambda 2 (i.e., .02 vs. .03).

TABLE 3

Reliability estimates for tests with three dimensions (Dimension One: all items with loading .3, Dimension Two:1/3 items with loading .6, 2/3 items with loadings 0, Dimension Three: 1/3 items with loading .6, 2/3 items with loadings 0)

	Population Value	Sample Estimates				
		n=50	n=100	n=500	n=1000	n=2000
#items=6	True reliability=.604					
Alpha	.490	.470 (.121)	.478 (.083)	.488 (.035)	.489 (.025)	.489 (.018)
Lambda2	.542	.550 (.089)	.544 (.063)	.543 (.028)	.543 (.020)	.542 (.014)
glb	.604	.702 (.073)	.670 (.056)	.632 (.027)	.624 (.021)	.618 (.015)
#items=12	True reliability=.753					
Alpha	.688	.675 (.069)	.682 (.047)	.687 (.020)	.688 (.014)	.688 (.010)
Lambda2	.727	.728 (.055)	.728 (.039)	.726 (.017)	.727 (.012)	.726 (.009)
glb	.753	.871 (.032)	.839 (.027)	.792 (.015)	.781 (.011)	.772 (.008)

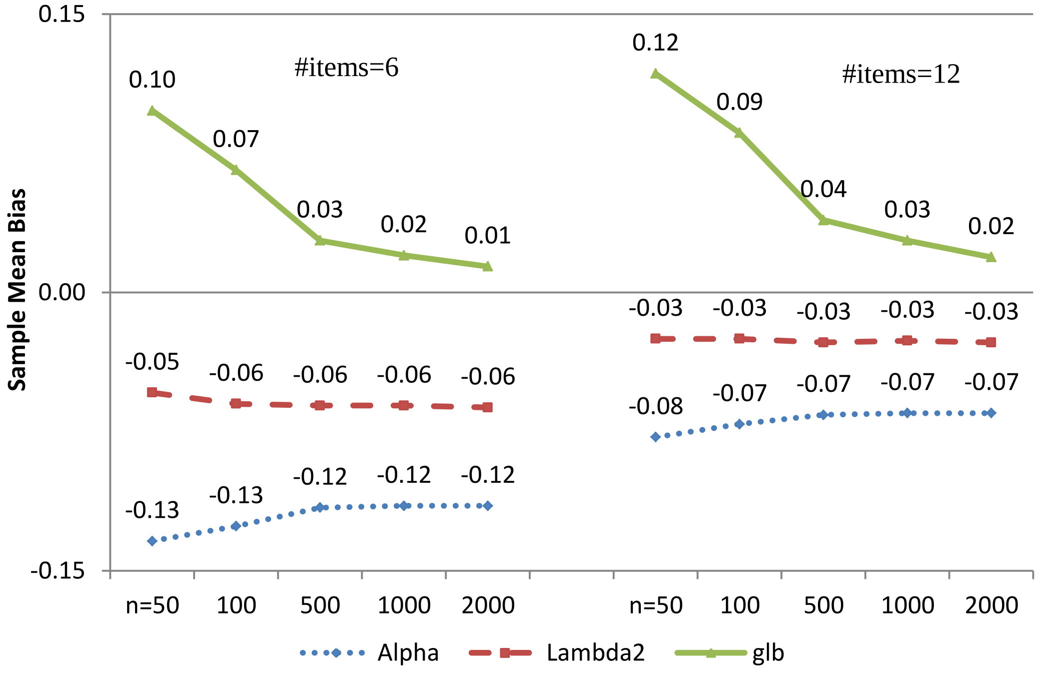

FIGURE 3. *The comparison of sample biases in the three reliability coefficients under the condition of three dimensions*

Discussions

Three factors were manipulated in the present simulation study, including the dimensionality of the test, sample size, and test length. The effect of each manipulated factor on the biases of the three reliability coefficients will be discussed below.

The dimensionality affects the reliability coefficients in this study mainly because the increase of dimensionality is associated with the higher degree of heterogeneity in items. Results show that the more heterogeneous the items are, the more bias in alpha, which are incongruent with the previous research on coefficient alpha (Cortina, 1993; Green & Hershberg, 2000; Green & Yang, 2009; Miller, 1995; Raykov, 1998; Sijtsma, 2009). Similarly, the dimensionality of the test also shows a negative impact on Lambda 2. On the other hand, the sample estimates of the glb become more accurate as items are more heterogeneous.

Among the three coefficients considered in this study, Lambda 2 is the only coefficient whose sample mean bias increases as the sample size become larger, although the effect of sample size is not substantial. Similarly, sample size does not affect the bias in alpha significantly. However, sample size is a critical factor for the bias of the glb, meaning that large sample size is required especially when the number of items is large.

In terms of test length, the accuracy of the sample estimates of coefficient alpha and Lambda 2 improves as the test includes more items. However, for the glb, the effect of test length is negative, which is consistent with the findings from Shapiro and Ten Berge (2000) and Ten Berge and Sočan (2004).

Summary and Conclusions

This study is designed to evaluate and compare two alternatives for coefficient alpha: the greatest lower bound (glb) and Lambda 2. Although some researchers

(Green & Yang, 2009; Sijtsma, 2009; Ten Berge &Sočan, 2004) recommended the use of the glb to replace coefficient alpha, we argue that Lambda 2 may be a better choice. In light of our simulation results, lambda 2 is recommended because it shows the least amount of bias under most of the simulation conditions. Although from the theoretical point of view the population glb is the *greatest* lower bound estimate to reliability, its sample estimate is often severely biased. The glb outperforms Lambda 2 in terms of the absolute bias only when test length is six and sample size is 2000 for one-dimension tests (1000 or above for two-dimension tests). However, the differences are hardly noticeable, .02 at most. When the dimension of the test increases to three and therefore test items become most heterogeneous, the glb has the least absolute bias only when test length is six and sample size is 500 or large. And the difference of the sample biases between the glb and Lambda 2 is up to .05. One may consider the difference of .05 as nontrivial. However, the sample estimates of the glb are always positively biased, which may lead to the unintended consequences where test developers falsely believe that the test is of higher quality than it truly is. In comparison, the negative sample bias of Lambda 2 seems less consequential in the sense that it could challenge test developers to further improve the quality of the test. In addition, Lambda 2 is easily accessible to researchers and practitioners as it is built into SPSS. Therefore, we recommend Lambda 2 to be used as the alternative reliability estimate for coefficient alpha, especially when test items are more heterogeneous and therefore alpha is more severely biased.

The current simulation study considered tests with normally distributed responses. Further studies are needed to examine the properties of the sample bias of Lambda 2 and the glb for non-normally distributed item responses. In addition, only continuous items were generated in this study. In future research, the performance of Lambda 2 needs to be evaluated for tests with Likert scale and dichotomous item responses that are commonly used in educational and psychological measurement.

Another line of future research is to investigate ways of constructing confidence interval for Lambda 2. Building confidence intervals for reliability coefficients has important implications for evaluating the accuracy of the sample estimate of reliability and for comparing different tests, scoring rubrics, or training procedures for raters or observers (Haertel, 2006). Few, if any, studies have been conducted to examine how to construct the confidence intervals for Lambda 2. One potential approach is to use the nonparametric bootstrapping procedures (Efron, 1987; Efron & Tibshirani, 1993) to derive its sampling distribution. The advantage of bootstrapping procedure is that it does not rely on strict distributional assumptions of data. Raykov (1998) proposed to use the nonparametric bootstrapping procedure to approximate empirically the sampling distribution of coefficient alpha. And the simulation study by Cui and Li (2012) suggests that the bootstrapping procedure for constructing the confidence intervals of coefficient alpha shows considerably better performance than different parametric procedures (Feldt, 1965; Hakstian & Whalen, 1976; van Zyl, Neudecker, & Nel's, 2000), especially when item responses are discrete. Future research is needed to evaluate the performance of the bootstrapping procedure for constructing the confidence interval for Lambda 2.

References

Bentler, P.M. (1972). A lower-bound method for the dimension-free measurement of internal consistency. *Social Science Research, 1,* 343-357.

Bentler, P.M. & Woodward, J.A. (1980). Inequalities among lower-bounds to reliability: With applications to test construction and factor analysis. *Psychometrika, 45,* 249-267.

Berge, J.M.F. T., &Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69(4),* 613-625.

Bollen, K. (1989). *Structure Equations With Latent Variables.* New York: Wiley & Sons.

Callender, J. C., &Osburn, H. G. (1979). An empirical comparison of coefficient alpha, guttman's lambda-2, and MSPLIT maximized split-half reliability estimates. *Journal of Educational Measurement, 16*(2), 89-99.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78(1),* 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests.*Psychometrika*, *16*, 297-334.

Cronbach (1988) Internal consistency of tests. *Psychometrika, 53,* 63-70.

Cui, Y. & Li, J.C. (2012). Evaluating the performance of parametric and nonparametric procedures of constructing confidence interval for coefficient alpha: A simulation study. *Journal of British Mathematical and Statistical Psychology.* Advanced online publication. DOI: 10.1111/j.2044-8317.2012.02038.x

Cureton, E. E. (1958). The definition and estimation of test reliability. *Educational & Psychological measurement*, 18, 715-738.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171–185.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357-370.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7*(2), 251-270.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale.*Psychometrika, 74*(1), 121-135.

Guttman, L. (1945). A basis for analyzing test-retest reliability.*Psychometrika, 10*, 255-282.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). (pp. 65-110). Westport, CT: Praeger.

Hakstian, A. R. & Whalen, T. E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika, 41*, 219-231.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items. I: Algebraic lower bounds. *Psychometrika, 42*(4), 567-578.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric test. *Psychometrika, 36,* 109-133.

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

McDonald, R. P. (1999). *Test Theory: A Unified Treatment.* Mahwah, NJ: Erlbaum.

McDonald, R. P. (1978). Generalizability in factorable domains: “Domain validity and generalizability." Educational and Psychological Measurement, *38,* 75-79.

Miller,M.B. (1995). Coefficient alpha: a basic introduction form the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2,* 255-273.

Muchinsky, P.M. (1996) The correlation for attenuation. *Educational & Psychological measurement,* 56, 63-75.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5*(3), 343-355.

Raykov, T. (1998). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement, 22,* 369-374.

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22*(4), 375-385.

Raykov & Shrout, (2002) Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research, 37,* 89-103.

Shapiro, A., & Ten Berge, J. M. F. (2000). The asymptotic bias of minimum trace factor analysis, with applications to the greatest lower bound to reliability. *Psychometrika, 65*(3), 413-425.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74(1),* 107-120.

Spearman, C. (1904). The proof and measurement of association between two things. A*merican Journal of Psychology, 15,* 72-101.

Ten Berge, J. M., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43*(4), 575-579.

Ten Berge, J.M.F., Snijders, T.A.B., & Zegers, F.E. (1981). Computational aspects of the greatest lower bound to reliability and constrained minimum trace factor analysis. *Psychometrika, 46,* 357-366.

Ten Berge, J. M. F., & Kiers, H.A.L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix. *Psychometrika,56,* 309-315.

Ten Berge, J. M. F., & Kiers, H.A.L. (2003). *The minimum rank factor analysis program MRFA* (Internal report). Department of Psychology, University of Groningen, The Netherlands.

Woodhouse, B. & Jackson, P.H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II. A search procedure to locate the greatest lower bound. *Psychometrika, 42,* 579-591.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach'sα, Revelle'sβ, and McDonald'sω_h: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70(1),* 123-133.