Article

Educational and Psychological Measurement
2016, Vol. 76(2) 205–230

DOI: 10.1177/0013164415588133
epm.sagepub.com

Evaluating the Impact of Guessing and Its Interactions With Other Test Characteristics on Confidence Interval Procedures for Coefficient Alpha

Insu Paek[1]

Abstract

The effect of guessing on the point estimate of coefficient alpha has been studied in the literature, but the impact of guessing and its interactions with other test characteristics on the interval estimators for coefficient alpha has not been fully investigated. This study examined the impact of guessing and its interactions with other test characteristics on four confidence interval (CI) procedures for coefficient alpha in terms of coverage rate (CR), length, and the degree of asymmetry of CI estimates. In addition, interval estimates of coefficient alpha when data follow the essentially tau-equivalent condition were investigated as a supplement to the case of dichotomous data with examinee guessing. For dichotomous data with guessing, the results did not reveal salient negative effects of guessing and its interactions with other test characteristics (sample size, test length, coefficient alpha levels) on CR and the degree of asymmetry, but the effect of guessing was salient as a main effect and an interaction effect with sample size on the length of the CI estimates, making longer CI estimates as guessing increases, especially when combined with a small sample size. Other important effects (e.g., CI procedures on CR) are also discussed.

Keywords

guessing, confidence intervals for coefficient alpha, bootstrap confidence intervals for coefficient alpha, reliability

[1]Florida State University, Tallahassee, FL, USA

Corresponding Author:
Insu Paek, Educational Psychology & Learning System, Florida State University, 3204D Stone Building, 1114 W. Call Street, Tallahassee, FL 32306-4453, USA.
Email: ipaek@fsu.edu

Coefficient alpha (Cronbach, 1951) is often used as a point estimate of reliability in practice. (precisely speaking, it is the test score reliability for an essentially tau-equivalent item test, Novick & Lewis, 1967.) Similar to other statistical indices, coefficient alpha has its sampling variation. Reporting just a point estimate does not convey the degree of uncertainty as an estimate. In this regard, reporting the coefficient alpha's interval estimate (i.e., confidence interval [CI]) along with the point estimate is useful. Using such information, one can examine the level of uncertainty associated with the alpha estimate. Several researchers have derived sampling distribution(s) of coefficient alpha and its interval estimation procedures (e.g., Feldt, 1965; Hakstian & Whalen, 1976; Kristof, 1963; Woodruff & Feldt, 1986). Those previous studies used rather a restrictive assumption of compound symmetry for the covariance matrix of items. This is the essentially parallel item assumption that posits equal item variance and the same covariance between all pairs of items. (In the essentially parallel tests (or items) the true score for one test is related by a constant shift to that for the other test while the error score variances for the two tests are the same.) van Zyl, Neudecker, and Nel (2000) derived an asymptotic normal distribution of the maximum likelihood estimator for coefficient alpha under normality without assuming any restrictive assumptions in the covariance matrix for items. Because their asymptotic sampling distribution is free from any constraints in the item covariance matrix, Duhachek, Coughlan, and Iacobucci (2005) and Wainer, Bradlow, and Wang (2007) recommended the use of the 95% CI based on van Zyl et al.'s asymptotic standard error (*SE*) for coefficient alpha.

In this study, the impact of guessing and its interactions with other test characteristics (e.g., in multiple-choice dichotomous-scoring tests) on four CI procedures for coefficient alpha including van Zyl et al.'s was investigated via simulations as a primary investigation of the study. The used CI procedures were van Zyl et al.'s CI (denoted by CI_V), Feldt's (1965) procedure (denoted by CI_F), a nonparametric bootstrap percentile procedure (denoted by CI_{B1}), and a nonparametric bootstrap bias corrected and accelerated procedure (denoted by CI_{B2}). (see for the bootstrap CIs, Efron, 1987, and Efron & Tibshirani, 1994; CI_{B2} was called ''BC_a'' when Efron, 1987, introduced it.) The deleterious effect of guessing on the point estimate of coefficient alpha has been investigated in the measurement literature (MacCann, 2004; Paek, 2015; Zimmerman & Williams, 2003). However, the impact of guessing and its interactions with other variables, in dichotomously scored response data under the influence of guessing, on the coefficient alpha's interval estimates have not been fully examined. Also, widespread use of coefficient alpha in educational achievement testing where multiple-choice tests are popular, and a need for reporting coefficient alpha's uncertainty, provide additional rationale for studying the effect of guessing on the coefficient alpha's interval estimates.

CI_V does not assume any particular structure for the item covariance matrix and is based on large-sample theory, while CI_F is based on an exact sampling distribution under the compound symmetry assumption. The common assumption for both procedures is normality of item response data. Simulated responses in this study are

dichotomous, not normally distributed, and do not assume any specific covariance matrix structure among items. In addition, the effect of guessing is introduced when item responses are generated. Thus, for CI_V and CI_F, assumptions are violated and their utilities are evaluated under more realistic conditions for a multiple-choice test where item responses are contaminated with guessing. Two bootstrap CI procedures are free from distributional and structural assumptions on the item responses and the item covariance structure. CI_{B1} is simpler than CI_{B2} in terms of conceptual understanding and computation. CI_{B2} is $O(N^{-1})$, that is, second-order accurate (as well as second-order correct) whereas CI_{B1} is $O(N^{-\frac{1}{2}})$, that is, first-order accurate. This means that the error for CI_{B2} in matching the true CI approaches zero at the rate of $1/N$ (N being a sample size) whereas the error for CI_{B1} goes to zero at the rate of $1/\sqrt{N}$ (Efron & Tibshirani, 1994). So, CI_{B2} will provide better results than CI_{B2} in general as N increases, but with some increase of computational complexity in its estimation. The four CI procedures are further explained in the ''Method'' section.

The use of the percentile bootstrap method for interval estimates and standard error of reliability under the congeneric condition was suggested in Raykov's (1998) study using a single example for illustration. Pandey and Hubert (1975) explored the Jackknife approaches to estimate the coefficient alpha's CI. Raykov (2002) proposed an analytic standard error of coefficient omega under the congeneric condition, derived via the delta method. Raykov and Marcoulides (2011) suggested the logistic transformation of the omega coefficient and applying the delta-method-based *SE*. Studies by Pandey and Hubert (1975), Raykov (1998, 2002), and Raykov and Marcoulides (2011) including van Zyl et al. (2000), all used the normality assumption to generate their item response data. Yuan, Guarnaccia, and Hayslip (2003) compared two asymptotic theory-based CI estimates for coefficient alpha including CI_V with CI_{B1} and CI_{B2} using a real data set from a 4-point Likert-type style test with a sample size of 419. They indicated that CI_{B2} should be the choice for their sample data. Recently, Padilla, Divers, and Newton (2012) and Padilla and Divers (2013) compared the performances of bootstrap and nonbootstrap CIs of coefficient alpha and omega, respectively, with data generated by the ordinal response factor analysis model. Maydeu-Olivares, Coffman, and Hartmann (2007) also used simulated data based on ordinal response factor analysis model to compare two large-sample theory-based CI procedures for coefficient alpha. Romano, Kromrey, and Hibbard (2010) investigated the performance of eight different nonbootstrap CI procedures of coefficient alpha with data generated by the three-parameter logistic model (3PLM). Studies by Maydeu-Olivares et al. (2007), Padilla et al. (2012), Padilla and Divers (2013), and Romano et al. (2010) are contrasting to the above cited four investigations (Pandey & Hubert, 1975; Raykov, 1998, 2002; Raykov & Marcoulides, 2011; and van Zyl et al., 2000) in that ordinal item responses were simulated in their studies. Although there are a few studies indicating potential robustness of the coefficient alpha's sampling distribution against nonnormality (e.g., Bay, 1973; Feldt, 1965; Yuan & Bentler, 2002), there is no evaluative information available yet for CI_V, CI_F, $CI_{B1,}$ and CI_{B2} with dichotomous data influenced by guessing. Most of all,

none of the studies cited above investigated the impact of guessing and its interactions with other psychometric aspects of a test on the four CI procedures. (The Romano et al. study used the three-parameter model for data simulation, but the guessing factor was not part of their investigation purposes or analyses of the results. In addition, Romano et al. did not include any bootstrap methods in their CI procedures.)

Data under the influence of guessing were generated using the 3PLM (Birnbaum, 1968) with the variations of sample size, test length, the level of coefficient alpha, and the level of guessing. (Details of the study design and simulation factors are described in the ''Method'' section.) Some advantages of using the 3PLM for simulating data influenced by guessing are worthy of noting, though they may not be completely new to those who are familiar with item response theory (IRT) modeling, as they can contrast the approach taken here with most of the previous studies regarding interval estimates for coefficient alpha. First, each item is allowed to have its own guessing level. Though obvious in IRT, this simple realization was not made explicit in the past studies. Second, the guessing realization via the lower asymptote in the 3PLM takes into account not only the simple random guessing behavior but also partial knowledge guessing behavior (Waller, 1989). This better reflects actual guessing behaviors in real data. Third, when item response are dichotomous, bound by 0 and 1, in a usual test where items have different discrimination, difficulty, and guessing levels, the true scores of items are not linearly related to one another, violating the congeneric measurement condition which subsumes the essentially tau-equivalent condition. (See also Zimmerman (1975) for the discussion of the assumptions of the classical test theory [CTT] framework and its interpretation under a probability space.) Fourth, in the binary response case, the error variance of the total test score is not independent of the total score (McDonald, 1999). Simulation by the 3PLM accounts for this functional dependency efficiently and provides flexible and realistic data simulations (in that items have different discriminations, difficulties, and guessing levels) for the current study in the investigation of the impact of guessing on the interval estimates of coefficient alpha.

In summary, the main investigation of this study was on the extent of the effect of guessing and its interactions with other simulation factors on coefficient alpha's CI estimates. This was completed by comparing four different CI procedures (a large-sample–based CI procedure [CI_V], an exact CI [CI_F] under compound symmetry, and two bootstrap CIs [CI_{B1} and CI_{B2}]) in the context of dichotomous item response data under the influence of guessing. Note that item responses in this study were generated such that they were contaminated by guessing, but the data-generating model adhered to the uncorrelated error structure, that is, the current study does not involve the correlated error structure among the observed item scores. To supplement the case of dichotomous data with guessing which violates the essentially tau-equivalent condition, additional data following the essentially tau-equivalent condition were simulated and the behaviors of the four interval estimators of coefficient alpha were investigated for readers who may be curious of the case where the essentially tau-equivalent condition holds.

Method

Coefficient Alpha and Reliability

Coefficient alpha is equal to test score reliability when items in a test (or subtests) follow the essentially tau-equivalent condition,

$$U_j = \tau + c_j + \epsilon_j, \tag{1}$$

where U_j is the jth item response, $\tau + c_j$ is the true score for the jth item response, and ϵ_j is the error for the jth item response. The true scores of the essentially tau-equivalent items are related by a restricted linear relationship, that is, a true score of an item is related to another true score of a different item by a constant c_j. The error variances of essentially tau-equivalent items are not required to be the same, that is, $\sigma^2_{U_j} \neq \sigma^2_{U_k}$ $(j \neq k)$ is permitted. In a factor analysis approach where the observed variable U_j is modeled as $d_j + \lambda_j f + r_j$ where λ_j is the loading, d_j is the intercept, f is a latent factor score, r_j is the residual, and the covariance matrix for U_js is analyzed, the (essentially) tau-equivalent condition is tantamount to the case where one-factor model (with uncorrelated residuals) has equal factor loadings or a unit value for the factor loadings (in which case the variance of the assumed population factor score distribution can be freely estimated). Coefficient alpha underestimates the population reliability when items follow the congeneric condition where the true scores of items are related to one another in a linear fashion, that is,

$$U_j = m_j\tau + c_j + \epsilon_j. \tag{2}$$

In Equation (2), m_j is not necessarily the same across j. The congeneric condition indicates that item loadings are not necessarily the same in the one-factor model (with uncorrelated errors) that uses $U_j = d_j + \lambda_j f + r_j$ with the covariance matrix U_js. Coefficient ω (McDonald, 1999) is a well-known estimator to provide the population reliability in this congeneric condition and coefficient $\omega \geq$ coefficient alpha under the congeneric condition. When the independence assumption of errors is violated, it is easily demonstrated that both coefficient alpha and coefficient ω can overestimate the population reliability (Kano & Azuma, 2003).

Suppose that $f(\cdot)$ is measurable and invertible. Borrowing from the generalized linear model framework, the observed response can be expressed as

$$U_j = f^{-1}\left(a_j\tau + c_j\right) + \epsilon_j, \tag{3}$$

where $f^{-1}(\cdot)$ is an inverse link function and the link function relates the $E\left(U_j\right) = \tau_j$ to linear predictor(s), that is, $f\left(\tau_j\right) = a_j\tau + c_j$. In Equation (2) under the congeneric condition, $f(\cdot)$ is an identity function, thus $\tau_j = a_j\tau + c_j$ and ϵ_j is not related to τ_j. The use of identity link function is commonly made when U_j is a continuous, typically normal variable in generalized linear model. In CTT, no particular parametric distributional assumption is imposed when an observed score is presented by the true score plus error although the normality assumption seems to be imposed sometimes

when a CI is constructed with the intention of stating a certain probability level using the conventional CTT standard error of measurement formula. When U_j is a dichotomous variable following the Bernoulli distribution, the typical canonical link function which brings a satisfactory analysis is the logit link function (see, McCullagh & Nelder, 1989). Thus, $f(\tau_j) = \text{logit}(\tau_j) = \log[\tau_j/(1-\tau_j)]$. Note that τ_j is the probability of correct answer for U_j in this dichotomous case. This means $\tau_j = \exp(a_j\tau + c_j)/[1+\exp(a_j\tau + c_j)]$, that is, the item true scores are nonlinearly related to each other. In addition, ϵ_j is not independent of τ_j any longer because the variance of error becomes the function of τ_i, that is, $var(\epsilon_j) = \tau_j(1-\tau_j)$. For a dichotomous U_j, Equation (2) may be viewed as a linear approximation to the nonlinear true score relationship that exists in the dichotomous items.

When dichotomous item responses are from a multiple-choice test, examinee guessing may exist and can contaminate item responses. The 3-parameter logistic (3PL) IRT model (Birnbaum, 1968) provides an explicit way to accommodate guessing in dichotomously scored categorical data. Because of the guessing parameter modeling for each item, the 3PLM cannot be expressed using the logit link function under the generalized linear model framework. The expected item score or the true item score under the 3PLM is

$$\tau_j = P_j(\theta) = g_j + (1-g_j)\exp(a_j(\theta - b_j))/[1+\exp(a_j(\theta - b_j))], \quad (4)$$

where a_j, b_j, and g_j are item slope (or discrimination), difficulty, and guessing (or lower asymptote) parameter, respectively, and θ is a person latent trait. Note that the true score under IRT is decided by item guessing, discrimination, difficulty, and person latent trait score. Because the true score in IRT is a monotonic nonlinear transformation of θ, and because in a multiple-choice test usually item guessing, discrimination, and difficulty vary across items, the item true scores are not linearly related. This nonlinearity also implies that the correlation between the two true item scores is not necessarily one, which is a violation of a corollary (i.e., $\rho_{\tau_j\tau_k} = 1$ where $j \neq k$) in the congeneric condition. If all items have the same guessing ($g_j = g$), discrimination ($a_j = a$), and difficulty ($b_j = b$), then the implied population observed score covariance matrix can be straightforwardly shown to be compound symmetric, which is also the implied covariance matrix by the essential parallel condition in CTT.

For a nonessentially tau-equivalent item test, coefficient alpha is an inconsistent reliability estimator (Raykov, 2012), that is, increasing sample size does not make coefficient alpha converge to the population reliability in probability. In this theoretical sense, coefficient alpha is not an optimal choice as a reliability estimator for dichotomous response data. However, even a reliability based on the congeneric condition, such as coefficient ω, is not an appropriate reliability either for dichotomous data having examinee guessing, from the theoretical point of view discussed above. Although psychometric researchers may exercise flexibility in formulating and calculating a specific reliability formula that fits the data best in the sense of dimensionality, item response function form, population latent trait distribution, and level of

measurement in the observed response data (e.g., categorical or at least approximately continuous item response data), the current practice for reporting reliability, especially in educational achievement testing programs which use often multiple-choice tests, is still dominated by the use of coefficient alpha (e.g., California Department of Education, 2014; Integrated Louisiana Educational Assessment Program, 2014; Michigan Department of Education, 2013; Nebraska State Accountability [NeSa], 2014; New York State Education Department, 2013). The investigation of the interval estimators for coefficient alpha in this study was based on practical grounds rather than theoretical consideration, namely that coefficient alpha is almost always the choice for reporting reliability (although we do know it is not an optimal reliability estimator) for multiple-choice tests in many educational assessment programs, and that there exists a practical desire to report the level of uncertainty related to the point estimate of coefficient alpha in an operational testing program setting.

Study Design

This study consists of two parts. The first investigation was when data are dichotomous responses contaminated by guessing, which is the major focus. As discussed in the previous section, having differential guessing, discrimination, and difficulty across items in dichotomous response data violates the congeneric condition, therefore the essentially tau-equivalent condition is also violated. The second investigation was conducted when data follow essentially tau-equivalent condition, thus coefficient alpha is a theoretically legitimate reliability estimator for the population reliability.

When Data Are Dichotomous Responses Under the Influence of Guessing. In addition to the different levels of guessing, different sample sizes (N = 100, 300, and 500), test lengths (P = 21 and 41), and different coefficient alpha levels were used to generate simulation data. The levels of coefficient alpha were represented by the population coefficient alpha values when there is no guessing, which were .8 and .9. The guessing factor was represented by overall guessing amount in a test, an average item guessing. Here g = 0 (no guessing), .25 (e.g., for a 4-option multiple-choice item test), and .33 (which is a high-guessing case, e.g., a 3-option multiple-choice item test). Thus, there were a total of 36 data simulation conditions (guessing × sample size × test length × coefficient alpha = 3 × 3 × 2 × 2 = 36). The four CI procedures (CI_V, CI_F, CI_{B1}, and CI_{B2}) were applied to each of these 36 conditions. In each condition, 1,000 data sets were replicated. For CI_{B1}, and CI_{B2}, the number of bootstrap samples for a given data set was 2,000. (See, e.g., Davison & Hinkley, 1997, and Efron & Tibshirani, 1994, for the number of bootstrap samples.) Also, the simulation of the above design was independently repeated 10 times. The means of the simulation summary measures, which are described below, were obtained for each condition in each of the 10 independent runs and were used to analyze the simulation results using analysis of variance (ANOVA) and an effect size approach.

As a major performance summary measure for the four CI procedures, coverage rate (CR) was calculated, which represents the proportion of the time for which a CI estimate includes a population coefficient alpha. When there is guessing, it is known that coefficient alpha decreases (e.g., MacCann, 2004; Paek, 2015; Zimmerman & Williams, 2003). The population value of coefficient alpha as a function of guessing (and other psychometric properties of a test) was calculated using the following formula (Paek, 2015),

$$C_\alpha = \left(\frac{P}{P-1}\right)\left(1-\frac{\sum \sigma_j^2}{\sigma_X^2}\right) = \left(\frac{P}{P-1}\right)\left(1-\frac{\sum\left[\int P_j(\theta)h(\theta)d\theta \int Q_j(\theta)h(\theta)d\theta\right]}{\int \sum P_j(\theta)Q_j(\theta)h(\theta)d\theta + \int\left[\sum P_j(\theta)\right]^2 h(\theta)d\theta - \left[\int \sum P_j(\theta)h(\theta)d\theta\right]^2}\right), \quad (5)$$

where $P_j(\theta)$ is the 3-parameter logistic (3PL) item response function with the scaling factor $D = 1.7$; $Q_j(\theta) = 1 - P_j(\theta)$; and $h(\theta)$ is a person ability population distribution, which was the standard normal distribution here. For each condition, the value of the population coefficient alpha was obtained by evaluating the above formula using numerical integration (θ ranging from -4 to 4 by 0.01 increment) with the simulation data-generating parameters.

To aid understanding of the behavior of the CI estimates under the influence of guessing, two additional summary measures, Length and Shape, were calculated. Length and Shape measures were used by Efron and Tibshirani (1994) and can deliver more information on the behaviors of different CI procedures. CR, Length, and Shape summary measures are defined as follows:

$$\text{CR} = 100\times \frac{\sum_k I[C_\alpha \in CI_k | \text{Nominal Type I error} = \lambda]}{K}, \quad (6)$$

where C_α is a population coefficient alpha, CI_k is the kth replication CI ($k = 1, 2, \ldots, K$), and $I[\cdot]$ is an indicator function (1 if $C_\alpha \in CI_k$; 0 otherwise). Under the nominal Type I error, $\lambda = .05$, a good CI procedure should produce a CR value close to $100(1-\lambda)\%$, that is, 95% in this case:

$$\text{Length}_k = \hat{\xi}_{Uk} - \hat{\xi}_{Lk}, \quad (7)$$

where $\hat{\xi}_{Uk}$ and $\hat{\xi}_{Lk}$ are the respective lower bound and the upper bound estimates for CI_k.

$$\text{Shape}_k = \frac{\hat{\xi}_{Uk} - \hat{C}_\alpha^k}{\hat{C}_\alpha^k - \hat{\xi}_{Lk}}, \quad (8)$$

where $\hat{C}_{\alpha}^{k}$ is the coefficient alpha estimate for the kth replication. The Shape measure indicates the degree of asymmetry of the CI interval around the point estimate of coefficient alpha. A value of the Shape measure greater than 1 shows that the distance between the upper bound CI estimate and the coefficient alpha point estimate is larger than that between the lower bound CI estimate and the estimated coefficient alpha.

CI_V is symmetric about $\hat{C}_{\alpha}^{k}$ while CI_F, CI_{B1}, and CI_{B2} can be asymmetrical. Therefore, the value of the Shape measure for CI_V is always 1 across all conditions. As mentioned before, average Length and average Shape across K replications were calculated and denoted by Length and Shape without the subscript k.

When Data Follow the Essentially Tau-Equivalent Condition. The same data simulation factors (N = 100, 300, and 500; P = 21 and 41; and C_{α} = .8 and .9) were used. Note that there is no item guessing factor in this simulation. The τ and c_j values were randomly drawn from the standard normal distribution. The error (ϵ_j) was randomly selected from $N(0, \sigma_{\epsilon_j}^2)$. For the ease of description of the simulation procedure for the error, let $X = \sum U_j (j = 1, 2, \ldots, L)$ and $X = T + E$. The one-factor model congeneric condition indicates $\sigma_X^2 = \sigma_T^2 + \sigma_E^2 = L^2 + \sigma_E^2$ for the essentially tau-equivalent condition. Based on this with C_{α} = .8 or .9, σ_E^2/L was calculated and used as a guideline to set up the minimum and maximum of uniform distributions where the value of $\sigma_{\epsilon_j}^2$ was randomly selected from a uniform distribution: $\sigma_{\epsilon_j}^2 \sim unif(2, 2.6)$ for C_{α} = .9 and L = 21; $\sigma_{\epsilon_j}^2 \sim unif(4.9, 5.6)$ for C_{α} = .8 and L = 21; $\sigma_{\epsilon_j}^2 \sim unif(4.2, 4.9)$ for C_{α} = .9 and L = 41; and $\sigma_{\epsilon_j}^2 \sim unif(9.9, 10.6)$ for $C_{\alpha\ =\ .8}$ and L = 41. The tau-equivalent item responses were generated using Equation (1) with the selected τ, c_j, and ϵ_j values from the above steps.

The four CI procedures were applied to those essentially tau-equivalent data and the same summary measures (CR, Length, and Shape) were calculated for each condition.

Four CI Procedures

CI_V is given by

$$[\hat{C}_{\alpha} - Z_{(\lambda/2)}\hat{S}, \hat{C}_{\alpha} + Z_{(1-\lambda/2)}\hat{S}], \tag{9}$$

where $Z_{(\lambda/2)}$ and $Z_{(1-\lambda/2)}$ are the $100(\lambda/2)$th and $100(1-\lambda/2)$th quantiles of the standard normal distribution, and $\hat{S}$ is the standard error estimate for the coefficient alpha defined by

$$\hat{S} = \sqrt{(1/N)\left[\frac{2P^2}{(P-1)^2(\mathbf{1}'\hat{\mathbf{\Sigma}}\mathbf{1})^3}\right]\left[(\mathbf{1}'\hat{\mathbf{\Sigma}}\mathbf{1})(tr\hat{\mathbf{\Sigma}}^2 + tr^2\hat{\mathbf{\Sigma}}) - 2(tr\hat{\mathbf{\Sigma}})(\mathbf{1}'\hat{\mathbf{\Sigma}}^2\mathbf{1})\right]}, \tag{10}$$

were N (sample size) and P (test length) are defined as before, $\mathbf{1}$ is a $P\times$ 1 vector of 1s, ''tr'' represents trace, and $\hat{\mathbf{\Sigma}}$ is a $P\times P$ estimated interitem covariance matrix:

CI_F is given by

$$[1-(1-\hat{C}_\alpha)F_{(1-\lambda/2)}, 1-(1-\hat{C}_\alpha)F_{(\lambda/2)}], \tag{11}$$

where $F_{(1-\lambda/2)}$ and $F_{(\lambda/2)}$ are the 100(1 $-\lambda/2$)th and 100($\lambda/2$)th quantiles of the central F distribution with ($N-$ 1) and ($N-$ 1)($P-$ 1) degrees of freedom.

CI_{B1} and CI_{B2} are constructed based on a nonparametric bootstrap method where a desired number of samples are independently drawn from an empirical distribution function with replacement. Let a $N \times P$ data matrix be $\mathbf{X}=(\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x}_N)'$ where $\mathbf{x_i}$ is a 1 $\times$ P row vector which is the ith person's response vector (i = 1, 2, . . ., N). A bootstrap sample, $\mathbf{X}_*=(\mathbf{x}_1^*, \mathbf{x}_2^*, \cdots, \mathbf{x}_N^*)'$ was constructed by drawing a random sample of response vectors from $\mathbf{X}$ with replacement. For the rth bootstrap sample, $\mathbf{X}_*^r$, (r = 1, 2, . . ., R) where R is the desired number of bootstrap samples, the bootstrap replication of coefficient alpha, $\hat{C}_\alpha^r$ is computed. With all obtained values of $\hat{C}_\alpha^r$, the bootstrap distribution function of $\hat{C}_\alpha^r$, denoted by $\hat{G}$, can be constructed. Then CI_{B1} is defined by

$$[\hat{G}^{-1}_{(\lambda/2)}, \hat{G}^{-1}_{(1-\lambda/2)}], \tag{12}$$

where $\hat{G}^{-1}_{(\lambda/2)}$ and $\hat{G}^{-1}_{(1-\lambda/2)}$ are the 100(λ/2)th and 100(1 $-\lambda/2$)th quantiles of $\hat{G}$. CI_{B2} is an adjusted percentile CI procedure for which bias correction is made and acceleration is taken into account, which is the rate of change of the standard error of $\hat{C}_\alpha^r$ relative to the true coefficient alpha C_α, that is, adjusting for the skewness of the sampling distribution of coefficient alpha. CI_{B2} is given by

$$[\hat{G}^{-1}_{(\lambda_1)}, \hat{G}^{-1}_{(\lambda_2)}], \tag{13}$$

where $\lambda_1$and λ_2 are defined by

$$\lambda_1=\Phi\left(\hat{z}_0+\frac{\hat{z}_0+z_{(\lambda/2)}}{1-\hat{k}(\hat{z}_0+z_{(\lambda/2)})}\right), \tag{14}$$

and

$$\lambda_2=\Phi\left(\hat{z}_0+\frac{\hat{z}_0+z_{(1-\lambda/2)}}{1-\hat{k}(\hat{z}_0+z_{(1-\lambda/2)})}\right), \tag{15}$$

where $\Phi(\cdot)$ is the standard normal distribution function and $z_{(\lambda/2)}$ and $z_{(1-\lambda/2)}$ are the 100(λ/2)th and 100(1 $-\lambda/2$)th quantiles of a standard normal distribution. $\hat{z}_0$ and $\hat{k}$ are the bias correction and the acceleration factor, respectively. When $\hat{z}_0$ and $\hat{k}$ are zero, CI_{B2} is equivalent to CI_{B1}. $\hat{z}_0$ and $\hat{k}$ are estimated by

$$\hat{z}_0=\Phi^{-1}[\hat{G}(\hat{C}_\alpha)], \tag{16}$$

and

$$\hat{k} = \frac{\sum_i \left(\hat{C}_{\alpha(\cdot)} - \hat{C}_{\alpha(-i)}\right)^3}{6\left[\sum_i \left(\hat{C}_{\alpha(\cdot)} - \hat{C}_{\alpha(-i)}\right)^2\right]^{3/2}}, \tag{17}$$

where $\Phi^{-1}(\cdot)$ is the inverse function of $\Phi(\cdot)$ and $\hat{G}(\cdot)$ is again the bootstrap cumulative distribution function. There are different ways to estimate $\hat{k}$. In this study, the Jackknife approach to estimate $\hat{k}$ (Efron & Tibshirani, 1994), shown in Equation (17), was used. $\hat{C}_{\alpha(-i)}$ is a coefficient alpha estimate without the ith response vector in the data matrix. $\hat{C}_{\alpha(\cdot)}$ is defined as

$$\hat{C}_{\alpha(\cdot)} = (1/N)\sum_i \hat{C}_{\alpha(-i)}. \tag{18}$$

Data-Generating Parameters in the 3PLM

Item difficulty parameters (b_j; j = 1, 2, . . ., J) were selected such that they ranged from −2 to 2 by .2 (21-item test) and by .1 (41-item test). Guessing parameters (g_j) were generated from truncated normal distributions: $g_j \sim N(.25, 0.1^2$, lower = 0, upper = .5) for g = .25, where the lower and the upper represent the cutoffs; $g_j \sim N(.33, .1^2$, lower = 0, upper = .5) for g = .33; and g_j = 0 when g = 0. Item slope parameters (a_j) were also generated from truncated normal distributions such that the selected slope parameters (with the other selected b_j and $g_j g_j$ values) lead to the population coefficient alpha values of .8 and .9 when there is no guessing. The truncated normal distributions used for selecting the item slope parameters were $a_j \sim N(.7, .2^2$, lower = .3, upper = 1.5) for the population coefficient α = .8 with no guessing and the test length = 21; $a_j \sim N(1.6, .2^2$, lower = .3, upper = 1.5) for the population coefficient α = .9 with no guessing and the test length = 21; $a_j \sim N(.4, .2^2$, lower = .2, upper = .8) for the population coefficient α = .8 with no guessing and the test length = 41; and $a_j \sim N(.8, .2^2$, lower = .3, upper = 1.5) for the population coefficient α = .9 with no guessing and the test length = 41. The ability θs were generated from $N(0, 1)$. Item responses were generated following the standard IRT data generation procedure.

Analysis of Simulation Results

The present study used five factors which are CI (CI_V, CI_F, CI_{B1}, and CI_{B2}), guessing (g = 0, .25, and .33), sample size (N = 100, 300, and 500), test length (P = 21 and 41), and different population coefficient alpha levels (C_α = .8, and C_α = .9 when there is no guessing) for the first part of the investigation. (The second part of the investigation, when data are from the essentially tau-equivalent condition, does not have the guessing factor.) A major performance measure of the CI procedures was CR. In addition, Length and Shape measures were calculated to assist investigating

the behavior of CI estimates. The same data set was analyzed by the different CI procedures. The CI factor is a within-subjects variable while all the other factors are between-subjects variables. The full factorial ANOVA for one within-subjects and four between-subjects variable repeated measures design (and for one within-subjects and three between-subjects design for the second part of the investigation) was conducted for CR, Length, and Shape, respectively. Because of some degree of violation of the repeated measures ANOVA assumptions (e.g., normality, homoscedasticity, and sphericity), the interpretation of the obtained p values and their statistical significance is not always straightforward (and also many effects were statistically significant under the .05 significance level). Therefore, an effect size approach was used as a primary criterion to select important effects from ANOVA. The effect size considered here was eta square (η^2), which is the ratio of sum of squares for an effect to the total sum of squares. The eta square is computationally efficient with ease of understanding and provides a measure of unique contribution by an effect in the total variation in the present study design. Cohen (1988) provided .01, .06, and .14 for the classification of small, medium, and large η^2 values, respectively. However, this classification was based on a simple independent two-group comparison case (i.e., a single-factor design having two levels). The present study design is more complex and involves more than two levels. In this study, the η^2 was classified as small, medium, and large, whose cutoffs were .01 (1%), .09 (9%), and .25 (25%), respectively, based on Cohen's (1988) classification of correlation as effect size (.1, .3, and .5 as small, medium, and large) and the relationship between eta squares and the coefficient of determination.

All computations including the programming of the CI procedures, data simulations, and analyses were done using the R language (R Development Core Team, 2014).

Results

When Data are Dichotomous Responses Under the Influence of Guessing

Coverage Rate (CR). The ANOVA table for CR is presented in Table 1. The total variation in this ANOVA analysis is divided into the variation due to between subjects and the variation due to within subjects.

Three main effect terms and one interaction term showed small or large effects. The largest effect was CI ($\eta^2 = .596$) and explained more than 59% of the total variation. C_α and sample size (N) had small effect sizes ($\eta^2 = .043$ and .011, respectively). The interaction between CI and C_α also showed a small effect ($\eta^2 = .028$). The guessing factor and its interactions with other factors did not show any salient effects on CR.

The marginal means for CI were .944, .947, .966, and .963 for CI_{B1}, CI_{B2}, CI_F, and CI_V, respectively. On average, CI_{B1} and CI_{B2} were lower than .95 while CI_F and CI_V were higher than .95. Of the two nonbootstrap CIs, CI_V was closer to .95 on average than CI_F. The marginal mean of CR, when $C_\alpha = .9$, was .957 slightly higher than

Table 1. Effect Sizes and ANOVA Result for CR When Item Responses are Dichotomous Under the Influence of Guessing.

Source	*df*	SS	MS	*F*	*p* Value	η^2
Between subjects	359	0.067				
g	2	0.001	0.001	3.853	.022	.005
C_α	1	0.009	0.009	65.925	<.01	.043*
P	1	<0.001	<0.001	3.093	.080	.002
N	2	0.002	0.001	8.718	<.01	.011*
g:C_α	2	0.001	<0.001	3.294	.038	.004
g:*P*	2	0.001	0.001	4.085	.018	.005
C_α:*P*	1	<0.001	<0.001	0.849	.358	.001
g:*N*	4	<0.001	<0.001	0.269	.898	.001
C_α:*N*	2	0.001	<0.001	3.206	.042	.004
P:*N*	2	0.001	<0.001	2.662	.071	.003
g:C_α:*P*	2	0.001	0.001	4.119	.017	.005
g:C_α:*N*	4	<0.001	<0.001	0.662	.619	.002
g:*P*:*N*	4	0.001	<0.001	1.448	.218	.004
C_α:*P*:*N*	2	0.001	<0.001	3.115	.046	.004
g:C_α:*P*:*N*	4	<0.001	<0.001	0.572	.683	.002
Residuals	324	0.046	<0.001			
Within subjects	1,080	0.151				
CI	3	0.130	0.043	4400.828	<.01	.596***
CI:*g*	6	0.001	<0.001	15.163	<.01	.004
CI:C_α	3	0.006	0.002	203.853	<.01	.028*
CI:*P*	3	0.001	<0.001	16.865	<.01	.002
CI:*N*	6	0.002	<0.001	27.424	<.01	.007
CI:*g*:C_α	6	0.001	<0.001	8.803	<.01	.002
CI:*g*:*P*	6	0.001	<0.001	10.296	<.01	.003
CI:C_α:*P*	3	<0.001	<0.001	2.263	.080	<.001
CI:*g*:*N*	12	<0.001	<0.001	1.585	.090	.001
CI:C_α:*N*	6	<0.001	<0.001	3.307	<.01	.001
CI:*P*:*N*	6	<0.001	<0.001	1.094	.364	<.001
CI:*g*:C_α:*P*	6	<0.001	<0.001	5.158	<.01	.001
CI:*g*:C_α:*N*	12	<0.001	<0.001	0.697	.756	<.001
CI:*g*:*P*:*N*	12	<0.001	<0.001	0.789	.662	<.001
CI:C_α:*P*:*N*	6	<0.001	<0.001	1.467	.186	<.001
CI:*g*:C_α:*P*:*N*	12	<0.001	<0.001	0.552	.880	<.001
Residuals	972	0.010	<0.001			
Total	1,439	0.219				

Note. CI = confidence interval; CR = coverage rate; MS, mean square; SS = sum of squares. "*," "**," and "***" represent small, medium, and large effect size, respectively.

.952, which is the marginal means of CR when C_α = .8. The effect of sample size was also small and made only slight differences among the marginal means of CR (.953, .956, and .955) for N = 100, 300, and 500. The interaction of CI with C_α is presented in Figure 1.

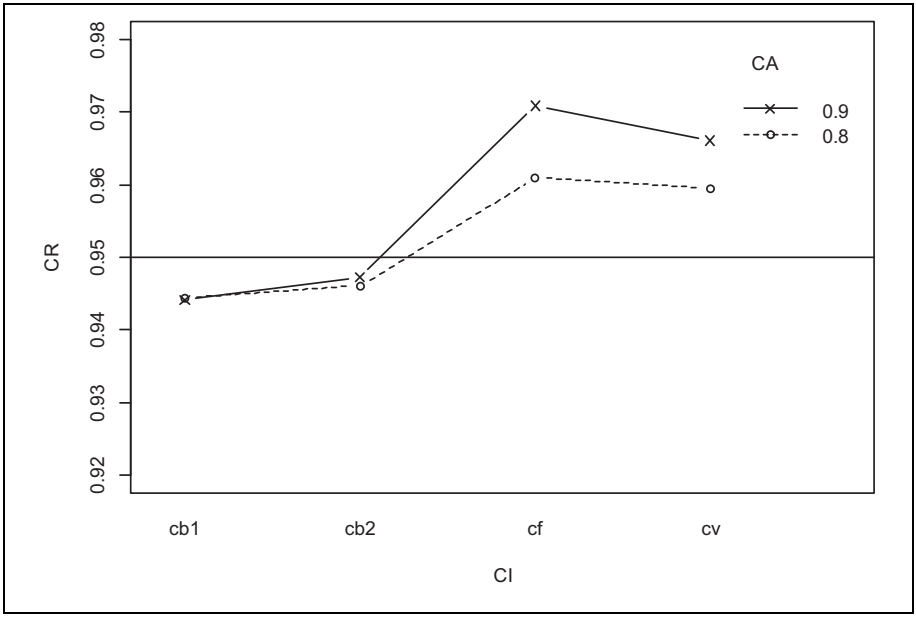

Figure 1. Interaction between CI and C_α on CR when item responses are dichotomous under the influence of guessing.

Note. "cb1," "cb2," "cf," and "cv" represent CI_{B1}, CI_{B2}, CI_F, and CI_V. "CR" represents coverage rate while "CA" means C_α.

The differences in the four CI procedures depended on the value of C_α. When C_α = .9, larger differences among the CI procedures were observed than when C_α = .8. Also, from Figure 1, it is clear that CI_{B1} and CI_{B2} were mimimally affected by the level of C_α while CI_F and CI_V exhibited their dependency on the level of C_α. CI_F showed the largest performance difference when C_α changed from .8 to .9.

Length (LE). The same (one within-subjects and four between-subjects) ANOVA for repeated measures design was applied to each of the Length and Shape summary measures. Table 2 shows the ANOVA result summaries and the effect-size classification for Length and Shape. For the ANOVA results, the degrees of freedom (*df*), sum of squares (SS), and *p* values were reported.

For Length, five terms showed at least small or larger effect size: sample size (*N*, large effect size, η^2 = .438), guessing (*g*, large effect size, η^2 = .264), C_α (medium effect size, η^2 = .214), interaction between guessing and sample size (*g:N*, small effect size, η^2 = .034), and interaction between C_α and *N* (C_α:*N*, small effect size, η^2 = .028). Guessing made noticeable differences in Length as a main effect and a small effect size on Length as an interaction effect with sample size, but its interactions with CI or other factors were not salient enough to be considered important.

Table 2. Effect Sizes and ANOVA Results for Length and Shape When Item Responses Are Dichotomous Under the Influence of Guessing.

		Length			Shape		
Source	*df*	SS	*p* Value	η^2	SS	*p* Value	η^2
Between subjects	359	3.374			1.939		
g	2	0.893	<.01	.264***	<0.001	<.01	<.001
C_α	1	0.724	<.01	.214**	0.001	<.01	<.001
P	1	0.030	<.01	.009	<0.001	<.01	<.001
N	2	1.481	<.01	.438***	1.937	<.01	.157**
g:C_α	2	0.009	<.01	.003	<0.001	<.01	<.001
g:*P*	2	0.015	<.01	.004	<0.001	<.01	<.001
C_α:*P*	1	0.003	<.01	.001	<0.001	<.01	<.001
g:*N*	4	0.116	<.01	.034*	<0.001	<.01	<.001
C_α:*N*	2	0.094	<.01	.028*	<0.001	<.01	<.001
P:*N*	2	0.004	<.01	.001	<0.001	.550	<.001
g:C_α:*P*	2	0.001	<.01	<.001	<0.001	<.01	<.001
g:C_α:*N*	4	0.001	<.01	<.001	<0.001	<.01	<.001
g:*P*:*N*	4	0.002	<.01	.001	<0.001	<.01	<.001
C_α:*P*:*N*	2	<0.001	<.01	<.001	<0.001	.266	<.001
g:C_α:*P*:*N*	4	<0.001	<.01	<.001	<0.001	<.01	<.001
Residuals	324	<0.001			0.0006		
Within subjects	1,080	0.011			10.393		
CI	3	0.009	<.01	.003	9.45	<.01	.766***
CI:*g*	6	<0.001	<.01	<.001	0.004	<.01	<.001
CI:C_α	3	<0.001	<.01	<.001	0.006	<.01	<.001
CI:*P*	3	<0.001	<.01	<.001	0.003	<.01	<.001
CI:*N*	6	0.001	<.01	<.001	0.926	<.01	.075*
CI:*g*:C_α	6	<0.001	<.01	<.001	<0.001	<.01	<.001
CI:*g*:*P*	6	<0.001	<.01	<.001	<0.001	<.01	<.001
CI:C_α:*P*	3	<0.001	<.01	<.001	0.003	<.01	<.001
CI:*g*:*N*	12	<0.001	<.01	<.001	<0.001	<.01	<.001
CI:C_α:*N*	6	<0.001	<.01	<.001	<0.001	<.01	<.001
CI:*P*:*N*	6	<0.001	<.01	<.001	<0.001	<.01	<.001
CI:*g*:C_α:*P*	6	<0.001	<.01	<.001	<0.001	<.01	<.001
CI:*g*:C_α:*N*	12	<0.001	<.01	<.001	<0.001	<.01	<.001
CI:*g*:*P*:*N*	12	<0.001	<.01	<.001	<0.001	<.01	<.001
CI:C_α:*P*:*N*	6	<0.001	<.01	<.001	<0.001	<.01	<.001
CI:*g*:C_α:*P*:*N*	12	<.001	<.01	<.001	<.001	<.01	<.001
Residuals	972	<0.001			0.001		
Total	1,439	3.385			12.332		

Note. SS = sum of squares; CR = coverage rate. "*," "**," and "***" represent small, medium, and large effect size, respectively.

The marginal means of Length for guessing were .056, .098, and .114, respectively, for $g = 0$, .25, and .33. Larger guessing led to a larger length in the CI estimates. The marginal means of $N = 100$, 300, and 500 were .133, .076, and .058. Increasing a sample size accompanied shorter CI estimates, which follows an

expected effect of an increase in sample size on reducing uncertainty due to sampling variability. The marginal means of Length for C_α were .112 (when C_α = .8) and .067 (when C_α = .9). A higher coefficient alpha made a shorter CI estimate. The interaction effects between guessing and sample size and between C_α and sample size are presented in Figure 2.

Figure 2 shows that differences in Length for the three levels of guessing became larger as the sample size decreased from 500 to 100. Also, the difference between two levels of C_α became larger as the sample size decreased from 500 to 100.

Shape (SH). The last three columns in Table 2 contain the ANOVA summary for Shape. CI showed a very large effect (η^2 = .766), explaining 77% of the total variation. In part, this is expected because CI_V is symmetric and the value of the Shape measure for CI_V is always 1 while the other CI procedures do not enforce symmetry. Sample size had a medium effect (η^2 = .157) and interaction between the CI and sample size showed a small effect (η^2 = .075). Guessing and its interaction with other factors did not have any salient effects.

The marginal means of CI_{B1}, CI_{B2}, CI_F, and CI_V were .773, .874, .904, and 1, respectively. A Shape value less than one means the distance between the upper bound and the point estimate is shorter than that between the point estimate and the lower bound. The further the Shape value is from 1, the more asymmetric a CI estimate is. The most asymmetric CI procedure on average was CI_{B1}, while CI_F was the closest to symmetry. The marginal means of N = 100, 300, and 500 were .837, .902, and .923, respectively. As sample size increases, CI estimates tended to approach symmetry. Interaction between CI and sample size is displayed in Figure 3.

When the sample size was 100, differences in Shape among the CI procedures were larger than when sample size was 300 or 500. Also, of the four CI procedures, CI_{B1} was the most sensitive to the sample size variation in its Shape value. That is, the extent of asymmetry in CI_{B1} was more severe than the other CI procedures when the sample size decreases.

When Data Follow the Essentially Tau-Equivalent Condition

Coverage Rate (CR). Table 3 summarizes the results for CR, Length, and Shape. (For ANOVA results, only the p value and degrees of freedoms are presented in the table.) The sample size (N) and the CI had medium effect sizes (η^2 = .103 and η^2 = .164, respectively). (These two effects were also detected as salient in when data were dichotomous with guessing.) The marginal means for N = 100, 300, and 500 were .943, .947, and .949, which are not big differences but with a pattern of being closer to .05 as N increases. The four CI procedures' marginal means were .942, .944, .950, and .949 for CI_{B1}, CI_{B2}, CI_F, and CI_V, respectively. The largest difference from .95 was found in CI_{B1} while, on average, CI_F accurately produced .95 coverage. In the condition where data follow the tau-equivalent condition, the analytical CI procedures produced better results than the bootstrap CIs, though differences are not large. (Note that when data are binary with guessing, the margin means for CIs were .944,

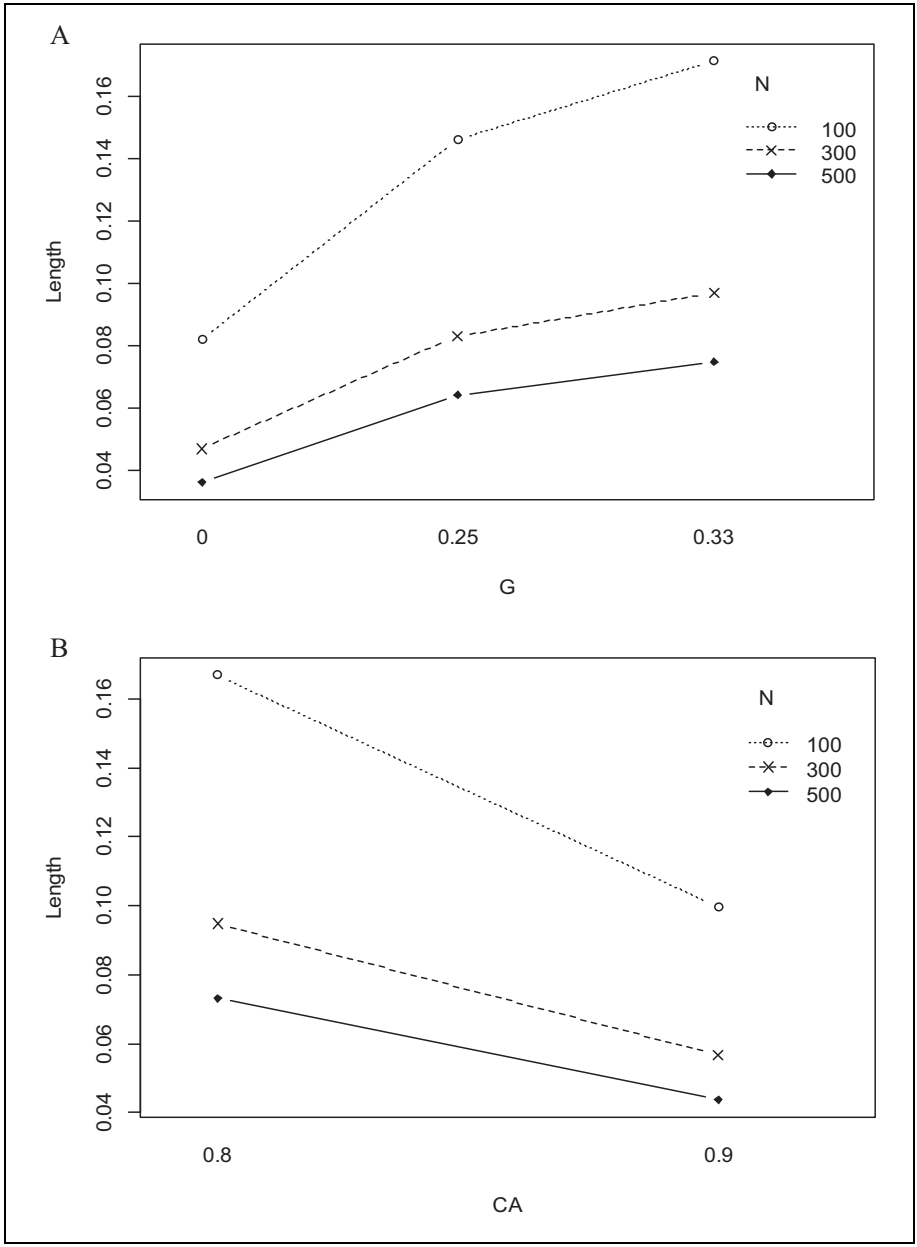

Figure 2. Interactions effects on Length when item responses are dichotomous under the influence of guessing. (A) Interaction between guessing and sample size. (B) Interaction between C_α and sample size.

Note. "CA," and "N" represent C_α and sample size, respectively.

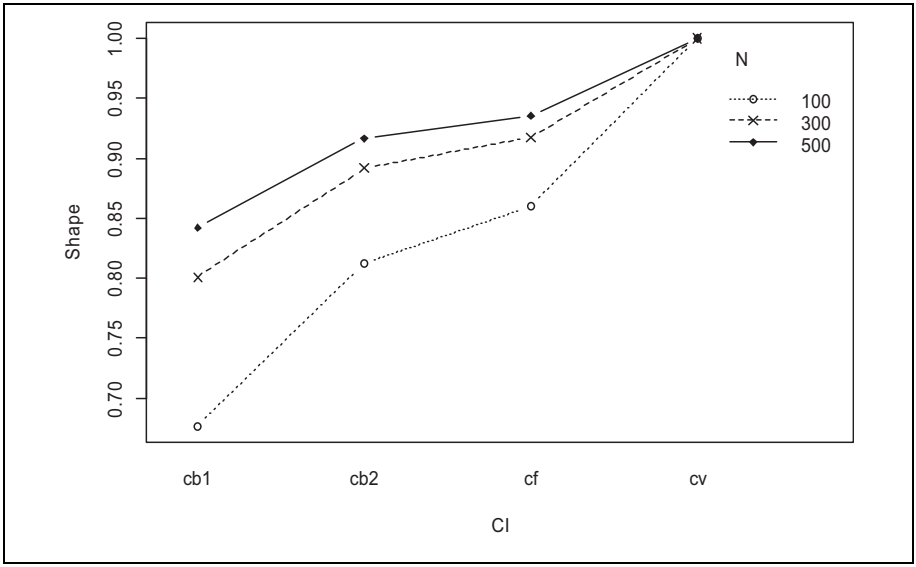

Figure 3. Interaction effect between confidence interval (CI) and sample size on Shape when item responses are dichotomous under the influence of guessing.
Note. "cb1," "cb2," "cf," "cv," and "N" represent CI_{B1}, CI_{B2}, CI_{F}, CI_{V}, and sample size, respectively.

Table 3. Effect Sizes and ANOVA Results for CR, Length, and Shape When Item Responses Follow the Tau-Equivalent Condition.

		CR		LE		SH	
Source	*df*	*p* Value	η^2	*p* Value	η^2	*p* Value	η^2
Between subjects	119						
C_α	1	.747	.001	<.01	.189**	.563	<.001
P	1	.747	.001	<.01	.024*	<.01	<.001
N	2	<.01	.103**	<.01	.678***	<.01	.148**
C_α:*P*	1	.324	.005	<.01	.020*	.628	<.001
C_α:*N*	2	.751	.003	<.01	.002	.422	<.001
P:*N*	2	.723	.003	<.01	.044*	<.01	<.001
C_α:*P*:*N*	2	.494	.007	<.01	.042*	.336	<.001
Residuals	108						
Within subjects	360						
CI	3	<.01	.164**	<.01	<.001	<.01	.777 ***
CI:C_α	3	.439	.001	<.01	<.001	.293	<.001
CI:*P*	3	.653	.000	<.01	<.001	<.01	<.001
CI:*N*	6	<.01	.059*	<.01	<.001	<.01	.075*
CI:C_α:*P*	3	.624	.000	<.01	<.001	.7147	<.001
CI:C_α:*N*	6	.864	.001	<.01	<.001	.4065	<.001
CI:*P*:*N*	6	.181	.002	<.01	<.001	<.01	<.001
CI:C_α:*P*:*N*	6	.977	.000	<.01	<.001	.0467	<.001
Residuals	324						
Total	479						

Note. ANOVA = analysis of variance; LE = Length; SH = Shape; *df* = degrees of freedom; CR = coverage rate. "*," "**," and "***" represent small, medium, and large effect size, respectively.

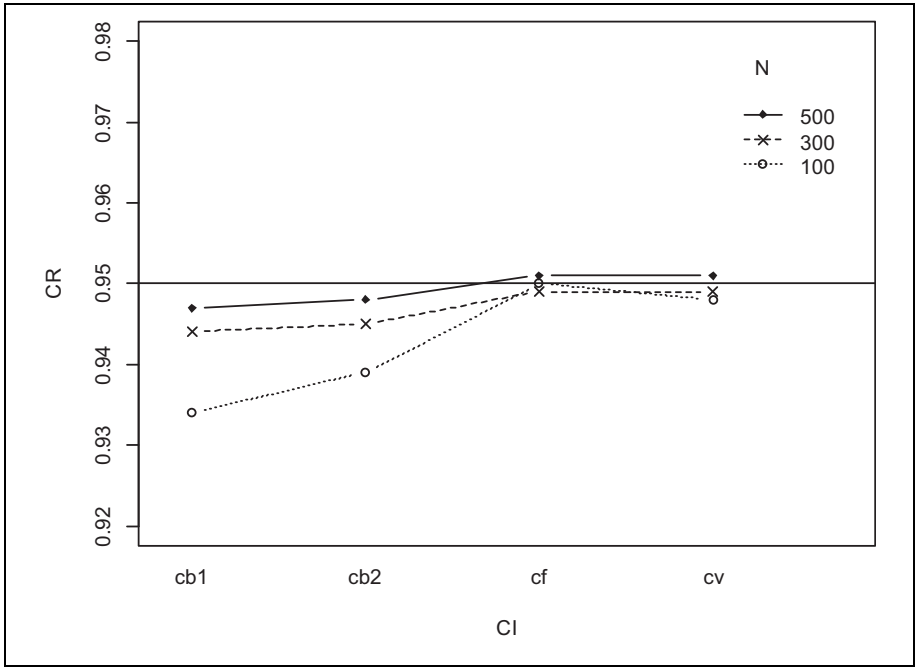

Figure 4. Interaction between confidence interval (CI) and *N* on coverage rate (CR) when coefficient alpha is reliability when item responses follow the tau-equivalent condition.
Note. "cb1," "cb2," "cf," "cv" and "N" represent CI_{B1}, CI_{B2}, CI_F, CI_V, and sample size.

.947, .966, and .963 for CI_{B1}, CI_{B2}, CI_F, and CI_V, respectively.) CI interacted with the sample sizes and produced a small effect (η^2 = .059). CI differences depended on the sample size condition. (The CI:*N* interaction was not salient when data were dichotomous with the effect of guessing.)

Figure 4 shows the interaction plot between CI and *N*. When the sample size was small, more differences from .95 were found in the bootstrap CIs than the two analytical CIs. A larger sample size made these differences decrease.

Length (LE). CI did not show salient differences. There was no salient interaction between CI and other variables, as was the case when data were dichotomous with guessing. Sample size had a very large effect (η^2 = .678 and the marginal means for *N* = 100, 300, and 500 were .101, .049, and .038), decreasing Length as *N* increases, which is the usual behavior of increasing sample size effect on the decrease of uncertainty related to an estimate. C_α had a medium effect (η^2 = .189 and the marginal means for C_α = .8 and C_α = .9 were .077 and .048, respectively). A high C_α was associated with a shorter Length. (The salient effects of C_α and *N* were also observed when data were dichotomous with guessing.) The four effects (*P*, C_α:*P, P:N*, and C_α:*P:N*) showed small effects, which did not have any salient effect when data were

dichotomous with guessing. A longer test length made a shorter length (.068 for P = 21 and .058 for P = 41). The P:N interaction indicated that the effect of a longer test on Length (i.e., decreasing Length) became larger when N was small than when N is large. The C_α:P interaction showed that the impact of a higher C_α on LE (i.e., decreasing LE) became larger as the test length increased.

Shape (SH). Exactly the same patterns were found as in the case where data were dichotomous with guessing. Sample size (N), CI, and CI's interaction with sample size (CI:N) showed medium, large, and small effects, respectively. In the condition where data were dichotomous with guessing, these three were also detected as salient effects and their effect sizes are similar to those when data follow the tau-equivalent condition. Increasing the sample size made the Shape values closer to one, that is, a larger sample made CIs more symmetric. (The marginal means of Shape were .836, .903, and .925 for N = 100, 300, and 500, respectively.) The CI procedures were different in their degree of symmetry (again noting that CI_V is symmetric). The marginal means of Shape were .756, .801, .904, and 1 for CI_{B1}, CI_{B2}, CI_F, and CI_V, which indicates that CI_{B1} produced the most asymmetric estimates. The pattern shown in the CI:N interaction effect is almost the same as that shown in Figure 3. The reduction of the degree of asymmetry in C_{B1} was most dramatic compared with other CI procedures as sample size increased.

Summary and Discussion

Given the lack of investigation of the impact of guessing and its interactions with other variables on interval estimators for coefficient alpha, the primary focus of this study was on the examination of the behavior of the four CI procedures for coefficient alpha with dichotomous item response data under the influence of guessing which exists in a multiple-choice test (e.g., in educational testing). In addition, the interval estimates of coefficient alpha when data follow the tau-equivalent condition were investigated as a supplement to the above case of dichotomous item response data with guessing. The common salient effects found in both binary data with guessing and tau-equivalent data cases were sample size and the choice of a CI procedure on CR, the value of population coefficient alpha and the number of items in a test on Length, and sample size, CI procedures, and the interaction of CI and sample size on Shape. For data following the tau-equivalent condition, the two analytical CI procedures (CI_F and CI_V) outperformed the two bootstrap CI procedures on average by a small margin, though the differences in the four procedures diminished as sample size increased. This is in contrast to when data are dichotomous with guessing, where the two bootstrap methods (CI_{B1} and CI_{B2}) performed better on average than the two analytical CI procedures.

Because the central investigation of interest was when item responses are dichotomous with examinee guessing, a summary of the results from the case of binary data with guessing and discussion of them are given below with more details.

With the criterion of CR as a statistical performance index, the four CI procedures showed differences. The two bootstrap CIs, CI_{B1} and CI_{B2} showed slight downward bias for CR while CI_F and CI_V showed a bit of upward bias for CR under .05 significance level. The two bootstrap CIs performed better than CI_F and CI_V, and the best performer on average was CI_{B2}. Guessing and its interaction with other factors (particularly with CI) did not make salient impacts on CR. Instead, a salient interaction of CI with the different levels of coefficient alpha (C_α) was found: The performance (CR) of CI_F and CI_V changed as C_α changed from .8 to .9 while the two bootstrap CIs were minimally affected by C_α. CI_F and CI_V performed better at C_α = .8 than at C_α = .9.

When Length was examined, not surprisingly, increasing sample size had a large effect and led to shorter CI estimates. Guessing also showed a large effect, particularly with a small sample size, although guessing's interactions with other factors did not show any salient effects. Increasing guessing resulted in increased Length, which indicates that guessing causes more uncertainty in item responses in addition to uncertainty due to sample size. When nonzero guessing (g = .25 or .33) was involved with a small sample size (e.g., N = 100), Length of CI estimates became much larger than when there was no guessing. The level of coefficient alpha had a medium effect size on Length: A higher C_α (= .9) made a shorter CI estimate than a lower C_α (= .8). Also, C_α and sample size showed an interaction effect: Differences in Length between the two C_α levels became larger as sample size decreased.

The four CI procedures had large differences regarding Shape: CI_{B1}, CI_{B2}, CI_F, and CI_V in the order of asymmetry from the most asymmetric to symmetric. (Note again that CI_V imposes symmetry around the point estimate of coefficient alpha while the others do not.) Sample size made a difference on Shape as a main effect and an interaction effect with CI: Increasing sample size decreased asymmetry in CI_{B1}, CI_{B2}, and CI_F estimates, and CI_{B1} showed the largest difference in the degree of asymmetry as sample size increased. There was no salient effect of guessing and its interaction on Shape.

Unlike the well-known detrimental effect of guessing on the point estimate of coefficient alpha, the effects of guessing and its interactions with other factors used for binary data with guessing in this study were not salient on CR and Shape. In a sense, this may be viewed as positive because one does not have to worry about a large deviation from what is expected in terms of the degree of asymmetry and the statistical performance for the CI estimates of coefficient alpha in spite of test takers' guessing in their item responses. However, the analysis on Length clearly indicates that the CI estimates were affected largely by the extent of guessing, particularly, even more greatly when guessing is combined with a relatively small sample size. Guessing creates extra uncertainty in item responses, which is reflected in an increased length of the CI estimates.

Within the boundary of the current simulation conditions, CI_{B2} showed the least amount of bias in terms of CR for binary data with guessing. CI_F showed the largest (positive) bias regarding CR. (Again, the marginal means of CR were 0.944, 0.947,

0.966, and 0.963 for CI_{B1}, CI_{B2}, CI_F, and CI_V.) Given the violation of normality of item responses and compound symmetry in the item covariance matrix in the study conditions, the biases in CR shown in CI_F and CI_V might be treated as not-large or even nearly robust though they were larger than the biases shown by the two bootstrap CI procedures (see also Yuan & Bentler, 2002, on robustness of the asymptotic distribution of coefficient alpha based on normality). The slightly smaller bias compared with CI_V than CI_F may be because of the more stringent assumption used by CI_F on the item covariance matrix (i.e., compound symmetry while no restriction is required in CI_V).

Although this study used several variables which are important in actual testing situations for a multiple-choice test, limitations exist that should be addressed. First, the present study used the 3PLM to generate item response data that have guessing. Some advantages of using the 3PLM in this simulation study were mentioned earlier, but the 3PLM is not the sole representation of modeling item responses and guessing, nor is it the only mechanism of how guessing is introduced in item responses. Other ways to produce item responses with guessing, such as a mixture IRT model (Mislevy & Verhelst, 1990) for instance, could be considered to generate item responses with guessing in a future study. Second, data generation and the two bootstrap CI procedures were based on what Lord (1955) called Type 1 sampling, which considers random sampling of test takers. In a testing situation, the sources of random variations may be test takers, items, or both test takers and items, each of which was described as Type 1, Type 2, and Type 12 by Lord (1955). Data in this study were generated under Type 1 sampling. (Some may argue that Type 12 fits real testing situations better, but others may not completely agree. For example, in educational achievement testing, specifications of the number of items, their content domains, and even the number of items that belong to [at least approximately] easy, medium, and hard items are elaborated by a test blue print, which cannot be equated to a simple random selection of an item set from a population of items.) The CIs in C_{B1} and CI_{B2} in this study were also constructed with the Type 1 sampling framework. Feldt (1965) derived CI_F under Type 12 sampling, but indicated that CI_F works well even in the Type 1 sampling scenario. Pandey and Hubert (1975) compared CI_F with 10 Jackknife variants suitable for Type 12 sampling and concluded that the ''row elimination'' technique, which is essentially consistent with Type 1 sampling, produced more adequate results. Thus, the use of the two bootstrap CI procedures under Type 1 sampling in this study might be thought of as robust regardless of Type 1 or Type 12 sampling schemes. However, Barchard and Hakstian (1997) showed that results from Type 1 sampling cannot be generalized to Type 12 sampling when essential parallelism is violated. Further studies are requested with data generation and bootstrap procedures suitable for Type 12 sampling. Third, the levels of C_α were restricted to .8 and .9, which were often found in multiple-choice tests in educational achievement testing. Different levels of C_α may be used in a future study. The bootstrap procedure becomes inconsistent when a parameter of interest is on the boundary of the parameter space (Andrews, 2000). It is not clear whether the

bootstrap CIs including CI_F and CI_V (which imposes symmetry) behave poorly as C_α approaches one. In relation to the choice of the level of C_α, the present study used different sets of item slope values to manipulate C_α for binary data with guessing. To vary the levels of C_α, other methods may be used. For example, mismatches between test difficulty and person ability distributions (e.g., item difficulties generated from $N(-0.5, 1)$ and abilities generated from $N(0, 1)$) can bring about the effect of lowering C_α. The current study assumed that the overall test difficulty matches the ability distribution, which is often the case in many educational assessment testing programs. Fourth, item parameters were generated from truncated normal distributions and the person ability distribution was assumed to be a normal distribution. These do not present all possible scenarios in practice. Different distributions for item parameters and investigating the impact of nonnormal ability distributions under the influence of guessing may be considered in future studies. Fifth, C_α in this study is a biased estimate of the true population reliability for binary data with guessing. The data generation here violates the essentially tau-equivalent condition completely. Some empirical results show that the difference between C_α and the true population reliability is small when data follow the unidimensional 3PLM with θ being from a normal distribution (see, Paek, 2015, for this). It may be, therefore, conjectured that the CRs for the population coefficient alpha reported here do not deviate largely from those for the true population reliability. However, future studies are required to draw solid conclusions on the behavior of the CI estimates regarding both the population coefficient alpha and the true population reliability. Last, more extreme test length (e.g., less than 20, though not common in educational testing because of the test score reliability issue) and sample sizes, different significance levels (e.g., .01 and .10) for the CI estimate calculation, and other CI procedures such as asymptotically distribution-free CI (Maydeu-Olivares et al., 2007; Yuan et al., 2003) and different bootstrap CIs having the second-order properties (e.g., bootstrap-*t* and ABC methods), which were not included in this study, may be considered in future investigations.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, *68*, 399-405.

Barchard, K. A., & Hakstian, A. R. (1997). The effects of sampling model on inference with coefficient alpha. *Educational and Psychological Measurement*, *57*, 893-905.

Bay, K. S. (1973). The effect of non-normality on the sampling distribution and standard error of reliability coefficient estimates under an analysis of variance model. *British Journal of Mathematical and Statistical Psychology*, *26*, 45-57.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

California Department of Education. (2014). *California Standard Tests Technical Report Spring 2013 Administration*. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/cst13techrpt.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

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

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge, England: Cambridge University Press.

Duhachek, A., Coughlan, A. T., & Iacobucci, D. (2005). Results on the standard error of the coefficient alpha index of reliability. *Marketing Science*, *24*, 294-301.

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

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall.

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

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

Integrated Louisiana Educational Assessment Program. (2014). *Integrated Louisiana Educational Assessment Program 2014 Technical Report*. Retrieved from https://www.louisianabelieves.com/docs/default-source/assessment/ileap-technical-summary.pdf?sfvrsn=6

Kano, Y., & Azuma, Y. (2003). Use of SEM programs to precisely measure scale reliability. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 141-148). Tokyo, Japan: Springer-Verlag.

Kristof, W. (1963). The statistical theory of stepped-up reliability when a test has been divided into several equivalent parts. *Psychometrika*, *28*, 221-238.

Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, *20*, 1-22.

MacCann, R. G. (2004). Reliability as a function of the number of item options derived from the ''knowledge or random guessing'' model. *Psychometrika*, *69*, 147-157.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*, 157-176.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (2nd ed.). London, England: Chapman & Hall.

McDonald, R. P. (1999). *Test theory—A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Michigan Department of Education. (2013). *Michigan English Language Proficiency Assessment Final Technical Report 2013 Administration: Kindergarten Through Grade 12*. Retrieved from http://www.michigan.gov/documents/mde/2012_ELPA_Tech_Report_final_4-23-2013_421032_7.pdf

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195-215.

Nebraska State Accountability (NeSa). (2014). *Nebraska State Accountability Reading, Mathematics, and Science Technical Report*. Retrieved from http://www.education.ne.gov/assessment/pdfs/Final_2014_NeSA_Technical_Report.pdf

New York State Education Department. (2013). *New York State Alternate Assessment Technical Report 2012-13*. Retrieved from http://www.p12.nysed.gov/assessment/reports/nysaa/nysaa-tr-13w.pdf

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1-13.

Padilla, M. A., & Divers, J. (2013). Coefficient omega bootstrap confidence intervals: Nonnormal distribution. *Educational and Psychological Measurement*, *73*, 956-972.

Padilla, M. A., Divers, J., & Newton, M. (2012). Coefficient alpha bootstrap confidence interval under nonnormality. *Applied Psychological Measurement*, *36*, 331-348.

Paek, I. (2015). An investigation of the impact of guessing on coefficient α and reliability. *Applied Psychological Measurement*, *39*, 264–277.

Pandey, T. N., & Hubert, L. (1975). An empirical comparison of several interval estimation procedures for coefficient alpha. *Psychometrika*, *40*, 169-181.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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

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

Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. H. Holyle (Ed.), *Handbook of structural equation modeling* (pp. 472-492). New York, NY: Guilford Press.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Romano, J. L., Kromrey, J. D., & Hibbard, S. T. (2010). A Monte Carlo study of eight confidence interval methods for coefficient alpha. *Educational and Psychological Measurement*, *70*, 376-393.

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*, 271-280.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement*, *13*, 233-243.

Woodruff, D. J., & Feldt, L. S. (1986). Tests for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika*, *51*, 393-413.

Yuan, K.-H., & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika*, *67*, 251-259.

Yuan, K.-H., Guarnaccia, C. A., & Hayslip, B. (2003). A study of the distribution of sample coefficient alpha with the Hopkins symptom checklist: Bootstrap versus asymptotics. *Educational and Psychological Measurement*, *63*, 5-23.

Zimmerman, D. W. (1975). Proability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, *40*, 395-412.

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, *27*, 357-371.