# University of Alberta

# Evaluating the Performance of the Uncorrected and Corrected Reliability Alpha for Range Restriction and the Confidence Intervals in a Single and Meta-Analytic Study

by

Johnson Ching Hong Li

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Measurement, Evaluation, and Cognition

Department of Educational Psychology

Spring 2013
Edmonton, Alberta

# Abstract

Range restriction has long been a methodological problem in educational and psychological research (Hunter & Schmidt, 2004), and this usually leads to a downward-biased estimate of a statistic. Even though much research has examined the performance of Pearson's correlation under range restriction in both single and meta-analytic studies (e.g., Li, Chan, & Cui, 2011a), the assessment of reliability coefficients (e.g., coefficient alpha) under range restriction is relatively limited. Regarding a single study, Fife, Mendoza, and Terry's (2012) have recently examined the performance of the uncorrected and bias-corrected coefficient alpha; as an extension, the performance of the confidence intervals (CIs) and widths also need to be examined. Regarding a meta-analytic study, Rodriguez and Maeda (2006) have proposed a framework for conducting a meta-analysis of coefficient alpha; as an extension, the accuracy of the bias-corrected mean alpha as well as the associated CIs also need to be evaluated.

In light of these unexamined issues, this dissertation sought to evaluate the performance of the uncorrected and bias-corrected alphas—as well as their CI—in both single and meta-analytic study research situations. This provides a comprehensive assessment of reliability under range restriction, thereby providing guidelines about the treatment of biases that come from range restriction. The Monte Carlo results showed that the uncorrected alpha suffered as a function of the selection ratio and the correlation between the test and the selection variable in both single and meta-analytic studies. By contrast, the bias-corrected alpha could adjust for the bias appropriately. Moreover, the bootstrap CIs constructed for the bias-adjusted alpha in both single and meta-analytic studies were generally accurate across different simulation conditions, including sample size, item number, etc. Application of the correction procedure and CI construction in a real study is demonstrated. Based on these results, conclusions, discussions, and recommendations are also presented.

# Acknowledgement

Completing my PhD is definitely one of the most important and challenging tasks in my life. Without the guidance and help from many others, this dissertation would not have been finished. First of all, I would like to express my sincere gratitude to my academic advisor and dissertation supervisor Dr. Ying Cui, for her continuous support for my research, and for her enthusiasm and immense knowledge in the area of educational measurement and psychometrics. With her expertise, she has helped me to improve the quality of writing, execution of the Monte Carlo studies, and presentation of empirical results. Second, I am heartily thankful to my supervisory committee members, Drs. Mark Gierl and Cheryl Poth, for their continuous support and advice on my study. They have coached me how to make the study to be more applicable in practice and more manageable in terms of the research design. Third, I would like to say thank you to my peers in the Centre for Research in Applied Measurement and Evaluation (CRAME). They have been very helpful and supportive.

In addition, I wish to thank my friends and parents for their sincere support over these years. I am heartily thankful to Ms. Virginia Tze, who has helped me to go through many difficult times, and for all the emotional support, entertainment, and caring provided. I also wish to express my gratitude to my parents, who have encouraged me to pursue my PhD degree. Lastly, I offer my blessings to all of those who have supported me in any respect during the completion of my degree.

## Table of Contents

## List of Tables

## List of Figures

# Chapter 1 - Introduction

In educational and psychological research, reliability plays a central role in evaluating the consistency of test scores across replications (Brennan, 2006). It is defined as the correlation between total scores on two independent administrations or two parallel forms of a test (Gulliksen, 1987). In many practical situations, however, test scores across replications are often unavailable. A reliability coefficient is, therefore, often estimated based on scores from a single test administration. Among these single-test reliability coefficients, coefficient alpha (Cronbach, 1951; Guttman, 1945) is regarded as one of the most frequently used and reported indices in the education and psychology literature (Bonett, 2010).

## Coefficient Alpha

Coefficient alpha characterizes the reliability of measurement based on the average correlation of *k* test parts. According to Bonett (2010), the *k* test parts can represent *k* raters, *k* occasions or testing situations, *k* alternative forms, or *k* questionnaire/test items. When the *k* test parts refer to the *k* questionnaire/test items, coefficient alpha is a measure of internal consistency reliability of the scores in a test. Note that coefficient alpha is regarded as one of the lower bound estimates of the reliability; these estimates are often used to evaluate the minimum reliability level of a test because neither the true score nor measurement error of each person is known, and hence statistical algorithms can only be derived to approximate the lowest possible reliability value of a test. Some authors (e.g., Sijtsma, 2009) have suggested that researchers should also provide other lower bound estimates [e.g., McDonald's (1999) omega] in addition to coefficient alpha. Details will be discussed in Chapter 2.

**Range Restriction**

Despite its common use in evaluating reliability, coefficient alpha is a downward-biased estimate of its true parameter value when a sample is subject to range restriction. Selection of participants is a common cause of range restriction (Hunter, Schmidt, & Le, 2006). Selection is not problematic itself, but it becomes problematic when one estimates the alpha based on a restricted sample, and seeks to generalize the result to the unrestricted population. Given that coefficient alpha depends on the variance of the restricted scores and selection often causes a smaller variance, the observed alpha is often downward-biased.

In light of the adverse effect of range restriction, many researchers (e.g., Chan & Chan, 2004; Hunter & Schmidt, 2004; Li, Chan, & Cui, 2011a) suggested that one can adjust for a biased sample estimate of a statistic using the standard deviation obtained from an unrestricted sample (e.g., study report, technical manual). These studies, however, focused primarily on the correction procedures for Pearson's correlation. To my knowledge, there are two empirical studies—Sackett, Laczo, and Arvey (2002) and Fife, Mendoza and Terry (2010)—which have examined the issue of reliability under range restriction. Both studies suggested that one should use correction procedures to adjust for bias whenever the correction factors (e.g., the standard deviation of scores in an unrestricted sample) are available.

## Confidence Intervals

In addition to the point estimate of the alpha, the associated confidence interval (CI) has important implications for evaluating the accuracy of the sample estimate of coefficient alpha, and for comparing different tests, scoring rubrics, or training procedures for raters or observers (Haertel, 2006). Indeed, many publication manuals have stated that the CI surrounding a statistical estimate should be provided in research studies. For example, Section 2.07 of the manual of the American Psychological Association (2010, p. 34) suggested that “because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy.” Thus this dissertation also examines the performance of the CIs surrounding the uncorrected and corrected alpha.

## Meta-analysis

As an extension, evaluating reliability based on a single study may not be sufficient, as argued by Vacha-Haase (1998) and Rodriguez and Maeda (2006). Vacha-Haase suggested that “[b]ecause tests are not reliable, it is important to explore score reliability in virtually all studies” (p. 6). Researchers are therefore encouraged to evaluate reliability based on multiple studies, and this can be achieved through a quantitative technique called meta-analysis. Meta-analysis is a statistical procedure that synthesizes the quantitative findings (e.g., correlation, coefficient alpha) reported in each single study conducted by independent researchers (Hunter & Schmidt, 2004). Meta-analysis produces a mean estimate of these measures, thereby providing a summary of these measures in a research domain (also known as the “typical” reliability). In addition, CIs are often generated in a meta-analytic study, so that researchers can evaluate the associated precision and sampling error of the mean reliability. Rodriguez and Maeda (2006) have recently proposed a framework specific to the meta-analysis of coefficient alpha. Since then, a large number of studies have examined the mean alpha reliability level of educational and psychological scales, including Vassar and Bradley (2010), Vassar and Crosby (2008), and Warne (2011).

As in a single study, however, a meta-analysis of coefficient alpha may be biased when the alpha reported in each single study is subject to range restriction. It appears to be a common practice, especially in the personnel psychology literature, to adjust for the correlations in each single study before they are used to compute the mean correlation in meta-analysis (Hunter & Schmidt, 2004). However, a consensus about how to conduct a meta-analysis of coefficient alpha when it is subject to range restriction has not yet been reached, given different concerns and suggestions (e.g., Blixt & Shama, 1986; Bonett, 2010; Feldt & Qualls, 1999; Hunter & Schmidt, 2004; Lord, 1984; Rodriguez & Maeda, 2006), which will be discussed in Chapter 2.

**Summary**

In sum, the primary purpose of this dissertation is to evaluate the performance of the uncorrected and bias-corrected coefficient alpha—as well as their CIs—in single and meta-analytic study research scenarios, respectively. This purpose can be separated into four goals below.

**Goal 1 (Monte Carlo Study 1): To evaluate the accuracy of the uncorrected and corrected coefficient alphas in a single study.** In the literature, much attention has been devoted to the correction procedures for Pearson's correlation under range restriction. Given that the effect of range restriction should be comparable to other statistics (e.g., reliability), this dissertation seeks to extend the framework of range restriction to reliability coefficients. Coefficient alpha is selected in this dissertation because it is one of the most frequently reported reliability coefficients, and hence its application should be more relevant to applied users and researchers, especially for meta-analytic studies. Two conventional correction procedures for coefficient alpha can be found in Gulliksen (1987) and Schmidt, Hunter, and Urry (1976), but their performances may need further examination. Thus, the first goal of this dissertation is to conduct a Monte Carlo study—a widely used strategy for evaluating the robustness of a data-analytic method across replications—to evaluate the performance of the uncorrected and two bias-corrected coefficient alphas in a single study.

**Goal 2 (Monte Carlo Study 1): To examine the performance of the non-parametric bootstrap confidence intervals in a single study.** In addition to the accuracy of the point estimate of the uncorrected and corrected alphas, the construction of the confidence intervals (CIs) has important implications for evaluating the sampling error and making statistical inferences. Thus, this dissertation evaluates the performance of the non-parametric bootstrap CIs surrounding the uncorrected and corrected alphas. The non-parametric bootstrap procedure is selected because it has recently been found to construct an adequate CI for the correlation corrected for range restriction (e.g., Chan & Chan, 2004; Li et al., 2011a; Mendoza, Hart, & Powell, 1991). By drawing (or bootstrapping) successive samples with replacement from an observed dataset, one can approximate the sampling behavior of a test statistic, so as to make the statistical inferences about the estimate. Hence the second goal of this dissertation is to evaluate the performance of the bootstrap CIs. The results of this Monte Carlo study provide empirical evidence about the bias of the range-restricted alpha and the accuracy of the two bias-corrected alphas across different data conditions, which applied researchers and users may encounter in practice.

**Goal 3 (Monte Carlo Study 2): To evaluate the accuracy of the uncorrected and corrected coefficient alphas in a meta-analytic study.** As noted by Bonett (2010), the CI constructed for coefficient alpha in a single study may not be sufficient (i.e., it is usually too wide), and this may lead to an inaccurate evaluation of the sampling error surrounding the alpha. Rather, evaluating the CI based on a meta-analytic study of coefficient alpha is an alternative. However, the alphas reported in each single study may also be affected by range restriction. Some authors (e.g., Rodriguez & Maeda, 2006; Botella, Suero, & Gambara, 2010) argue that researchers cannot evaluate the mean alpha and its CI because of the range-restriction bias in each alpha, while others (e.g., Hunter & Schmidt, 2004) recommend that researchers should adjust for the alpha to an unrestricted SD from a large sample (e.g., technical manual) before evaluating the mean alpha and the associated CI.

In light of these different views, the third goal of this dissertation is to evaluate the performance of the uncorrected and corrected mean alphas through the use of a second Monte Carlo study. Results of this study provide empirical evidence about the accuracy of the uncorrected and corrected mean alphas in a meta-analytic study. This study also provides empirical results so that researchers can consider the merits and drawbacks that come from the uncorrected and corrected mean alphas, and hence sheds light on the most appropriate practice of conducting a meta-analysis of coefficient alpha.

**Goal 4 (Monte Carlo Study 2): To examine the performance of the non-parametric bootstrap confidence intervals in a meta-analytic study.** As in a single study, the performance of the CIs surrounding the mean alphas are also important for making statistical inferences about the typical alpha level in a research domain. Rodriguez and Maeda (2006) suggested that one can use the conventional parametric method to construct the CI for the mean alpha in a meta-analytic study. Recently, Li, Cui, and Chan (in press) found that the non-parametric bootstrap CI built for the mean correlation corrected for range restriction outperformed the conventional CI. Such an improvement is expected to be found for the case of the mean corrected alpha, but no study to date has examined this potential improvement. Hence the fourth goal of this dissertation is to examine the performances of the non-parametric bootstrap CIs surrounding the mean uncorrected and corrected coefficient alphas in the second Monte Carlo study.

This dissertation is organized into five chapters. Chapter 2 is the literature review section, which discusses the background related to reliability, range restriction, correction procedures, and meta-analysis. Chapter 3 is the method section, which describes the design of the two Monte Carlo studies. Chapter 4 presents the results of the two studies, in order to provide empirical evidence for the corrected and uncorrected alpha in both single and meta-analytic studies. In Chapter 5, application of the correction procedures in a single and meta-analytic study of coefficient alpha in reference to the Spence's Children Anxiety Scale (SCAS; Spence, 1997) is discussed. Based on these findings, discussion, conclusions and recommendations are presented in Chapter 6.

# Chapter 2 – Literature Review

The issue of reliability in psychometric testing can be traced back to the early 20th century when researchers, including Spearman, Guttman and Cronbach, sought to examine the consistency of examinee test scores. Reliability is defined as the correlation between total scores on two independent administrations or two parallel forms of a test (Gulliksen, 1987). In practice, however, test scores across replications are rarely available. A reliability coefficient is, therefore, often estimated based on scores from a single test administration. Among different single-test reliability coefficients, coefficient alpha (or α; Cronbach, 1951; Guttman, 1945) is regarded as one of the most frequently reported and used indices in the education and psychology literature. The next section will review different reliability coefficients.

## Background of Reliability Coefficients

**Spearman's test-retest reliability coefficient ($\rho_S$).** Spearman (1904) was the first one who sought to evaluate the reliability of total scores on two independent administrations or two parallel forms of a test. He suggested that one can measure the same persons with two alternative or parallel forms of a test, so that the correlation of their total scores between the two forms represents the reliability level. The reliability of a test $Y$, denoted by $\rho_{YY}$, can be measured by

$$\rho_{YY} = \rho_S = r_{jq}, \tag{2.1}$$

where $r_{jq}$ is the Pearson correlation *r* between the total scores in the test forms *j* and $q$. The two forms are comparable in terms of the content domains, variances and means, as assumed in classical test theory (CTT).

**Spearman-Brown split-half reliability coefficient ($\rho_{SB}$).** Although Spearman (1904) developed a procedure to measure reliability, it is generally impractical for one to construct two strictly parallel or alternative forms of tests. Even if one can do so, the true score of person $i$, $T_i$ (e.g., learning motivation), is expected to vary after the first administration; this affects the consistency of the scores across replications. To deal with this problem, Spearman (1910) and Brown (1910) proposed to measure reliability based on a single-test administration. Specifically, one can split the items in half to mimic a test-retest scenario. The correlation between the two halves represents the reliability of the scores across two replications. However, given that this correlation is based on two-half forms, which consist of only half the items of the original test, it is a downward-biased estimate of the true test-retest correlation and needs to be adjusted. Spearman and Brown proved that the reliability coefficient can be estimated based on the correlation between the total scores of the two halves, $r_{oe}$ ($o$ indicates the odd items and $e$ represents the even items), by

$$\rho_{YY} = \rho_{SB} = 2r_{oe}/(1 + r_{oe}). \tag{2.2}$$

As in Equation 2.1 the assumption that the two halves are comparable in terms of the content domains, variances and means is required for this equation.

**Rulon-Flanagan-Guttman's split-half reliability coefficient ($\rho_{RFG}$).** Based on Spearman and Brown's work, Rulon (1939), Flanagan (1937), and Guttman (1945), developed a reliability coefficient for the split-half situation, in which the assumption of the equality of the variances for the two halves is relaxed. They proposed that a reliability coefficient, $\rho_{YY}$, can be estimated by

$$\rho_{YY} = \rho_{RFG} = 2[1 - (\sigma_o^2 + \sigma_e^2)/\sigma_Y^2], \tag{2.3}$$

where $\sigma_o^2$ is the variance of the scores of the odd items, $\sigma_e^2$ is the variance of the scores of the even items, and $\sigma_Y^2$ is the variance of the scores in the full test.

**Coefficient alpha ($\rho_\alpha$).** Although the split-half strategy provides a way of evaluating the reliability in a single test administration, there are many ways to split a test into two halves. If a test has *k* items, then we have $k!/2[(k/2)!]^2$ ways of dividing a test into two halves. Cronbach (1951) described that one can split a test into *k* parts, where *k* is the number of test items. By doing so, there is only one way to split a test into *k* parts, thereby removing the ambiguity of how to split a test into two halves. Conceptually, coefficient alpha for *k* parts represents the average correlation of all possible items in the test; this indicates the amount of internal consistency reliability of item scores of a test. Coefficient alpha is by far the most commonly reported and used reliability coefficient (e.g., Bonett, 2010; Rodriguez & Maeda, 2006).

***General framework and assumptions of coefficient alpha***. Suppose we have *k* items (i.e., $j = 1, 2, \cdots, k$) measuring a psychological construct (e.g., learning motivation) for *n* persons (i.e., $i = 1, 2, \cdots, n$). On the basis of the linear model, an observed score for person *i* on item *j* can be expressed as

$$Y_{ij} = \mu + I_j + T_i + e_{ij}, \tag{2.4}$$

where $\mu$ is the mean item score across the entire population of persons and items, $I_j$ is the effect of item $j$, $T_i$ is the (true score) effect of person $i$, and $e_{ij}$ is the error of measurement for person *i* on item *j*. According to Cronbach (1972), five assumptions are required: (a) the $n$ persons are a random sample from the population of persons; (b) the $k$ items are a random sample from the population of items; (c) the true scores, $T_i$, are normally distributed over the population of persons; (d) the errors of measurement, $e_{ij}$, are normally distributed over the entire persons-by-items matrix, and independent of each other and of $T_i$; and (e) the error variances, $\sigma_{e_{ij}}^2$, are assumed to be identical for any infinite subpopulation of persons and items, i.e.,, $\sigma_{e_{ij}}^2 = \sigma_e^2$. Given assumptions (d) and (e), the variance of the observed scores on item $j$, $\sigma_{Y_j}^2$, consists of two components,

$$\sigma_{Y_j}^2 = \sigma_T^2 + \sigma_e^2, \tag{2.5}$$

where $\sigma_T^2$ is the (true score) person effect variance and $\sigma_e^2$ is the error variance. The covariance between observed scores on items $j$ and $q$ can be expressed as

$$\sigma_{Y_jY_q} = \sigma_T^2. \tag{2.6}$$

When $\sigma_T^2$ and $\sigma_e^2$ are held constant across items, the observed scores on $k$ items, $Y_j$ ($j = 1,2, \dots, k$), are assumed to have equal variances and covariances. These items are regarded as *essentially parallel* measurements, in which the covariance matrix of the $k$ items is assumed to be compound symmetric (i.e., equal variances and covariances).

Data that meet the essentially parallel condition are important for an accurate estimation of the true reliability. However, violating such a condition is unavoidable in practice (Barchard & Hakstian, 1997a, 1997b). It is generally impractical to construct interchangeable items with equal variances and covariances, and these items are measuring the same underlying construct. A more relaxed condition, known as the essentially tau-equivalent condition, is characterized by unequal variances and equal covariances (Lord & Novick, 1968).

For this condition, error variances $\sigma_e^2$ can vary across items. In addition, the effect of person $i$, namely $T_i$ in Equation (2.4), is not constant across items. Rather than that, for any two items $j$ and $q$, the person effect for item $j$ is equal to the person effect for item $q$ plus a constant $C_{jq}$ across all persons, i.e. $T_{iq} = T_{ij} + C_{jq}$. Given that variances are unaffected by the addition of a constant to a variable, the person effect variance is still constant across items. Therefore, items that are essentially tau-equivalent take the properties of unequal variances (i.e., $\sigma_{Y_j}^2 = \sigma_T^2 + \sigma_{e_j}^2$) but equal covariances (i.e., $\sigma_{Y_jY_q} = \sigma_T^2$). Thus the covariance matrix of the *k* items does not have equal diagonal elements (variances), but the off-diagonal elements (covariances) are equal.

In addition to the essentially tau-equivalent condition, the *k* items can be neither equal variances nor covariances in real situations. Hence a third level of measurement data, namely the congeneric condition (Jöreskog, 1971), is introduced. This level represents the most relaxed condition because the covariance matrix of *k* items has neither equal diagonal elements (variances) nor equal off-diagonal elements (covariances). For any two items $j$ and $q$, the person effect for item $q$ is equal to the person effect for item $j$ multiplied by a positive constant $b_{jq}$ and plus a constant $C_{jq}$ across all persons, namely $T_{iq} = b_{jq}T_{ij} + C_{jq}$. This leads to a covariance matrix with unequal item variances and covariances, i.e., $\sigma_{Y_j}^2 = \sigma_{T_j}^2 + \sigma_{e_j}^2$, and $\sigma_{Y_jY_q} = b_{jq}\sigma_{T_j}^2$.

Given any one of the three aforementioned data conditions, coefficient alpha ($\rho_\alpha$) estimates the reliability of *k* items in a test through

$$\rho_{YY} = \rho_{\alpha} = \frac{k}{k-1}\left[\frac{\sigma_{Y^t}^2 - \sum_{j=1}^{k} \sigma_{Y_j}^2}{\sigma_{Y^t}^2}\right], \tag{2.7}$$

where *k* is the number of items, $\sigma_{Y_j}^2$ is the variance of observed scores on item $j$, $Y_i^t = \sum_{j=1}^{k} y_{ij}$ is the total item score for person *i*, and $\sigma_{Y^t}^2$ is the variance of the total item scores across *n* persons (Cronbach, 1951; Guttman, 1945). Note that coefficient alpha is only one of the lower bound reliability estimates, and it may become less accurate, especially when the congeneric condition is violated (Cui & Li, 2012). Some studies (e.g., Sijtsma, 2009; Revelle & Zinbarg, 2009) suggested that researchers should report other greater lower bound estimates of reliability [e.g., McDonald’s (1999) omega] in additional to coefficient alpha. Hence researchers are encouraged to check the assumption of observed scores before using coefficient alpha. Graham (2006) has discussed the details about how to check these assumptions with the aid of a structural equation modeling (SEM) package.

**Range Restriction**

Range restriction has long been recognized as a common phenomenon by researchers and applied users since Pearson (1903) developed the formula for the correlation $r$. To date, a great number of studies have examined the performance of various bias-correction formulae for the range-restricted correlation (e.g., Alexander, Hanges, & Alliger, 1985; Andre & Hegland, 1998; Chan & Chan, 2004; Darlington, 1998; Gulliksen, 1987; Hunter & Schmidt, 2004; Li et al., 2011a; Sackett & Yang, 2000; Schmidt et al., 1976; Thorndike, 1949). These studies generally regarded range restriction as a phenomenon due to participants’ selection. On the other hand, Hunter and Schmidt (2004) have conceptualized range restriction as a type of study artifacts when researchers conduct a meta-analysis. To offer a better understanding of the relationship between range restriction and reliability, the next section will first present Hunter and Schmidt’s framework of range restriction, and then discuss its relationship with reliability.

**Range restriction as a special type of study artifacts.** According to Hunter and Schmidt (2004), study artifacts are defined as spurious observations or findings that come from investigative procedures made by human beings. They identified a list of study artifacts which may change the value of statistical outcomes or estimates. These artifacts include range restriction, sampling error, reporting or transcriptional error, and attrition artifacts (Hunter & Schmidt, 2004). For example, due to range restriction, the reliability of a study is expected to be systematically lower than the true reliability in the unrestricted population. Regarding sampling error, the observed reliability tends to vary randomly from the true population value because of the sampling error, which in part depends on the number of participants in the study. For reporting or transcriptional error, it is expected that the observed reliability differs from the true population value due to a variety of reporting problems, such as “inaccuracy in coding data, computational errors, errors in reading computer output, typographical errors by secretaries or by printers” (Hunter & Schmidt, 2004, p. 35). Thus, the observed reliability is systematically lower than its true value to the extent that there is a systematic attrition in the number of participants on the study. As noted by Hunter and Schmidt, these artifacts may work in conjunction to induce “quantitative errors so large as to create qualitatively false conclusions” (Hunter and Schmidt, 2004, p. 34), and hence researchers are encouraged to adjust for these artifacts whenever correction factors are available.

According to Hunter et al. (2006), the artifacts described above are expected to affect the observed reliability of a test in a sample. That is, a general attenuation formula for the observed reliability can be expressed as

$$\rho_{YY_o} = A\rho_{YY}, \tag{2.8}$$

where $\rho_{YY_o}$is the observed reliability of test $Y$, $\rho_{YY}$ is the true reliability of test $Y$, and $A$ is the artifact multiplier. For example, suppose we know that the reporting or transcriptional error leads to 10% reduction in reliability. We can adjust the observed reliability for this error by

$$\rho_{YY} = \rho_{YY_o}/A, \tag{2.9}$$

where $A = 90\%$ in this example. However, for artifacts due to sampling error or reporting or transcriptional error, it is generally impossible for one to know the exact value of the artifact multiplier $A$ in practice because these artifacts tend to affects the observed reliability $\rho_{YY_o}$nonsystematically, meaning that we do not know the degree to which or even the direction in which they affects the reliability.

By contrast, the artifact due to range restriction is relatively more manageable, given that the extent of this artifact can be estimated whenever the correction parameters are available and the type of selection process can be inferred (Hunter & Schmidt, 2004). In fact, the importance of reporting and evaluating the reliability adjusted for range restriction has received increasing attention in the literature in education, psychology, and business (e.g., Bonett, 2010; Botella, Suero, & Gambara, 2010; Brennan, 2006; Rodriguez & Maeda, 2006; Sackett et al., 2002). Despite their recent use, the correction procedures for range restriction were mainly developed for Pearson's correlation. Hence the next section will first present four different cases of range restriction for Pearson's correlation, and then discuss the comparable cases of range restriction for coefficient alpha.

## Thorndike's Three Conventional Cases of Range Restriction for Pearson's Correlation

**Case I correction procedure for correlation.** Thorndike (1949) proposed three well-known procedures adjusting for a biased correlation for range restriction. Among them, Thorndike's Case I and II (direct) range restrictions present a fundamental selection process, in which selection occurs based on the rank-ordered *X* scores. For example, a researcher may examine the correlation between variable *X* (e.g., SAT) used in selection of college applicants and variable *Y* (e.g., learning motivation) measured in a group of restricted students. Given that selection occurs based on the rank-ordered *X* scores, and this causes range restriction on *Y*, the measured correlation between *X* and *Y* is often downward-biased. Note that the difference between the two cases lies in the availability of the unrestricted standard deviation (SD). That is, Case I assumes that the unrestricted SD of the criterion *Y* is known whereas Case II assumes that the unrestricted SD of the predictor *X* is known.

Figure 1 shows a conceptual model for the Case I restriction for Pearson's correlation, in which selection occurs based on the rank-ordered *X*, and the unrestricted SD of *Y* (i.e., $S_Y$; usually through a large sample estimate from a technical manual) is available for making a correction. In equation form, the data can be expressed as

$$\left[\begin{array}{cc} \boldsymbol{X_r} & \boldsymbol{Y_r} \\ \hline \boldsymbol{X_m} & \boldsymbol{Y_m} \end{array}\right], \tag{2.10}$$

where $\boldsymbol{X_r}$ and $\boldsymbol{Y_r}$ are the vectors that contain the observed *X* and *Y* scores, respectively, for a restricted sample, and $\boldsymbol{X_m}$ and $\boldsymbol{Y_m}$ are the vectors that contain the observed *X* and *Y* scores, respectively, for an unselected (or missing) sample. Case I assumes that the unselected sample of *Y* (i.e., $\boldsymbol{Y_m}$) is available but the unselected sample of *X* (i.e., $\boldsymbol{X_m}$) is unobservable. Hence the data become

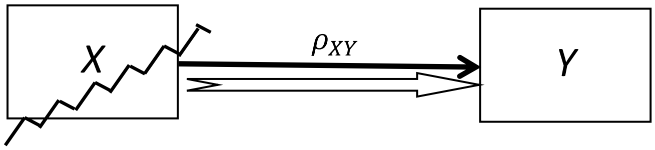
$$\left[\begin{array}{cc} \boldsymbol{X_r} & \boldsymbol{Y_r} \\ \hline \vdots & \boldsymbol{Y_m} \end{array}\right]. \tag{2.11}$$

According to Chan and Chan (2004), when the regression of *Y* on *X* is both linear and homoscedastic (which will be explained in the following section), the Case I correction equation for Pearson's correlation can be written as

$$r_{c1} = r_{XY_u} = \sqrt{1 - u_Y^2\left(1 - r_{XY_r}^2\right)}, \tag{2.12}$$

where $r_{XY_u}$ is the unrestricted correlation between *X* and *Y*, $r_{XY_r}$ is the restricted correlation between *X* and *Y*, and $u_Y = s_Y / S_Y$ is the ratio of the restricted to the unrestricted SD of *Y* (Thorndike, 1949). Note that in practice the unselected *Y* scores (i.e., $\boldsymbol{Y_m}$) are not necessarily available but the unrestricted SD of *Y* (i.e., $S_Y$) is required, as implied in Equation 2.12.

*Figure 1.* A conceptual model for the Case I and II range-restricted correlation.



*Note*: ⇨ indicates the direction of range restriction

→ denotes the relationship between two variables (no causal relationship is assumed).

*X* is the observed score of variable *X*. *Y* is the observed score of variable *Y*. $\rho_{XY}$ is the population correlation between *X* and *Y*.

Although it is statistically sound, the Case I correction procedure is usually not practical in educational and psychological studies because it is difficult for researchers to gather information on the unrestricted SD of *Y*. For example, assume the college students in a study have been selected according to their SAT (*X*) scores. The unselected students may not have a chance to participate in the study and provide their learning motivation (*Y*) scores. Hence an estimate of the unrestricted SD of *Y* is generally not available (Stauffer & Mendoza, 2001).

**Case II correction procedure for correlation.** The correction procedure for the Case II restriction is generally more practical. In fact, it had been commonly used in meta-analytic studies that involved restricted job incumbents before the development of Hunter et al.'s (2006) Case IV correction procedure. Although Case II shares the same conceptual model as in Case I (i.e., selection occurs based on *X*; see Figure 1), Case II assumes that the unrestricted SD of *X* ($S_X$) is known. Thus, the observed data matrix is different from that in Case I, i.e.,

$$\left[\begin{array}{cc} \boldsymbol{X_r} & \boldsymbol{Y_r} \\ \hline \boldsymbol{X_m} & \vdots \end{array}\right]. \tag{2.13}$$

When the regression of *Y* on *X* is both linear and homoscedastic, the Case II correction equation for correlation can be expressed as

$$r_{c2} = r_{XY_u} = \left(\frac{1}{u_X}\right) r_{XY_r} \Big/ \sqrt{\left(\frac{1}{u_X^2} - 1\right) r_{XY_r}^2 + 1}, \tag{2.14}$$

where $r_{XY_u}$ and $r_{XY_r}$ are defined above, and $u_X = s_X/S_X$ is the ratio of the restricted to the unrestricted SD of *X* (Thorndike, 1949).

In many research situations, $S_X$ (i.e., the unrestricted SD of *X*) is easier to be obtained than $S_Y$. For example, a researcher can get access to a university database, which may contain the SAT (*X*) scores of both the selected and unselected students. In another example, a human resources manager in a company may be interested in the relationship between a cognitive-skill test (*X*) and job performance (*Y*). The manager may get access to a database, which contains the *X* scores of both the selected and unsuccessful job applicants.

**Case III correction procedure for correlation.** In addition to Case I and Case II, Thorndike (1949) proposed a third type of range restriction, namely Case III (indirect) range restriction which occurs when the correlation between *X* and *Y* is range-restricted by a third variable *Z*, as shown in Figure 2. A university, for example, may use a variable *Z* (e.g., SAT) to select students for admission. Later, if a researcher is interested in the predictive validity of a cognitive skill test (i.e., *X*) to a learning motivation test (i.e., *Y*), the measured correlation between *X* and *Y* using a restricted sample is often downward-biased. In this case, participants are selected based on the rank-ordered *Z* scores, and this causes range restriction on both *X* and *Y*, thereby reducing the variances of observations in *X* and *Y*. In equation, the data can be expressed as

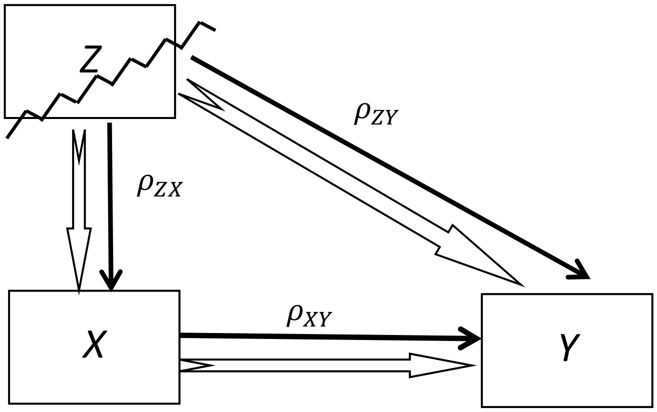
$$\left[\begin{array}{ccc} \boldsymbol{Z_r} & \boldsymbol{X_r} & \boldsymbol{Y_r} \\ \hline \boldsymbol{Z_m} & \vdots & \vdots \end{array}\right], \tag{2.15}$$

where $\boldsymbol{X_r}$ and $\boldsymbol{Y_r}$ are defined as above, $\boldsymbol{Z_r}$ is the restricted sample that contains the restricted *Z*  scores, and $\boldsymbol{Z_m}$ is the unselected or missing sample that contains the unselected *Z* scores. When the regression of *Y* on *X* is both linear and homoscedastic, the correlation corrected for the Case III restriction can be expressed as

$$r_{c3} = r_{XY_u} = \frac{r_{XY_r} + r_{XZ_r} r_{YZ_r} \left( \frac{1}{u_Z^2} - 1 \right)}{\sqrt{\left[ 1 + r_{XZ_r}^2 \left( \frac{1}{u_Z^2} - 1 \right) \right] \left[ 1 + r_{YZ_r}^2 \left( \frac{1}{u_Z^2} - 1 \right) \right]}}\ , \tag{2.16}$$

where $r_{XY_r}$, $r_{XZ_r}$, and $r_{YZ_r}$ are the restricted correlations between *X*-*Y*, *X*-*Z*, and *Y*-*Z*, respectively, and $u_Z = s_Z/S_Z$ is the ratio of the restricted to the unrestricted SD of *Z* (Thorndike, 1949).

*Figure 2.* A conceptual model for the Case III range-restricted correlation.



*Note*: ⇨ indicates the direction of range restriction

→ denotes the relationship between two variables (no causal relationship is assumed).

*X* is the observed score of variable *X*. *Y* is the observed score of variable *Y*. *Z* is the observed score of variable *Z*. $\rho_{XY}$ is the population correlation between $X$ and *Y*. $\rho_{ZX}$ is the population correlation between $Z$ and *X*. $\rho_{ZY}$ is the population correlation between $Z$ and *Y*.

## Hunter et al.'s (2006) Case IV correction procedure for correlation

Since Thorndike's (1949) work, a large number of studies have examined the biasing effects of range restriction on correlation in various disciplines, including education (e.g., Andre & Hegland, 1998; Brennan, 2006; Darlington, 1998; Gulliksen, 1950), psychology (e.g., Alexander, Hanges, & Alliger, 1985; Chan & Chan, 2004; Li et al., 2011a; Li, et al., in press; Mendoza & Mumford, 1987), and business (e.g., Burke, Normand, & Doran, 1989; Hunter & Schmidt, 2004; Hunter, Schmidt, & Le., 2006; Olson & Becker, 1983; Schmidt et al., 1976;

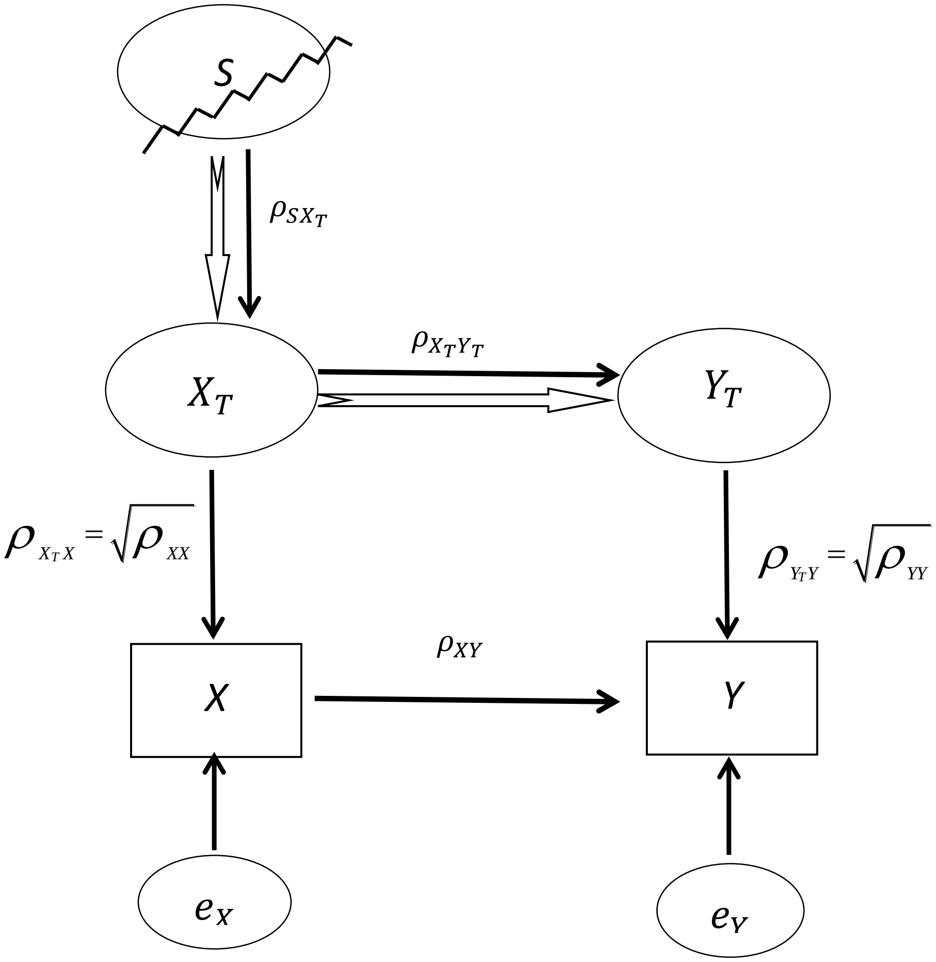
Yang, Sackett, & Nho, 2004). Among them, the recent papers discussed by Hunter, Schmidt, Le, and Oh (e.g., Hunter et al., 2006; Le & Schmidt, 2006; Schmidt, Le, & Oh, 2006) are regarded as the most influential, given that they provide a more general and practical range restriction model for correlation. They called this correction procedure the Case IV formula.

According to Schmidt et al. (2006), although the Case III correction equation (Equation 2.16) seems computationally manageable, it has not been commonly used in the education and psychology literature (Li, et al., 2011a). Schmidt et al. stated that "[i]n practice, this equation can rarely be used" (p. 284) because, first, it is unlikely to be true that participants have been restricted based on a single variable *Z*, as other variables including reference letters, experience, interview performance, etc. are not taken into consideration. Second, the five parameters—$r_{XY_r}$, $r_{XZ_r}$, $r_{YZ_r}$, $s_Z$, and $S_Z$ —are seldom available.

To deal with these limitations, Schmidt et al. (2006) proposed a new range-restriction model, which is known as the Case IV correction procedure. Unlike Case III, Case IV assumes that the third variable is a composite of several unobservable or unquantifiable variables used in selection (named suitability *S*; see Le & Schmidt, 2006). In this framework, a university may use a composite of several unquantifiable variables (e.g., achievement test scores, reference letters and interview performances, denoted as the suitability construct *S*) to recruit applicants, and those with an *S* score below a cutoff are not selected. Later, if a researcher examines the correlation between a cognitive skill test (i.e., *X*) and a learning motivation test (i.e., $Y$), the measured correlation between $X$ and $Y$ is downward-biased.

Figure 3 shows a conceptual model for the Case IV restriction. There are three characteristics. First, it assumes that the selection composite $S$ causes range restriction on the true score of $X$ (i.e., $X_T$), and this causes range restriction on the true score of $Y$ ($Y_T$). If the effect of range restriction from $S$ to $Y_T$ is fully mediated by $X_T$, then the ratio of the restricted to the unrestricted SD of $S$ (i.e., $u_s = s_s/S_s$) can be fully reflected by the ratio of the restricted to the unrestricted SD of $X_T$ (i.e., $u_{X_T} = s_{X_T}/S_{X_T}$). Hence one only needs to know $u_{X_T}$ in order to make a correction, and this can be estimated through Equation 2.19 below. Second, the effect of range restriction from $S$ should go to $X_T$ rather than $X$ as assumed in the Case III model (Hunter et al., 2006). Third, given the complex structure of the five variables or constructs (i.e., $S$, $X_T$, $X$, $X_T$, and $Y$), the effects of range restriction from $S$ to others should differ substantially. This means that the conventional Case I to III models ignore the complexity of the range-restriction effects implicit in these variables or constructs, thereby leading to a potentially over-simplified correction procedure.

*Figure 3.* A conceptual model for the Case IV range-restricted correlation.



*Note*: ⇒ indicates the direction of range restriction

→ denotes the relationship between two variables (no causal relationship is assumed).

*S* is the selection construct. *X* is the observed score of variable *X*. *Y* is the observed score of variable *Y*. $X_T$ is the true score of *X*. $Y_T$ is the true score of *Y*. $e_X$ is the measurement error of *X*. $e_Y$ is the measurement error of *Y*. $\rho_{SX_T}$ is the population correlation between *S* and $X_T$. $\rho_{X_TY_T}$ is the population correlation between $X_T$ and $Y_T$. $\rho_{X_tX}$ is the population correlation between $X_T$ and *X*, and its square indicates the population reliability of *X*, i.e., $\rho_{XX}$, as shown in Equation 2.22 below. Likewise, $\rho_{Y_tY}$ is the population correlation between $Y_T$ and *Y*, and its square indicates the population reliability of *Y*, i.e., $\rho_{YY}$.

In equation form, the data matrix under the Case IV restriction can be shown as

$$\left[\begin{array}{ccc} \boldsymbol{S_r} & \boldsymbol{X_r} & \boldsymbol{Y_r} \\ \hline \boldsymbol{S_m} & \vdots & \vdots \end{array}\right]. \tag{2.17}$$

where $\boldsymbol{X_r}$ and $\boldsymbol{Y_r}$ are defined as above, $\boldsymbol{S_r}$ is the restricted sample that contains the restricted $S$ scores, and $\boldsymbol{S_m}$ is the unselected (or missing) sample that contains the unselected $S$ scores. In practice, the $S$ scores are unobservable, and hence the data become

$$\left[\begin{array}{cc} \boldsymbol{X_r} & \boldsymbol{Y_r} \\ \hline \vdots & \vdots \end{array}\right]. \tag{2.18}$$

According to Schmidt et al. (2006), the quantity $u_{X_T} = s_{X_T}/S_{X_T}$ is defined as the ratio of the restricted to the unrestricted SD of the true score of $X$, which can be estimated through

$$u_{X_T} = \sqrt{\frac{u_X^2 - (1 - r_{XX_u})}{r_{XX_u}}}\,, \tag{2.19}$$

where $r_{XX_u}$ is the reliability coefficient of $X$ in a group of unrestricted applicants, and $u_X = s_X/S_X$ is defined as the ratio of the restricted to the unrestricted SD of $X$. Given Equation 2.19, the parameter $u_{X_T}$ can be substituted into $u_X$ in the Case II correction Equation 2.14, producing the Case IV correction procedure,

$$r_{c4} = \left(\frac{1}{u_t}\right) r_{XY_c} \Bigg/ \sqrt{\left[\left(\frac{1}{u_t^2}\right) - 1\right] r_{XY_c}^2 + 1}\,, \tag{2.20}$$

where $r_{XY_c}$ is the restricted correlation between $X$ and $Y$ corrected for unreliability, and this is computed by

$$r_{XY_c} = r_{XY_r} \Big/ \left( \sqrt{r_{XX_r}} \cdot \sqrt{r_{YY_r}} \right), \qquad (2.21)$$

where $r_{XY_r}$ is the restricted correlation between *X* and *Y*, $r_{XX_r}$ is the restricted reliability of *X*, and $r_{YY_r}$ is the restricted reliability of *Y*.

Since Hunter, Schmidt, Le, and colleagues' work, the procedure of adjusting the correlation for the Case IV restriction in both single and meta-analytic studies has received increasing attention in the literature. For example, Li et al. (2011a) conducted a Monte Carlo study, and found that the Case IV corrected correlation in a single study was accurate across different data conditions, thereby providing empirical evidence for the adequacy of the Case IV corrected correlation.

The Case IV corrected correlation has not only received more attention in a single study, but it is also becoming more popular in a meta-analytic study, especially in the area of personnel psychology. Generally, meta-analysis is a statistical procedure that synthesizes the correlations reported in each single study, thereby producing a mean correlation and summarizing a general pattern of the correlational effect in a research domain. For example, Le and Schmidt (2006) re-analyzed a previously published meta-analysis, which examined the validity of employment interviews. They corrected all of the range-restricted correlations for the Case IV restriction before these correlations were used for estimating the mean correlation for meta-analysis. Such a correction can also be found in recent meta-analyses such as Christian, Bradley, Wallace, and Burke (2009) and Banks, Batchelor, and McDaniel (2010).

## Two Conventional Correction Procedures for Reliability

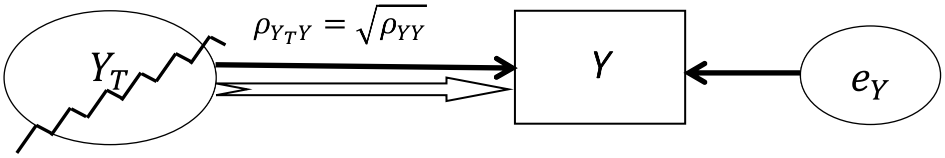
The effect of range restriction on reliability is expected to be comparable with that on Pearson’s correlation, given that both statistics depend on the variance of the sample. In fact, some researchers have regarded reliability as a type of correlation because reliability is conceptually considered as the squared correlation between true and observed scores, i.e.,

$$\rho_{Y_T Y} = \frac{Var(Y_T)}{SD(Y_T) \cdot SD(Y)} = \frac{SD(Y_T)}{SD(Y)} = \sqrt{\rho_{YY}} = \sqrt{\rho_{\alpha}}. \quad (2.22)$$

The next section will present two conventional (labeled Case I and III in this dissertation) correction procedures for reliability.

**Case I correction procedure for reliability**. Early discussion of the effect of range restriction on reliability can be found in Gulliksen’s (1950, 1987) handbook, which proposed a correction equation for the first type of selection process. Figure 4 shows a Case I conceptual model. When selection or range restriction occurs on total true scores of *Y*, the variance of these scores is usually reduced. Hence the measured reliability based on this sample is typically downward-biased. This type of restriction is in fact similar to the Case I correlation as displayed in Figure 1.

*Figure 4.* A conceptual model for the Case I restricted reliability coefficient.



*Note*: ⇨ indicates the direction of range restriction

→ denotes the relationship between two variables (no causal relationship is assumed).

$Y_T$ is the true score of variable *Y*. *Y* is the observed score of variable *Y*. $e_Y$ is the error of measurement of variable *Y*. $\rho_{Y_tY}$ is the correlation between $Y_T$ and *Y*, and its square indicates the reliability of *Y*, i.e., $\rho_{YY}$, as shown in Equation 2.22. Note that the ratio of the restricted to unrestricted SD of *Y* is known in Case I.

Conceptually, although the Case I correction procedure assumes that selection occurs based on the true score of *Y*, empirical studies showed that it appears to be robust to a situation in which selection is based on the observed score of *Y*. For example, Li, Cui, Gierl, and Chan (2012) simulated a situation, in which selection occurred based on the observed score of *Y*, and the Case I correction procedure was applied to adjust for the bias for reliability. Results showed that the Case I corrected alphas were consistently more accurate than the uncorrected alphas across different simulation conditions, including a highly stringent selection ratio (i.e., 10%) and a moderate sample size (i.e., 50).

In equation form, the data observed under the Case I restricted reliability can be expressed as

$$\begin{bmatrix} \boldsymbol{Y_r} \\ \hline \boldsymbol{Y_u} \end{bmatrix}, \tag{2.23}$$

where $\boldsymbol{Y_r}$ is the $n_r$ (restricted sample size) by *k* matrix containing the item scores of the restricted persons due to the rank-ordered total scores of $Y_T$, and $\boldsymbol{Y_u}$ is the $(n - n_r)$ by *k* matrix containing the item scores of *Y* of the unselected persons. When the data matrix $\boldsymbol{Y_u}$ is unavailable, Equation (2.23) becomes

$$\begin{bmatrix} \boldsymbol{Y_r} \\ \hline \vdots \end{bmatrix}. \tag{2.24}$$

For example, supposing that a researcher collects a sample of college students and estimates the reliability of an academic achievement test *Y* (e.g., SAT), but attempts to generalize this result to the general population of students. The estimated reliability of SAT is often downward-biased because the variance is much reduced with this restricted sample.

According to Gulliksen (1987), when the regression of *Y* on $Y_T$ is both linear and homoscedastic (which will be discussed in the following section), a reliability coefficient (e.g., coefficient alpha) corrected for the Case I range restriction ($r_{\alpha 1}$) can be expressed as

$$r_{\alpha 1} = r_{YY_u} = 1 - u_Y^2\left(1 - r_{YY_r}\right), \tag{2.25}$$

where $r_{\alpha 1}$ is a sample estimate of the coefficient alpha corrected for the Case I range restriction, $r_{YY_u}$ is the sample unrestricted reliability of $Y$, $r_{YY_r}$ is the sample restricted reliability of $Y$, and $u_Y = s_Y/S_Y$ is the ratio of the restricted to the unrestricted standard deviation (SD) of the total scores of *Y*. Gulliksen called this the homogeneity formula (i.e., Equation 20, p. 151), which lays a foundation correcting for the range-restricted reliability.

Note that Equation 2.25 is technically the same as Equation 2.12. Substituting $Y_T$ into *X* in Equation 2.12, we get

$$r_{Y_T Y_u} = \sqrt{1 - u_Y^2\left(1 - r_{Y_T Y_r}^2\right)}. \tag{2.26}$$

According to Equation (2.22), the population correlation between $Y_T$ and *Y* (i.e., $\rho_{Y_T Y}$) is equal to the square root of the population reliability (i.e., $\sqrt{\rho_{YY}}$), and thus Equation 2.26 can be expressed as

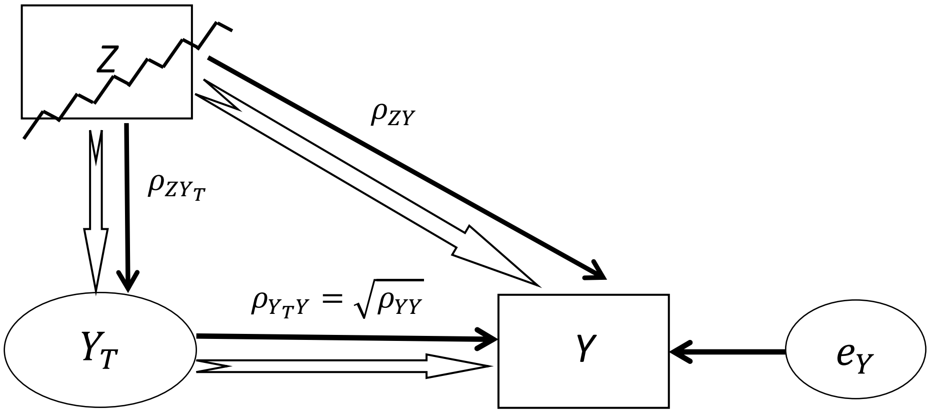
$$\sqrt{r_{YY_u}} = \sqrt{1 - u_Y^2\left(1 - r_{YY_r}\right)}. \tag{2.27}$$

Taking the square root on both sides, Equation 2.27 becomes the correction equation for the Case I restricted reliability (i.e., Equation 2.25).

**Case III correction procedure for reliability**. As in the Case III correction procedure for correlation, selection can occur based on another variable *Z*. Regarding reliability, suppose *Y* is a cognitive ability test that is evaluated by a researcher. When the researcher collects a sample of college students and estimates the reliability of *Y* but attempts to generalize the result to the entire population of students, the observed reliability is usually downward-biased, given that these students are selected based on another variable *Z* (e.g., SAT). In this case, the variance of *Y* is reduced by the selection process that occurs in *Z*. The correction procedure for this case of range restriction is called the Case III correction procedure for reliability in this dissertation.

Figure 5 shows a conceptual model for the range restriction that occurs in reliability, in which the SD ratio of the restricted to the unrestricted scores of $Z$ (i.e., $u_Z = s_Z/S_Z$) is used to adjust for the bias.

*Figure 5.* A conceptual model of the Case III range-restricted reliability.



*Note*: ⇨ indicates the direction of range restriction

→ denotes the structural relationship between two variables (no causal relationship is assumed).

$Y_T$ is the true score of variable *Y*. *Y* is the observed score of variable *Y*. *Z* is the observed score of variable *Z*. $\rho_{ZY}$ is the population correlation between $Z$ and *Y*. $\rho_{ZY_T}$ is the population correlation between $Z$ and $Y_T$. $\rho_{Y_tY}$ is the population correlation between $Y_T$ and *Y*, and its square indicates the reliability of *Y*, i.e., $\rho_{YY}$, as in Equation 2.22.

In equation form, the data matrix for the Case III restriction can be expressed as

$$\left[\begin{array}{ccc} \boldsymbol{Z_r} & \boldsymbol{Y_{T_r}} & \boldsymbol{Y_r} \\ \hline \boldsymbol{Z_m} & \vdots & \vdots \end{array}\right], \tag{2.28}$$

where $\boldsymbol{Z_r}$ is the restricted sample containing the restricted *Z* scores, $\boldsymbol{Z_m}$ is the unselected (or missing) sample containing the unselected *Z* scores, and $\boldsymbol{Y_{T_r}}$ and $\boldsymbol{Y_r}$ are the $n_r$ by *k* matrices containing the restricted true and observed *Y* scores as a result of the rank-ordered *Z* scores. In practice, the true scores are never available, and hence Equation 2.30 becomes

$$\left[\begin{array}{cc} \boldsymbol{Z_r} & \boldsymbol{Y_r} \\ \hline \boldsymbol{Z_m} & \vdots \end{array}\right], \tag{2.29}$$

According to Schmidt et al. (1976), when the regression of *Y* on *Z* is both linear and homoscedastic, the correction procedure for the Case III restricted reliability can be expressed as

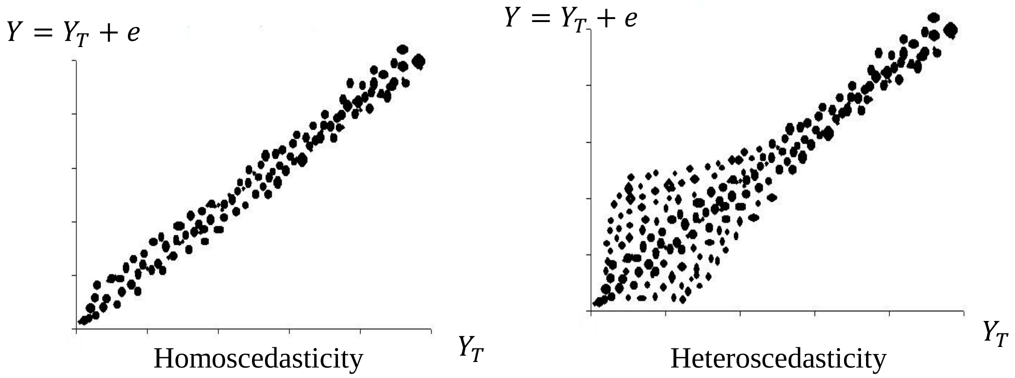
$$r_{\alpha 3} = r_{YY_u} = 1 - \frac{1-r_{YY_r}}{1-r^2_{YZ_r}[1-(1/u^2_Z)]}, \tag{2.30}$$

where $r_{\alpha 3}$ is the sample coefficient alpha corrected for the Case III restriction, $r_{YY_u}$ and $r_{YY_r}$ are the unrestricted and restricted sample reliability, $r_{ZY_r}$ is the sample restricted correlation between *Z* and *Y*, and $u_Z = s_Z/S_Z$ is the ratio of the restricted to the unrestricted SD of *Z*.

**A note on the assumption of the correction procedure for reliability.** Although the correction procedures for correlation and reliability have been discussed extensively in the literature and can be found frequently in measurement textbooks, it is important to note that these procedures are based on a key assumption; using the Case I corrected reliability as an illustration, this assumption requires that the regression of *Y* on $Y_T$ is both linear and homoscedastic. Some studies regarded this assumption as the homogeneity of the standard errors of measurement for the restricted and unrestricted groups of participants (e.g., Rodriguez & Maeda, 2006).

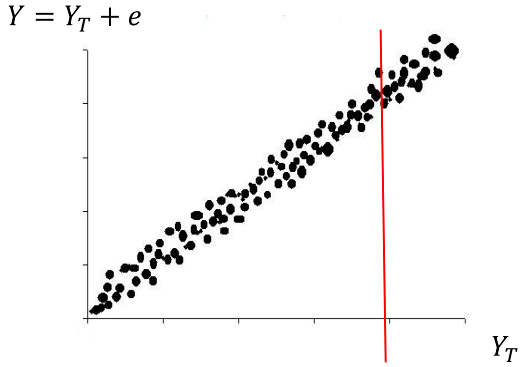
In statistics, the assumption of homoscedasticity means that the variance around the regression line is the same for all values of the predictor variable. Regarding the case of a reliability coefficient in the classical test theory, the horizontal axis indicates the true scores of $Y$ (i.e., $Y_T$), and the vertical axis refers to the observed scores of $Y$, as shown in Figure 6. According to classical test theory, $Y = Y_T + e$, where $Y$ is the total observed item scores, $Y_T$ is the total true item scores, and *e* is the standard error of measurement associated with $Y$. As indicated in Figure 6, the observed scores $Y$ are not the same as the true scores $Y_T$, and they should contain standard errors of measurement $e$. When $e$ is normally distributed and is independent of the true score of $Y$, the assumption of homoscedasticity is met. The opposite refers to heteroscedasticity, meaning that the standard error of measurement $e$ (and hence the observed score *Y*) depends on the true score of $Y$. For example, as shown in Figure 6, when one has a low score on $Y_T$, her or his observed score $Y$ should contain a larger standard error of measurement.

*Figure 6.* Diagrams showing the conditions of homoscedasticity and heteroscedasticity for a reliability coefficient in the classical test theory.



The reliability coefficients (e.g., the internal consistency reliability coefficients) developed from classical test theory are based on the assumption of homoscedasticity. Coefficient alpha is not an exception. This assumption implies the homogeneity of the standard errors of measurement for the restricted and unrestricted groups of participants. For example, supposing that a group of top $Y_T$ achievers is selected in Figure 7, the standard error of measurement of this restricted group is assumed to be the same as the unrestricted group.

*Figure 7*. A situation in which participants with the true scores of $Y$ above the cutoff are selected.



In light of the equivalency between the two assumptions (i.e., homoscedasticity and homogeneity of standard errors for the restricted and unrestricted groups), one may adjust a reliability coefficient for the artifact of range restriction, when the data meet the homoscedastic assumption presumed in the classical test theory. For the case of coefficient alpha, the data are assumed to meet at least the congeneric condition. That is, when the congeneric condition is met, the homoscedastic assumption is expected to be tenable.

Regarding the potential violation of the homoscedastic assumption, different studies found different results. For example, Blixt and Shama (1986) analyzed responses of 1,824 college freshmen on the Descriptive Tests of Mathematics Skills (DTMS) Elementary Algebra Skills Test (College Entrance Examination Board/Educational Testing Service, 1978), and found that the assumption of homoscedasticity was generally tenable. They concluded that "because of the simplicity of the classical approach in both calculation and practical use, it is recommended that the classical form of calculating standard error be retained" (p.

550). By contrast, Feldt and Qualls (1999) evaluated the data of the Iowa Tests of Educational Development (ITED; Feldt, Forsyth, Ansley, & Alnot, 1993) from 170 districts, and found that this assumption was generally not tenable. Fortunately, Fife et al. (2012) found that the corrected reliability still outperformed the uncorrected reliability substantially, even when this assumption is severely violated. In applied settings, researchers can check the standard error of measurement in their sample with the unrestricted sample (e.g., study report, technical report) before they use the correction procedure. For details, readers can refer to the aforementioned studies or Chapter 5 (i.e., a real study demonstration) in this dissertation.

### Previous studies evaluating the performance of the corrected reliability estimates

Although the majority of the studies have focused on the accuracy of the corrected correlation, some recent empirical studies have started to examine the performance of the uncorrected and corrected reliability coefficients.

**Sackett, Laczo, and Arvey (2002)**. In light of the fact that “range restriction affects estimates of reliability in the same manner as it affects estimates of criterion-related validity” (Sackett, Laczo, & Arvey, 2002, p. 56), Sackett et al. sought to examine the impact of the degree of range restriction on the (inter-rater) reliability coefficient by the use of a simulation study. In particular, they manipulated four levels of reliability (.60, .70, .80, and .90) and nine levels of selection ratio (.10 to .90 with an interval of .10), and evaluated the impact of range restriction on the reliability coefficient. They found that the downward biases ranged from 0.54% to 50.9% for the three scenarios, in which direct range restriction resulted in the largest biases when other factors were held constant. In addition, a smaller selection ratio led to a larger downward bias when other factors were held constant, because the variance of the variable is much more reduced given a smaller selection ratio. Generally speaking, these results suggested that the restricted reliability estimates are noticeably downward-biased, and they should be corrected for range restriction whenever the correction factors are available.

**Fife, Mendoza, and Terry (2012)**. In a recent study, Fife et al. (2012) examined the performance of different reliability coefficients (i.e., coefficient alpha, McDonald's omega, test-retest) under range restriction based on a Monte Carlo study. They generated dichotomized responses (i.e., 1 or 0) according to the item response theory (IRT) model, and evaluated the performance of these uncorrected and corrected reliability coefficients. Their results showed that all of the uncorrected reliability estimates were substantially downward-biased (with percentage biases ranging from -365.5% to -0.18%, i.e., Table 2) for the (more stringent) direct range restriction, while the corrected reliability coefficients were accurate (with percentage biases ranging from -0.18% to 13.76%, i.e., Table 4). Moreover, they provided a set of recommendations for dealing with the correction procedures for reliability. In particular, they suggested that when the sample is indirectly range-restricted by another variable, which is a usual case in practice, all of the three reliability coefficients (i.e., coefficient alpha, McDonald's $\omega$, test-retest) should be adjusted for range restriction. Fife et al. concluded that this can give more accurate estimates of the true reliability with reasonably small standard errors.

**Further issues**. This dissertation adds to the aforementioned studies by assessing the accuracy of the CIs surrounding the corrected reliability, and evaluating the issue of range restriction and reliability in the framework of meta-analysis. Both Sackett et al. (2002) and Fife et al. (2012) have examined the biases that come from the uncorrected reliability, and evaluated the performances of the correction procedures adjusting for the bias (i.e., Case I in Sackett et al., 2002, and Cases I and III in Fife et al., 2012). As an extension, this dissertation evaluates the performance of the CIs surrounding the uncorrected and corrected reliability. Fife et al. have assessed the standard error of the corrected reliability, and concluded that it is adequate because it produced a precise (or narrow) estimate. However, a narrow standard error does not necessarily mean a good estimate, given that it can produce too narrow CIs which adversely affects the coverage of the true reliability. Hence this dissertation also evaluates the coverage probability yielded by the CIs as well as their widths, in order to provide a more comprehensive assessment of the sampling error surrounding the uncorrected and corrected reliability. Furthermore, these studies did not examine the biases yielded by the uncorrected mean reliability and the benefits of reporting the corrected mean alpha in a meta-analytic study.

The following sections will first describe the CIs that can be built for the alpha in a single study, and then discuss the mean alpha and the associated CIs in a meta-analytic study.

## Confidence Intervals

In addition to the corrected coefficient alpha, one may also be interested in the associated sampling error and precision. The sampling error can be used to construct the confidence interval (CI) of the corrected alpha, which provides a way for statistical inferences about the estimate. For example, if one seeks to test whether a statistic (e.g., Pearson's correlation) is significant from zero at the .05 level, she or he can report the 95% confidence interval of the correlation estimate. If the interval does not contain the value of zero, this correlation estimate is significant at the .05 level. In the case of coefficient alpha, given that the value of zero means that there is no consistency of the test scores, one often seeks to evaluate coefficient alpha relative to a cutoff higher than zero. Often, in educational and psychological testing, a reliability of .70 is considered reasonable, and a reliability of .80 is considered good (Henson, 2001). Given these values, one can make statistical inferences about the reliability of a test relative to the pre-assigned value. Moreover, reporting the confidence intervals of different tests allows one to evaluate the accuracy of the sample estimate of alpha and different tests, scoring rubrics, or training procedures for raters or observers (Haertel, 2006). In addition, Section 2.07 of the manual of the American Psychological Association (2010) states that "because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy" (p. 34).

Considerable progress has been made on the parametric confidence interval (CI) estimation of coefficient alpha, thereby providing methods for making statistical inferences. For example, Feldt (1965) demonstrated that the $100(1 - \alpha_I)\%$ confidence interval for coefficient alpha, which is not subject to range restriction, can be estimated based on the central $F$-distribution, which is given by

$$[1 - (1 - r_\alpha)F_b, 1 - (1 - r_\alpha)F_a], \tag{2.31}$$

where $r_\alpha$ is the sample estimate of coefficient alpha, and $F_a$ and $F_b$ are the $100(\alpha_I/2)$ and $100(1 - \alpha_t/2)$ percentile points of the central $F$-distribution with $(n - 1)$ and $(n - 1)(k - 1)$ degrees of freedom.

Hakstian and Whalen (1976) used a normalizing transformation to approximate the confidence interval for coefficient alpha with standard normal distribution:

$$\left[\left\{1 - c^{*3}\left[(1 - r_\alpha)^{1/3} + z_{1-(\alpha_I/2)}\hat{\sigma}\right]^3\right\}, \left\{1 - c^{*3}\left[(1 - r_\alpha)^{1/3} - z_{1-(\alpha_I/2)}\hat{\sigma}\right]^3\right\}\right], \tag{2.32}$$

$$\text{where} \quad \hat{\sigma} = \sqrt{\frac{18k(n-1)(1-r_\alpha)^{2/3}}{k-1}}, \tag{2.33}$$

$$c^* = \frac{(9n-11)(k-1)}{9(n-1)(k-1)-2}, \tag{2.34}$$

and $z_{1-(\alpha_I/2)}$ is the $100(1 - \alpha_I/2)$ percentile point of the standard normal distribution.

van Zyl, Neudecker, and Nel (2000) derived the asymptotic normal distribution of sample coefficient alpha based on the assumption of the multivariate normal distribution of item responses. That is, as $n \to \infty$, $\sqrt{n}(r_\alpha - \rho_\alpha)$ is normally distributed with a mean of zero and a variance of

$$Q = \left[\frac{2k^2}{(k-1)^2(\boldsymbol{I}'\boldsymbol{\phi}\boldsymbol{I})^3}\right]\left[(\boldsymbol{I}'\boldsymbol{\phi}\boldsymbol{I})(tr\boldsymbol{\phi}^{\mathbf{2}} + tr^2\boldsymbol{\phi}) - 2(tr\boldsymbol{\phi})(\boldsymbol{I}'\boldsymbol{\phi}^{\mathbf{2}}\boldsymbol{I})\right], \tag{2.35}$$

where $\boldsymbol{I}$ is a $k \times 1$ vector of 1s, and $\boldsymbol{\phi}$ is the variance-covariance matrix of item responses. Given that the population coefficient alpha $\rho_{\alpha}$ is a constant, then the sample coefficient alpha is normally distributed with mean of $\rho_{\alpha}$ and variance of $Q/n$. Therefore, the $100(1-\alpha_I)\%$ confidence interval for coefficient alpha can be formed by

$$\left[r_{\alpha} \pm z_{1-(\alpha_I/2)}\sqrt{\frac{Q}{n}}\right]. \tag{2.36}$$

In a recent study, Cui and Li (2012) conducted a Monte Carlo study, which comprehensively evaluated the performance of these parametric CIs and three types of non-parametric bootstrap CIs [i.e., bootstrap standard interval (BSI), bootstrap percentile interval (BPI), and bootstrap bias-corrected and accelerated interval (BCaI). They found that the non-parametric bootstrap CIs, especially BSI, consistently outperformed the parametric CIs across data situations, including number of items, sample size, and item response distribution.

The aforementioned parametric methods, however, assume that the sample estimate of coefficient alpha is not subject to range restriction. Given that the sampling behaviors of the unrestricted and bias-corrected alphas should differ, these parametric procedures may not be applicable. For this reason, this dissertation evaluates the non-parametric bootstrap procedure in order to approximate empirically the sampling distribution of the corrected alpha so as to construct their bootstrap CIs. Generally, the bootstrap procedure is a non-parametric method for making statistical inferences. By drawing successive samples with replacement from an observed dataset, one can approximate the sampling behavior of a test statistic. According to Beasley and Rodgers (2009), the non-parametric bootstrap method has two unique advantages over the traditional parametric (distributional) method. The first advantage is that one can create an entirely new statistic without deriving the mathematical formula for estimating the standard error and confidence interval of the statistic. In this case, the sampling distribution of the corrected alpha is generally unknown; hence the non-parametric bootstrap procedure can be used to estimate the associated sampling error and confidence interval. A second advantage is that statistical inferences about nonnormal statistics are typically more accurate with the non-parametric bootstrap procedure. Because the maximum value of the corrected alpha is bounded by +1, its distribution is naturally nonnormal, especially when the true alpha of the test is high. Given these advantages, the non-parametric bootstrap procedure is used to derive the sampling error and estimate the confidence intervals of the alpha corrected for the two cases, thereby making statistical inferences about these statistics.

Indeed, using the bootstrap procedure to build the confidence interval for the bias-corrected correlation can be found in recent studies. For instance, to construct the confidence interval for the correlation corrected for the Case II restriction, Chan and Chan (2004) evaluated the performance of three types of bootstrap confidence intervals across different data situations, including the size of the correlation, selection ratio, and restricted sample size. They found that the bootstrap bias-corrected and accelerated interval (BCaI) produced good results systematically across the manipulated conditions. As an extension, Li et al. (2011a) conducted a simulation study in order to examine the accuracy of the same bootstrap confidence intervals for the correlations corrected for the Case III and Case IV restrictions. Their results showed that the BPI yielded accurate results across different data conditions. Given the reasonable results from these studies, this dissertation applies the bootstrap procedure to the alpha corrected for range restriction. Details will be discussed in Chapter 3.

### Meta-Analysis

As noted by Bonett (2010), the confidence width for coefficient alpha may be too wide for an accurate sampling error evaluation, especially when the sample size is small (e.g., less than 30) in a single study. An alternative is to evaluate the alpha of a test from multiple studies, and this technique is known as meta-analysis. Meta-analysis is a statistical procedure that synthesizes the quantitative findings provided in multiple studies conducted by independent researchers. In the literature, there are many approaches to meta-analysis, but most of them were based on a pioneer paper written by Glass (1976). Specifically, Glass was interested in the effect sizes (i.e., correlation) of four types of psychotherapy in relation to untreated control groups. The empirical findings reported in multiple studies were combined to a common effect size metric. Moreover, the distribution of these effect sizes was reported, and study level characteristics that could explain the distributions or variations of these effect sizes were discussed. Glass addressed the importance of meta-analysis in educational research—“[t]he need for the meta-analysis of research is clear. The literature on dozens of topics in education is growing at an astounding rate” (p. 3). Hence synthesizing empirical outcomes reported in each single study can provide a general pattern of these findings in a research domain, and evaluate the study characteristics (e.g., culture, gender) that cause the variability of these outcomes.

Vacha-Haase (1998) suggested that reliability evaluation of a test may be less accurate if it is based on a single study only. The meta-analysis of reliability coefficients appears to be a more trustworthy method, given that the empirical findings are based on multiple studies, which are expected to reflect the true reliability of a test more adequately. In particular, Vacha-Haase mentioned that "reliability generalization characterizes (a) the typical reliability of scores for a given test across studies, (b) the amount of variability in reliability coefficient for given measures, and (c) the sources of variability in reliability coefficients across studies" (p. 6).

**Meta-analysis of coefficient alpha.** Vacha-Hasse's (1998) framework has included different types of reliability coefficients such as test-retest, internal consistency, and inter-rater reliability. Rodriguez and Maeda's (2006) proposed and developed a framework specific to the meta-analysis of coefficient alpha. The framework is important and particularly relevant to the education and psychology literature, given that most authors have reported their coefficient alpha for the scores in their studies. Applied studies examining the mean alpha level of different educational and psychological scales can be found in Vassar and Bradley (2010), Vassar and Crosby (2008), and Warne (2011).

The following sections discuss the two common measures—the mean alpha and its CI—that are often generated from a meta-analysis of coefficient alpha.

**The mean alpha.** The meta-analysis of coefficient alpha is a statistical procedure that synthesizes the sample coefficient alphas $[r_\alpha(q), w = 1, \ldots, Q]$ reported in primary studies, thereby producing a mean alpha estimate ($\overline{r_\alpha}$) in a research domain. Given that coefficient alpha is non-normally distributed, Rodriguez and Maeda (2006) suggested that one needs to transform each singe alpha into the standardized alpha $T_\alpha$ by

$$T_\alpha(q) = |1 - r_\alpha(q)|^{1/3}. \tag{2.37}$$

These standardized alphas are assumed to be asymptotically normally distributed, and hence they can be used to make statistical inferences. The variance of each transformed alpha in study $q$ is

$$v(q) = \frac{18 \cdot k(q)[n(q)-1][1-r_\alpha(q)]^{2/3}}{[k(q)-1][9n(q)-11]^2}, \tag{2.38}$$

where $k(q)$ is the item number in study $q$, and $n(q)$ is the sample size in study $q$. The mean standardized alpha ($\overline{T_\alpha}$) is

$$\overline{T_\alpha} = \frac{\sum_{q=1}^{Q} w(q) T_\alpha(q)}{\sum_{q=1}^{Q} w(q)}, \tag{2.39}$$

$$\text{where } w(q) = 1/v(q) \tag{2.40}$$

is the reciprocal of the variance of each standardized alpha $T_\alpha(q)$. Consequently, the weighted mean alpha ($\overline{r_\alpha}$) is

$$\overline{r_\alpha} = \left|1 - \overline{T_\alpha}^3\right|, \tag{2.41}$$

which summarizes the mean alpha effect of the studies in a research domain. Note that if one seeks to evaluate the mean alpha corrected for range restriction, she or he can adjust the alphas for range restriction in each single study before they are used to estimate the mean alpha in Equation 3.24. Details will be discussed in Chapter 3.

**The CI surrounding the mean alpha.** As in a single study, meta-analysts may also be interested in the precision and sampling distribution of the mean alpha $\overline{r_\alpha}$. In the literature, there are two common types of intervals—confidence interval (CI) and credibility interval (CV)—that can be constructed around the mean coefficient alpha $\overline{r_\alpha}$ in a meta-analysis. Comparatively, CI is preferred when one seeks to make statistical inferences about the mean coefficient alpha $\overline{r_\alpha}$, whereas CV is preferred when one intends to evaluate the mean and variability of the population reliability underlying the primary studies. To provide a more comprehensive evaluation, meta-analysts have tended to report both intervals in recent studies (e.g., Crook et al., 2011; Ziegler, Dietl, Danay, Vogel, & Bühner, 2011).

Regarding the CI, meta-analysts usually construct a parametric CI surrounding the mean correlation estimate in validity generalization. This strategy was also proposed and discussed for reliability generalization in Rodriguez and Maeda (2006). The non-parametric bootstrap procedure for the mean alpha, which has not been examined in the literature, will be discussed in Chapter 3.

***Parametric CI.*** According to Rodriguez and Maeda (2006), the $100(1 - \alpha_I)\%$ confidence interval for the standardized mean alpha $\overline{T_\alpha}$ is

$$\overline{T_\alpha} \pm z_{\alpha_t/2}\sqrt{\bar{v}}, \tag{2.42}$$

$$\text{where } \bar{v} = 1/\textstyle\sum_{q=1}^{Q} w(q), \tag{2.43}$$

$$\text{and } w(q) = \left[\frac{18 \cdot k(q)[n(q)-1][1-r_\alpha(q)]^{2/3}}{[k(q)-1][9n(q)-11]^2}\right]^{-1}, \tag{2.44}$$

Consequently, the lower and upper limits of Equation 2.39 can be converted back into the metric of coefficient alpha, i.e.,

$$\overline{r_{\alpha}}(l) = |1 - T_{\alpha}(l)^{3}|, \tag{2.45}$$

$$\text{and } \overline{r_{\alpha}}(u) = |1 - T_{\alpha}(u)^{3}|, \tag{2.46}$$

where *l* is the lower limit and *u* is the upper limit.

**Summary**

This chapter reviewed the development of various reliability coefficients, discussed the background of range restriction, and presented the correction procedures for Thorndike’s (1949) three conventional cases and Hunter et al.’s (2006) fourth case of range restrictions for correlation. Moreover, it presented the two conventional cases of range restrictions for coefficient alpha, and discussed findings of the previous studies that examined their performances. In practice, researchers and applied users may also be interested in the confidence intervals (CIs) surrounding the corrected coefficient alpha. This chapter also introduced the parametric and non-parametric procedures for building these intervals. Given that the effect of range restriction on reliability is equally important in a meta-analytic study, this chapter also reviewed the background of estimating the mean alpha in a meta-analytic study, and discussed the methods for constructing the associated parametric and non-parametric CIs. Building on these concepts, the next chapter will present the design and methods in order to evaluate the empirical performance of the uncorrected and corrected coefficient alphas as well as the associated CIs in both single and meta-analytic studies.

# Chapter 3 – Method

The purpose of this chapter is to present the method that can be used to evaluate the four goals as discussed in Chapter 1. The first section presents the first Monte Carlo study, which evaluates the performance of the uncorrected and corrected alphas as well as their CIs in a single study; these correspond to Goals 1 and 2 as mentioned in Chapter 1. The second section presents the second Monte Carlo study, which evaluates the uncorrected and corrected mean alphas as well as their CIs in a meta-analytic study; these correspond to Goals 3 and 4 in this dissertation. As noted above, the Monte Carlo studies are regarded as a class of computational algorithms that generate empirical results across replications based on a repeated random sampling strategy. This can provide empirical evidence about the accuracy of uncorrected and corrected alphas in both single and meta-analytic research situations.

## Monte Carlo Study 1 – Single Study

### Goal 1: To Evaluate the Accuracy of the Uncorrected and Corrected Alphas in Single Study

Eight factors that may affect the accuracy of the uncorrected and corrected alphas were evaluated: type of measurement data ($\emptyset$), numbers of item ($k$), restricted sample size ($n_r$), selection ratio ($\pi$), unrestricted population coefficient alpha ($\rho_\alpha$), type of item responses ($\psi$), correlation between $Z$ and $Y_T$ for item $k$ [$\rho_{ZY_T}(k)$], and correlation between $Z$ and $Y$ for item $k$ [$\rho_{Z\mathrm{Y}}(k)$].

**Factor 1: Type of measurement data ($\emptyset$; three levels)**. Three types of measurement data—the essentially parallel, essentially tau-equivalent, and congeneric data conditions—were evaluated. As noted above, the essentially parallel condition is assumed to possess equal item variances and covariances, the essentially tau-equivalent condition is assumed to have unequal item variances but equal covariances, and the congeneric data condition is assumed to have items with unequal variances and covariances. Conventionally, coefficient alpha was developed based on the essentially parallel assumption. This dissertation evaluates whether the accuracy still exists when the data deviate from the most stringent essentially parallel condition to the less restrictive congeneric condition.

**Factor 2: Numbers of item (*k*; three levels)**. Three levels were controlled: 5, 10, and 20. Five items represent a typical short scale measuring an educational and psychological construct (e.g., the mastery goal orientation scale in the Achievement Goal Questionnaire [AGQ] by Elliot and Church, 1997), and 20 items indicate a relatively long scale (e.g., the hypochondriasis construct, i.e., concern with bodily symptoms, in the Minnesota Multiphasic Personality Inventory [MMPI-2]; Tellegen et al., 2003).

**Factor 3: Restricted sample size ($n_r$; three levels)**. Restricted sample sizes were manipulated at values of 30, 50 and 100. In a classic large-scale employee selection study, the median sample size is found to be 68 in 1500 validation studies (Lent, Aurbach, & Levin, 1971; Salgado, 1998). Recently, the sample size requirement has tended to increase. For example, the median size is found to be 138 across 36 studies in an employee selection meta-analysis study (Cass, Siu, Faragher, & Cooper, 2003). Hence, a restricted sample size of 100 is also controlled to evaluate whether or not a larger sample size is associated with better performance of the corrected alphas and the bootstrap procedure. The last level, 30, was used to evaluate whether the accuracy decreased with a small sample size.

**Factor 4: Selection ratio ($\pi$; five levels)**. Selection ratio is the ratio of the number of persons in the restricted sample over the unrestricted sample (i.e., $\pi = n_r/n$). Five levels were manipulated: .10, .30, .50, .70, and .90. The former (.10) represents a very stringent selection, meaning that only the top 10% of test achievers are selected. By contrast, the latter (.90) represents a very lenient restriction. These levels are comprehensive enough to represent different selection ratios in practice.

**Factor 5: Population coefficient alpha ($\rho_\alpha$; three levels)**. Three levels of coefficient alpha were evaluated: .70, .80, and .90.  Le and Schmidt (2006) stated that the mean of the unrestricted reliability was found to be .802 in the area of personnel psychology, according to previous meta-analyses and simulation studies. These levels also followed Nunnally's (1967) reliability benchmarks, i.e., .7 for tests in early development, .8 for basic research, and .9 to .95 for tests or instruments that are used to make important decision.

**Factor 6: Item responses ($\psi$; two levels)**. Two levels—continuous and dichotomous—were examined. The continuous item responses meet the assumption of the alpha. In practice, dichotomous item responses are even more common, and hence they were examined in this dissertation.

**Factor 7: Correlation between $Z$ and $Y_T$ for each item ($\rho_{ZY_T}$; two levels).** According to Hunter et al. (2006, p. 594), "range restriction in most data sets is indirect", meaning that range restriction usually occurs based on another variable as in the Case III model discussed in Chapter 2. Thus this dissertation followed Hunter et al.'s approach, in which a more general and realistic restriction model was used to generate the range-restriction effects. Specifically, the correlation between $Z$ and $Y_T$ is allowed to vary across items; this mimics a practical situation in which the effect of range restriction from another variable $Z$ on each item can be different. Following Le and Schmidt (2006), two levels of $\rho_{ZY_T}$—.30 and .60—were examined. These values represent a moderate to large relationship between $Z$ and $Y_T$, and they were frequently found and commonly used in both empirical and simulation studies (Hunter & Schmidt, 2004; Le & Schmidt, 2006). The associated SD was manipulated as one fifth of the population values.

**Factor 8: Correlation between $Z$ and $Y$ for each item ($\rho_{Z\mathrm{Y}}$; two levels).** The same two levels—.30 and .60—were evaluated as in the case of $\rho_{ZY_T}$. Likewise, the associated SD was manipulated as one fifth of the population values.

To summarize, the eight factors were combined to produce a design with $3 \times 3 \times 3 \times 5 \times 3 \times 2 \times 2 \times 2 = 3{,}240$ conditions. According to Mooney (1997), the minimum number of replications should be 1,000 for a Monte Carlo study. In addition, for a Monte Carlo study with a large number of design factors, Skrondal (2000) suggested that researchers can use fewer numbers of replications. Given that the present study involved 3,240 conditions and each condition consisted of 2,000 bootstrap samples, it was regarded as a complex design. Hence each condition was replicated 1,000 times.

**Data generation.** Generally, the data generation procedure followed Cui and Li’s (2012) study. First, for each combination of the simulated factors, without loss of generality, continuous normal item scores were generated according to the essentially parallel, essentially tau-equivalent, and congeneric data conditions, respectively. To generate the essentially parallel and essentially tau-equivalent data, the Type 12 sampling procedure proposed by Barchard and Hakstian (1997a) was used. This procedure assumes that both items and persons are a random sample from their corresponding populations (details will be discussed in the next section). To generate the congeneric data, the Type 12 sampling procedure was modified by removing a constraint, i.e., the person effects are assumed to be identical across items. By contrast, the person effects (i.e., person true scores, $T_i$ and $T_j$) for any two items *i* and *j* were assumed to take a linear relationship.

After continuous normal data are generated, the dichotomous item responses (i.e., 0 or 1) were obtained by dichotomizing the continuous data with the manipulated cut score of zero. Details will be discussed in the following section.

***Essentially parallel data.*** According to Barchard and Hakstian (1997a), the Type 12 sampling procedure for generating the essentially parallel data was based on the linear model, $Y_{ij} = \mu + I_j + T_i + e_{ij}$, as shown in Equation (2.4). To obtain observed scores for $Y_{ij}$, the four unknowns: $\mu$, $I_j$, $T_i$, and $e_{ij}$ were generated. First, without loss of generality, the grand mean $\mu$ was fixed at 0. Second, the item effects $I_j$ were assumed to vary across the items, which were generated from

$Normal{\sim}(0, 10^2)$. Third, the person effects (i.e., true scores), $T_1, T_2, \ldots, T_n$ were generated from a normal distribution with mean 0 and variance $\sigma_T^2$,

$$\sigma_T^2 = \frac{\rho_\alpha \sigma_{Y_j}^2}{k(1-\rho_\alpha)+\rho_\alpha}, \tag{3.1}$$

where $\rho_\alpha$ is the population coefficient alpha, $k$ is the number of items, and $\sigma_{Y_j}^2$ is the item variance, which is fixed at 100. Fourth, the measurement errors, $e_{ij}$, were generated for each item from $Normal{\sim}(0, \sigma_e^2)$, where $\sigma_e^2 = \sigma_{Y_j}^2 - \sigma_T^2$. Given the generated $I_j$, $T_i$, and $e_{ij}$ , and $\mu$ is fixed at 0, an observed score for person $i$ on item $j$ (i.e., $Y_{ij}$) was obtained by the linear equation in (2.4). By doing so, the generated data were multivariate normal and met the essentially parallel condition.

Dichotomous responses were generated through categorizing the continuous normal data using different cut scores. The *z* score of zero was selected as the cut point, and hence half of the generated normal data were assigned to the score of 0 and another half became the score of 1. Note that such a categorization led to changes in the values of the population item variance-covariance matrix, which resulted in a change in the population value of the coefficient alpha. To compute the population coefficient alpha for the dichotomous data, the population variances and covariances for items with $K$ categories scored from $0$ to $K-1$ were calculated as (Maydeu-Olivares, Coffman, & Hartmann, 2007; p. 175),

$$\sigma_{ii} = Var[W_i] = \left(\sum_{k=0}^{K-1} k^2 P(W_i = k)\right) - \mu_i^2, \tag{3.2}$$

and

$$\sigma_{iq} = Cov\left[W_i W_q\right] = \left(\sum_{k=0}^{K-1} \sum_{l=0}^{K-1} klP[(W_i = k) \cap \left(W_q = l\right)]\right) - \mu_i \mu_q, \tag{3.3}$$

where $P[(W_i = k) \cap (W_q = l)]$ is the probability that item $i$ takes the value $k$ and item $q$ takes the value $l$, and

$$\mu_i = E[W_i] = \sum_{k=0}^{K-1} kP(W_i = k). \tag{3.4}$$

***Essentially tau-equivalent data.*** $I_j$, $e_{ij}$, and $\mu$ were generated identically as in the essentially parallel condition, except for the person effects $T_i$. The difference between the essentially parallel and essentially tau-equivalent conditions lies in the unequal variances across items, $\sigma^2_{Y_j}$. Therefore, instead of setting item variance $\sigma^2_{Y_j}$ to be a constant and equal to 100 for all items, $\sigma^2_{Y_j}$ was generated from a normal distribution with mean 100 and variance 225. The person effect variance was, therefore, calculated by

$$\sigma^2_T = \frac{\rho_\alpha \mathrm{Mean}\left(\sigma^2_{Y_j}\right)}{k(1-\rho_\alpha)+\rho_\alpha}. \tag{3.5}$$

Although the essentially parallel condition assumes $T_{iq} = T_{ij} + C_{jq}$, without loss of generality, $C_{jq}$ was set to 0 because variances and covariances are unaffected by adding a constant to a variable. Therefore, the person effect $i$, $T_i$, was generated from a normal distribution with mean 0 and variance $\sigma^2_T$. For item $j$, measurement error $e_{ij}$ was generated from a normal distribution with mean 0 and variance $\sigma^2_{e_j}$, which was calculated by $\sigma^2_{e_j} = \sigma^2_{Y_j} - \sigma^2_T$. $I_j$ was generated in the same way for the essentially parallel and essentially tau-equivalent data. Finally, the observed score for person $i$ on item $j$ was obtained by $Y_{ij} = \mu + I_j + T_i + e_{ij}$. The generated data were normally distributed and met the essentially tau-equivalent condition.

***Congeneric data.*** The same procedure of generating the item effect $I_j$ for essentially parallel and essentially tau-equivalent data was used in the generation of congeneric data. However, the person effects vary across items for congeneric data. The person effects for any two items $j$ and $q$ were assumed to be linearly related, that is, $T_{iq} = b_{jq}T_{ij} + C_{jq}$, which implies $\sigma_{T_q}^2 = b_{jq}^2\sigma_{T_j}^2$. The variance of person effects for any item $j$, $\sigma_{T_j}^2$, was generated from a normal distribution with a mean equals to

$$\text{Mean}\left(\sigma_{T_j}^2\right) = \frac{\rho_\alpha \text{Mean}\left(\sigma_{Y_j}^2\right)}{k(1-\rho_\alpha)+\rho_\alpha}, \tag{3.6}$$

and SD equals to Mean $\left(\sigma_{T_j}^2\right)$ divided by 4. This SD is selected to ensure that negative values of $\sigma_{T_j}^2$ will rarely be generated. Next, person effects for item 1, $T_{i1}$ $(i = 1, 2, \cdots, n)$, were randomly generated from a normal distribution with mean 0 and variance $\sigma_{T_1}^2$. Person effects for any other item $j$ are calculated by $T_{ij} = b_{1j}T_{i1}$ $(i = 1, 2, \cdots, n)$, where $b_{1j}$ is equal to $\sqrt{\sigma_{T_j}^2/\sigma_{T_1}^2}$. It should be noted that $C_{jq}$ is again set to 0 without loss of generality. For any item $j$, measurement error $e_{ij}$ was generated from a normal distribution with mean 0 and variance $\sigma_{e_j}^2$, which was calculated by $\sigma_{e_j}^2 = \sigma_{Y_j}^2 - \sigma_{t_j}^2$. Finally, the observed score for person $i$ on item $j$ was obtained by $Y_{ij} = \mu + I_j + T_{ij} + e_{ij}$. The generated data were normally distributed and met the congeneric condition.

After generating the essentially parallel, essentially tau-equivalent, and congeneric data, the observed scores $Y_{ij}$ were obtained, and this formed a data matrix, $\boldsymbol{Y}$. Given the range-restriction model manipulated in this dissertation, it is another variable *Z* that causes range restriction on $Y_T$. Therefore, the observations in *Z* were generated by

$$Z' = b_1 Y_T + b_2 Y + e_Z, \tag{3.7}$$

where $b_1$ is the regression slope for $Y_T$ [i.e., $b_1 = \rho_{ZY_T} - \rho_{ZY}\sqrt{\rho_\alpha}/(1-\rho_\alpha)$], $b_2$ is the regression slope for *Y* [i.e., $b_2 = \rho_{ZY} - \rho_{ZY_T}\sqrt{\rho_\alpha}/(1-\rho_\alpha)$], and $e_Z$ is the error of measurement of *Z*, which was generated from a normal distribution with mean 0 and $SD = \sqrt{1-\rho^2_{ZY\cdot Y_T}}$, where $\rho_{ZY\cdot Y_T}$ is the correlation between *Z* and *Y* partial for $Y_T$, which is $\sqrt{(\rho^2_{ZY} + \rho^2_{ZY_T} - 2\rho_{ZY}\rho_{ZY_T}\sqrt{\rho_\alpha})/(1-\rho_\alpha)}$. By doing so, the mean correlation between *Z* and $Y_T$ is fixed at $\rho_{ZY_T}$ and the mean correlation between *Z* and *Y* is fixed at $\rho_{ZY}$, according to the manipulated values of .3 and .6, respectively. After generating the $Z'$ scores for all items, the selection composite variable *Z* for each examinee was obtained by averaging all of the $Z'(k)$ scores, i.e., $Z = \sum_{k=1}^{K} Z'(k)/K$.

Next, the generated $Y_T$ and *Y* scores were rank-ordered by the *Z* scores top down, thereby forming a restricted sample,

$$\left[\begin{array}{cc} \boldsymbol{Z_r} & \boldsymbol{Y_r} \\ \hline \boldsymbol{Z_m} & \vdots \end{array}\right]. \tag{3.8}$$

For the Case I correction, the required data matrix can be simplified as

$$\left[\begin{array}{c} \boldsymbol{Y_r} \\ \hline \vdots \end{array}\right]. \tag{3.9}$$

The sample estimates of the unrestricted parameters can be obtained from a simulated sample; these includes the sample unrestricted SD of total score of *Y* (i.e., $S_Y$) for Case I, and the sample unrestricted SD of *Z* (i.e., $S_Z$) for Case III.

The remaining simulation steps followed the bootstrap procedures from Equations (3.10) to (3.18), which will be discussed below. Consequently, the 95% parametric CIs, and the bootstrap standard interval (BSI), bootstrap percentile interval (BPI), and bootstrap bias-corrected and accelerated interval (BCaI) for the uncorrected and corrected alphas could be constructed. The number of the bootstrap samples was fixed at *B* = 2,000 to allow an accurate computation of a bootstrap percentile (Efron & Tibshirani, 1993). Each condition was replicated 1,000 times in order to evaluate the performance of the alphas and the CIs.

**Goal 2: To Examine the Performance of the Bootstrap CIs Surrounding the Uncorrected and Corrected Alphas in Single Study**

By drawing successive samples of data with replacement, the bootstrap procedure can derive the sampling distribution as well as the confidence interval of the statistics. In this section, the construction of the CI for a corrected alpha (i.e., $r_{\alpha c}$) is presented.

For illustrative purposes, only the Case I correction procedure is demonstrated here. Generally, the bootstrap procedure resampled the data matrix $\begin{bmatrix} \boldsymbol{Y_r} \\ \hline \vdots \end{bmatrix}$ *B* times in order to obtain *B* numbers of the resampled corrected alpha estimates, thereby constructing the associated sampling distribution and CI.

In particular, $n_r$ number of 1 by *k* vector (i.e., a vector containing the *Y* scores of a person) in $\boldsymbol{Y_r}$ (i.e., Equation 3.9) were randomly resampled with replacement to form a first bootstrap sample,

$$\begin{bmatrix} Y_r^*(1) \\ \hline \vdots \end{bmatrix}. \tag{3.10}$$

Given this data matrix, the first bootstrap corrected alpha $r^*_{\alpha c}(1)$ was obtained from Equation 2.25. We can convert this estimate into a standardized unit by $T_c(q) = |1 - r_{\alpha c}(q)|^{1/3}$. Repeating the bootstrap process $B$ times, $T^*_c(1), T^*_c(2), \ldots, T^*_c(B)$ were obtained. The bootstrap standard error is

$$SE_B = \sqrt{\sum_{b=1}^{B}[T^*_c(b) - T^*_c(.)]^2/(B-1)}, \tag{3.11}$$

where $T^*_c(.) = \sum_{b=1}^{B} T^*_c(b)/B$ is the mean of the $B$ number of $T^*_c$s.

**Bootstrap Standard Interval (BSI).** The 100(1 - $\alpha_I$)% BSI for the corrected alpha was then given by

$$T_c \pm z_{(1-\alpha_I/2)} \cdot SE_B \tag{3.12}$$

where $z_{(1-\alpha_I/2)}$ is the $100(1-\alpha_I/2)^{\text{th}}$ percentile point in a standard normal distribution, and $\alpha_I$ is the Type I error, which is fixed at 5%. This interval can be converted back on the alpha unit by $r_{\alpha c} = |1 - T_c^3|$.

**Bootstrap Percentile Interval (BPI).** To construct the BPI, $T^*_c(1), T^*_c(2), \ldots, T^*_c(B)$ were rank-ordered such that $T^*_c(1st) \leq T^*_c(2nd) \leq \cdots \leq T^*_c(Bth)$ . The 100(1 –$\alpha_I$)% BPI for $T_c$ is

$$(T^*_c[l],\ T^*_c[u]), \tag{3.13}$$

where $l = B(\alpha_I\ /2)$ and $u = B(1 - \alpha_I\ /2)$. This interval can be converted back on the alpha unit by $r_{\alpha c} = |1 - T_c^3|$.

**Bootstrap Bias-Corrected and Accelerated Interval (BCaI).** As reported in Chan and Chan (2004), the BPI tended to be biased, especially when the distribution of the bootstrap statistics is skewed. To correct for the skewness, Efron and Tibshirani (1993) proposed the bootstrap bias-corrected and accelerated percentile interval (BCaI). Two correction factors—*u* and *v*—were required. Of the two factors, *u* was used to adjust the median bias of the bootstrap corrected reliability coefficient,

$$u = \Phi^{-1}[\#(T_c^*(b) < T_c)/B], \tag{3.14}$$

where $\Phi^{-1}(.)$ is the inverse of the cumulative standard normal distribution function, and $\#[T_c^*(b) < T_c]/B$ is the proportion of the bootstrap $T_c^*$s below the original sample estimate of $T_c$. When the bootstrap distribution is symmetric, $\#[T_c^*(b) < T_c]/B$ will be close to 0.5 and hence *u* will be close to 0. The second correction factor *v* is regarded as the rate of change of the standard error of $T_c$ with respect to its true parameter value, which can be obtained by

$$v = \sum_{j=1}^{k}[T_c(.) - T_c(j)]^3/6\{\sum_{j=1}^{k}[T_c(.) - T_c(j)]^2\}^{\frac{3}{2}}, \tag{3.15}$$

where $T_c(j)$ is the jackknife value of $T_c$ generated by removing the *j*th row in Equation 3.30, and $T_c(.)$ is the mean of the $n$ jackknife estimates. With *u* and *v*, the lower and upper limits of BCaI became

$$\{T_c^*(B \cdot a_1), T_c^*[B \cdot (1 - a_2)]\}, \tag{3.16}$$

where $a_1 = \Phi\left\{u + \left[\left(u + z_{1-(\alpha_I/_2)}\right)/\left[1 - v\left(u + z_{1-(\alpha_I/_2)}\right)\right]\right]\right\}, \quad (3.17)$

and $a_2 = \Phi\left\{u + \left[(u - z_{1-(\alpha_I/_2)})/\left[1 - v(u - z_{1-(\alpha_I/_2)})\right]\right]\right\}. \quad (3.18)$

This interval can be converted back on the alpha unit by $r_{\alpha c} = |1 - T_c^3|$.

## Monte Carlo Study 2 – Meta-Analysis

### Goal 3: To Evaluate the Accuracy of the Uncorrected and Corrected Mean Alpha in Meta-Analysis

Studying the effect of range restriction on reliability is not only important in single study, it is also important in meta-analysis. As noted above, meta-analysis is a statistical procedure that synthesizes the alphas $[r_\alpha(q), q = 1, \ldots, Q]$ reported in primary studies, thereby producing a mean alpha estimate ($\overline{r_\alpha}$) in a research domain. In the literature, there are two data-generation models—the fixed-effects (FE) and random-effects (RE) models—which govern the distribution of the population coefficient alphas in multiple studies. This dissertation generated data based on the random-effects (RE) model, which is characterized by the heterogeneity of population coefficient alphas across studies. Assume the variance of the true unrestricted population alphas $\rho_{\alpha_u}$ across $k$ studies is $\sigma^2(\rho_{\alpha_u})$. The RE model is presumed to allow $\sigma^2(\rho_{\alpha_u}) > 0$. A second model, known as the fixed-effects (FE) model, is also available in the literature. It presumes the homogeneity of population coefficient alphas across studies, i.e., $\sigma^2(\rho_{\alpha_u}) = 0$. According to Schmidt, Oh, and Hayes (2009), the FE model is only a special case of the RE model. It often underestimates the variability of estimates in meta-analysis, and hence it is of limited use in practice.

Following the procedures outlined in Rodriguez and Maeda (2006), this dissertation examines and develops a meta-analytic procedure in order to synthesize the alphas corrected for range restriction, producing a summary of the mean corrected alpha in a research domain. Assume $\rho_{\alpha_u}(q)$ is the population unrestricted alpha across $W$ single studies, where $q = 1,2, \ldots, Q$. The RE model assumes that $\rho_{\alpha_u}(1) \neq \rho_{\alpha_u}(2) \neq \cdots \neq \rho_{\alpha_u}(Q)$, meaning that there is a variability of the true unrestricted alpha for each study. In practice, the sample in each study may be subject to range restriction due to another correlated variable $Z$. Hence, the alphas reported in these studies are typically the sample restricted alphas, i.e.,

$$r_{\alpha_r}(1), r_{\alpha_r}(2), \ldots, r_{\alpha_r}(Q). \tag{3.19}$$

These restricted alphas can be corrected for range restriction, producing a vector of sample corrected alphas,

$$r_{\alpha c}(1), r_{\alpha c}(2), \ldots, r_{\alpha c}(Q). \tag{3.20}$$

According to Rodriguez and Maeda (2006), they can be transformed to the study effects, $T(q)$, through

$$T_c(q) = |1 - r_{\alpha c}(q)|^{1/3}. \tag{3.21}$$

These study effects are assumed to be asymptotically normally distributed, and hence they can be used to make statistical inferences. The variance of each transformed alpha in study $q$ is

$$v(q) = \frac{18 \cdot k(q)[n_r(q)-1][1-r_{\alpha c}(q)]^{2/3}}{[k(q)-1][9n_r(q)-11]^2}, \tag{3.22}$$

where $k(q)$ is the item number in study $q$, and $n_r(q)$ is the restricted sample size in study $q$. The mean transformed study effect $T$ corrected for range restriction ($\overline{T_c}$) is

$$\overline{T_c} = \frac{\sum_{q=1}^{Q} w(q) T_c(q)}{\sum_{q=1}^{Q} w(q)}, \tag{3.23}$$

$$\text{where } w(q) = 1/v(q). \tag{3.24}$$

is the reciprocal of the variance of each study effect $T_c(q)$. Consequently, the weighted mean corrected alpha ($\overline{r_{\alpha c}}$) is

$$\overline{r_{\alpha c}} = \left|1 - \overline{T_c}^{\,3}\right|, \tag{3.25}$$

and this is a statistic which summarizes the mean alpha effect of the studies in a research domain. The next section will discuss the levels of the simulated conditions.

**Factor 1: Measurement data ($\emptyset$; three levels)**. The same three data conditions—essentially parallel, tau-equivalent, and congeneric—were evaluated.

**Factor 2: Number of items ($k$; two levels)**. The same two levels were evaluated: 5 and 20. The middle level of 10 in Study 1 was excluded to reduce the number of simulation conditions.

**Factor 3: Restricted sample size ($n_r$; two levels)**. Restricted sample sizes, $n_i(k)$, were generated from two normal distributions: *N*(100, 30) and *N*(300, 100). The mean of 100 was selected because the median incumbent sample size was 138 across 36 studies in a meta-analysis of employee selection (Cass, Siu, Faragher, & Cooper, 2003). The mean of 300 was used to evaluate whether a larger sample size would improve the accuracy. Regarding the variability, the sample-size distribution in meta-analysis tends to possess a large variability (i.e., 30 -150; Brannick, Yang, & Cafri, 2011). Thus the associated SDs were 30 and 100, which were approximately one third of the associated values. Note that these mean values were larger than those of 30 and 100 in Study 1 because this study needs to generate the variance of the restricted sample sizes. Hence a larger mean value is required to capture such the variance. The lower bound of the simulated sample size was fixed at 10.

**Factor 4: Selection ratio ($\pi$; five levels)**. The same five levels—.10, .30, .50, .70, and .90—were evaluated. In addition, given that the selection ratio should vary across studies in meta-analysis, the associated SD was controlled as one fourth of the mean levels. The lower bound was set at .03 and the upper bound was set at .97.

**Factor 5: Population coefficient alpha ($\rho_\alpha$; three levels)**. The same three levels of coefficient alpha were evaluated: .70, .80, and .90. The associated SD was controlled as one fifth of its unrestricted value.

**Factor 6: Item responses ($\psi$; two levels)**. Two levels—continuous and dichotomous—were examined. The continuous item responses lay a theoretical foundation for the performance of the corrected alpha. In practice, dichotomous item responses are even more common, and hence they were examined in this dissertation.

**Factor 7: Correlation between $\boldsymbol{Z}$ and $Y_T$ for each item ($\rho_{ZY_T}$; two levels).** The same two levels—.30 and .60—were examined. The associated SD was manipulated as one fifth of the corresponding value.

**Factor 8: Correlation between $\boldsymbol{Z}$ and $Y$ for each item ($\rho_{Z\mathrm{Y}}$; two levels).** The same two levels—.30 and .60—were evaluated. The associated SD was manipulated as one fifth of the corresponding value.

**Factor 9: Number of studies ($\boldsymbol{q}$; two levels)**. As noted in Field (2005), the minimum number of studies for a meta-analysis is around 15. In addition, the value of 30 is also evaluated to examine if an increased number of studies can improve the accuracy. Hence two numbers, 15 and 30, were selected and evaluated.

To summarize, the nine factors were combined to produce a design with $3 \times 2 \times 2 \times 5 \times 3 \times 2 \times 2 \times 2 \times 2 = 2{,}880$ conditions. Each condition was replicated 1,000 times.

**Data generation.** The data generation process followed the same procedure as in Study 1. The only difference lies in the generation of the data for $Q$ single studies in Study 2, rather than only one single study in Study 1. Therefore, the first step was to generate the unrestricted population coefficient alphas, $\rho_{\alpha_u}(q)$, $q = 1,2, \ldots, Q$, for multiple studies. Second, the data were generated for each single study based on the manipulated and generated $\rho_{\alpha_u}(q)$ in Step 1. The remaining procedures follow those in Study 1, except that the selection ratio for each single study was generated from their corresponding normal distributions rather than fixed in Study 1. After obtaining the observations for all the variables in each single study, they could be used to estimate the uncorrected and corrected coefficient alphas as in Equation 3.19. After that, the remaining procedures followed those outlined in Equations 3.20 – 3.25, and this produced the mean uncorrected and corrected alpha estimates for meta-analysis.

### Goal 4: To Examine the Performance of Bootstrap CIs Surrounding the Uncorrected and Corrected Mean Alphas in Meta-Analysis

As noted above, Adams et al. (1997) proposed a non-parametric bootstrap procedure so that one can resample the effect sizes reported in primary studies with replacement to derive the sampling distribution of the mean effect size. Li et al. (in press) also conducted a Monte Carlo study, and found that the bootstrap CIs surrounding the mean Case IV corrected correlation achieved better results than the conventional CIs. In this dissertation, suppose one collects $Q$ coefficient alphas, corrects for range-restriction biases, and converts them in a standardized unit, $T_c(q)$, by the use of $|1 - r_{\alpha c}(q)|^{1/3}$. One can apply a correction procedure in order to adjust for their range-restriction biases, and include the restricted sample sizes to form a matrix

$$\begin{bmatrix} T_c(1) & n_r(1) \\ \vdots & \vdots \\ T_c(Q) & n_r(Q) \end{bmatrix}. \tag{3.26}$$

Given this data matrix, one can resample $T_c$ and $n_r$ together [i.e., each row in Equation 3.26] with replacement to form a bootstrap sample,

$$\begin{bmatrix} T_c^*(1) & n_r^*(1) \\ \vdots & \vdots \\ T_c^*(Q) & n_r^*(Q) \end{bmatrix}. \tag{3.27}$$

With (3.27), the first bootstrap mean standardized alpha, $\overline{T_c}^*(1)$, can be computed by (3.23). Repeating this process *B* times (*B* can be fixed at 2,000, as suggested by Efron and Tibshirani, 1993), *B* numbers of bootstrap $\overline{T_c}^*$s are obtained

$$[\overline{T_c}^*(1), \overline{T_c}^*(2),\ldots, \overline{T_c}^*(B)]. \tag{3.28}$$

Next, we can estimate the associated standard error by measuring their SD empirically, i.e.,

$$\text{SE}^* = \sqrt{\textstyle\sum_{b=1}^{B}\left[\overline{T_c}^*(b) - \overline{T_c}^*(.)\right]^2/(B-1)}, \tag{3.29}$$

where $\overline{T_c}^*(.) = \sum_{b=1}^{B} \overline{T_c}^*(b)/B$ is the mean of the *B* numbers of $\overline{T_c}^*$s. Consequently, the 100(1 - $\alpha_I$)% bootstrap standard interval (BSI) is constructed by

$$\left[\overline{T_c} \pm z_{(1-\alpha_I/2)} \cdot \text{SE}^*\right], \tag{3.30}$$

where $z_{(1-\alpha_I/2)}$ is the $100(1-\alpha_I/2)$th percentile point in a standard normal distribution. This interval can be converted back on the alpha unit by $\overline{r_{\alpha c}} = \left|1 - \overline{T_c}^3\right|$.

To construct the bootstrap percentile interval (BPI), $\overline{T_c}^*(1), \overline{T_c}^*(2), \ldots, \overline{T_c}^*(B)$ are rank-ordered such that $\overline{T_c}^*[1st] \leq \overline{T_c}^*[2nd] \leq \cdots \leq \overline{T_{\alpha c}}^*[Bth]$. The 100(1 – $\alpha_I$)% BPI is

$$\{T_c{}^*[l],\ T_c{}^*[u]\}, \tag{3.31}$$

where $l = B(\alpha_I/2)$ and $u = B(1-\alpha_I/2)$. This interval can be converted back on the alpha unit by $\overline{T_{\alpha c}} = \left|1 - \overline{T_c}^3\right|$.

As in Study 1, the BPI tends to be biased, especially when the distribution of the bootstrap statistics is skewed. To correct for the skewness, two correction factors—$u$ and *v*—are required. Of the two factors, *u* is used to adjust the median bias of the bootstrap $\overline{T_c}$ estimates,

$$u = \Phi^{-1}\left[\#\left(\overline{T_c}^*(b) < \overline{T_c}\right)/B\right], \tag{3.32}$$

where $\Phi^{-1}(.)$ is the inverse of the cumulative standard normal distribution function, and $\#[\overline{T_c}^*(b) < \overline{T_c}]/B$ is the proportion of the bootstrap $\overline{T_c}^*$s below the original sample estimate of $\overline{r_{\alpha c}}$. When the bootstrap distribution is symmetric, $\#[\overline{T_c}^*(b) < \overline{T_c}]/B$ will be close to 0.5 and hence *u* will be close to 0. The second correction factor *v* is regarded as the rate of change of the standard error of $\overline{T_c}$ with respect to its true parameter value, which can be obtained by

$$v = \textstyle\sum_{k=1}^{K}[\overline{T_c}(.) - \overline{T_c}(k)]^3/6\{\sum_{k=1}^{K}[\overline{T_c}(.) - \overline{T_c}(k)]^2\}^{\frac{3}{2}}, \tag{3.33}$$

where $\overline{T_c}(k)$ is the jackknife value of $\overline{T_c}$ generated by removing the *k*th row in (8), and $\overline{T_c}(.)$ is the mean of the $n$ jackknife estimates. The lower and upper limits of BCaI become

$$\{\overline{T_c}^*(B \cdot a_1), \overline{T_c}^*[B \cdot (1 - a_2)]\}, \tag{3.34}$$

where $a_1 = \Phi\left\{u + \left[\left(u + z_{1-(\alpha_I/2)}\right)/\left[1 - v\left(u + z_{1-(\alpha_I/2)}\right)\right]\right]\right\}$, (3.35)

and $a_2 = \Phi\left\{u + \left[(u - z_{1-(\alpha_I/2)})/\left[1 - v(u - z_{1-(\alpha_I/2)})\right]\right]\right\}$. (3.36)

This interval can be converted back on the alpha unit by $\overline{T_{\alpha c}} = \left|1 - \bar{T}_c^{\,3}\right|$.

## Evaluation Criteria

**Criterion 1: Evaluating the accuracy of the uncorrected and corrected alphas**. To examine the accuracy of the corrected alpha, percentage bias ($\text{Bias}_\forall$) was used: $\text{Bias}_\forall = [(\forall - \rho_\alpha)/\rho_\alpha] \times 100\%$, where $\forall$ can be the restricted, Case I or III coefficient alphas, indicating their mean scores across 1,000 replications, respectively. As stated in Li et al. (2011a), a parameter estimate is considered good if a $\text{Bias}_\forall$ is within $\pm 5\%$, and reasonable if a $\text{Bias}_\forall$ is within $\pm 10\%$. To summarize the $\text{Bias}_\forall$s across the simulation conditions, Flores (1986) proposed a mean absolute percentage error: $\text{MAPE}_\forall = \sum_{i=1}^{N_m} |\text{Bias}_\forall(i)|/N_m$ , where $N_m$ is the number of model conditions. A $\text{MAPE}_\forall$ within 10% is considered an appropriate fit (Li et al., 2011a).

**Criterion 2: Evaluating the Performance of the CIs**. Because all of the CIs used in this study were based on the 5% significance level, the coverage across 1,000 replications should ideally be as close to the nominal value of 950 as possible (or .95 in terms of coverage probability). However, researchers seldom obtain an exact point estimate even if the true mean value is .95 underlying a distribution. To allow for sampling errors, an empirical coverage probability falling within [.922, .968] is considered appropriate, as stated in Li et al. (2011a). However, some researchers found that this criterion may still be too stringent to be fulfilled in practice. For example, Schmidt found that the average coverage probability of the CI surrounding Hunter et al.'s (2006) corrected correlation was around .89, and he suggested the use of the 10% cutoff as a criterion (Schmidt, personal communication, October 18th, 2010). Hence the 10% bias is still regarded as a reasonable result, meaning that an empirical coverage probability should be at least .855 [i.e., $.95 - (.95 \cdot 10\%)$], which is regarded as the lenient criterion of this study.

This study also evaluates the widths of the CIs. On one hand, it is desirable to have a narrower interval because it indicates a precise point estimate. On the other hand, an over-precise interval tends to produce an under coverage probability because it is too narrow and does not span the population value appropriately. In this sense, a CI with a desirable coverage probability and a reasonable width is the most adequate procedure.

# Chapter 4 –Results

## Monte Carlo 1 – Single Study

This section presents the results for the uncorrected and corrected coefficient alphas, as well as for their CIs, in a single study. These results seek to provide empirical findings about the biases that come from the uncorrected alpha, and examine whether or not the two corrected alphas and the associated CIs can improve the accuracy of these measures. Given that the patterns of the uncorrected and corrected alphas did not differ across the three data conditions (i.e., parallel, tau-equivalent, and congeneric data), the results based on the most relaxed congeneric condition are discussed. Hence, a total of 540 conditions (i.e., 3 levels of item number by 3 levels of restricted sample size by 3 levels of alpha by 2 levels of correlation between $Z$ and $Y_T$ by 2 levels of correlation between $Z$ and $Y$ by 5 levels of selection ratio) are presented in the following sections. The results are discussed based on the continuous and dichotomous responses, respectively.

## Results for Goal 1: Evaluating Coefficient Alpha in Single Study

### Continuous Responses

Figure 8 shows the means of the 1,000 replicated uncorrected and two corrected alphas across the 540 simulation conditions, and Figure 9 displays the associated percentage biases. Generally, the uncorrected alphas were inadequate and fluctuated substantially, especially across levels of selection ratio, correlation between $Z$ and $Y_T$, and correlation between $Z$ and $Y$. Of the 540 conditions, only 22 (or 4.1%) were within the stringent level of $\pm$5%, and 74 (or 13.7%) were within the lenient level of ±10%. The percentage biases ranged from -99.7% to -1.2%. Regarding the overall bias, the MAPE was 30.9%, which fell outside the nominal level of 10%. Hence the uncorrected alpha estimates were generally inadequate.

*Figure 8.* Means of the 1,000 replicated uncorrected and corrected coefficient alpha estimates across 540 simulation conditions for continuous responses.

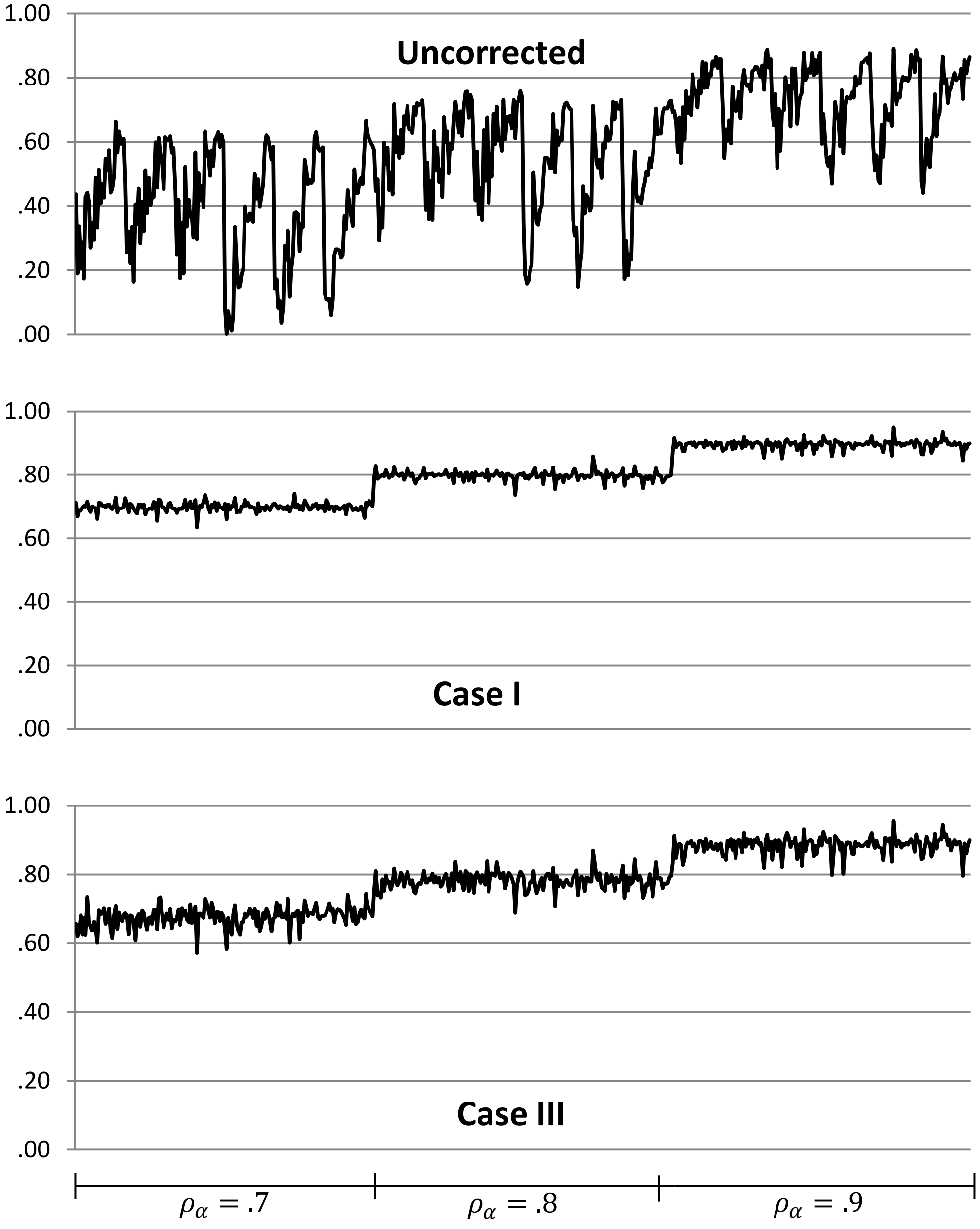
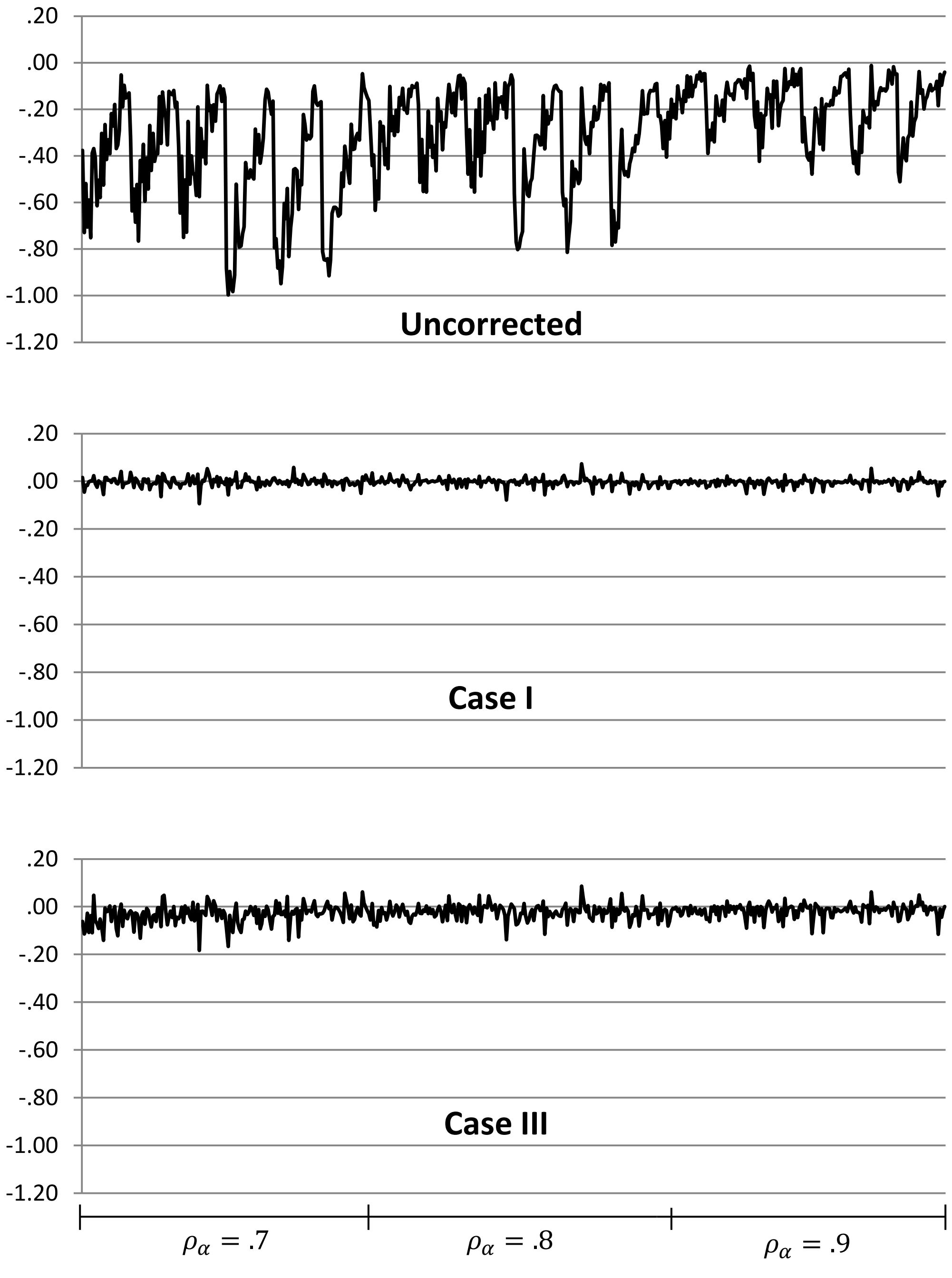


*Figure 9.* Means of the 1,000 replicated percentage biases obtained by the uncorrected and corrected alphas across 540 simulation conditions for continuous responses.



By contrast, the two bias-corrected alphas were more accurate than the uncorrected alpha. For Case I, the percentage biases ranged from -9.4% to 7.3%. Of the 540 conditions, 524 (or 97.0%) produced a percentage bias within the stringent level of $\pm 5\%$, and 540 (or 100%) were within the lenient level of $\pm 10\%$. Overall, the MAPE was 1.2%, which was within the criterion of 10%. For Case III, the percentage biases ranged from –18.3% to 8.5%. Of the 540 conditions, 427 (79.1%) were within the stringent level, and 522 (or 96.7%) were within the lenient level. The MAPE was 3.1% which was also within the nominal level of 10%. These findings are summarized in Table 1. They are comparable with the findings reported in Fife et al. (2012), in which the MAPE yielded by the corrected alpha was .75% (see Table 3 in their study). Their slightly smaller MAPE was probably, first, due to the relatively simpler range restriction process in their study, i.e., the correlation between $Z$ and $Y_T$ was the same for all items. Second, their study assumed that the true population unrestricted SD was known, while this study mimicked a more realistic situation in which only a sample estimate of the unrestricted SD from a simulated sample was obtained to make a correction, leading to a slightly larger fluctuations in the findings.

Table 1. *Summary of the percentage biases obtained by the uncorrected and corrected alphas for continuous responses.*

| Alpha | Mean | SD | [Max, Min] | MAPE | # within<br>±5% bias | # within<br>±10% bias |
|---|---|---|---|---|---|---|
| UnCor | -30.9% | 21.4% | [-99.7%, -1.2%] | 30.9% | 29(5.4%) | 92(17.0%) |
| Case I | -0.4% | 1.8% | [-9.4%, 7.3%] | 1.2% | 524(97.0%) | 540(100%) |
| Case III | -2.4% | 3.4% | [-18.3%, 8.6%] | 3.1% | 427(79.1%) | 522(96.7%) |

*Note*: UnCor is uncorrected. # is number of conditions.

This section discusses the specific effect of each factor on the findings. Given that the number of items and restricted sample size did not show an obvious effect on the uncorrected alpha, the results are presented based on a total of 60 conditions (i.e., 3 population alpha levels by 2 correlation levels between $Z$ and $Y_T$ by 2 correlation levels between $Z$ and $Y$ by 5 levels of selection ratio). Three key findings emerged, as shown in Table 2. First, when the selection ratio increased, the percentage bias decreased gradually. This is reasonable because a less stringent ratio produces a more heterogeneous sample, which represents an unrestricted sample more adequately. Second, when the correlations between $Z$ and $Y_T$ or between $Z$ and $Y$ (or both) became stronger, the biases increased gradually. It is because a stronger relationship between $Z$ and $Y_T$ (or between $Z$ and $Y$, or both) could generate a more homogeneous sample, resulting in a more downward-biased estimate of coefficient alpha. Third, when the alpha increased from .7 to .9, the bias decreased gradually. It appears that, when the scores are more internally consistent, the impact of range restriction on reliability is reduced.

Table 2. *Means of the 1,000 replicated percentage biases obtained by the uncorrected alpha across 60 selected simulation conditions when the item number =10, restricted sample size = 50, and items were continuously scored.*

| | | $\rho_\alpha = .7$ | | $\rho_\alpha = .8$ | | $\rho_\alpha = .9$ | |
|---|---|---|---|---|---|---|---|
| $\rho_{ZY}$ | $\pi$ | $\rho_{ZY_T} = .3$ | .6 | $\rho_{ZY_T} = .3$ | .6 | $\rho_{ZY_T} = .3$ | .6 |
| .3 | .1 | **-.542** | **-.685** | **-.330** | **-.552** | **-.291** | **-.331** |
| | .3 | **-.350** | **-.595** | **-.274** | **-.407** | **-.226** | **-.224** |
| | .5 | **-.303** | **-.424** | **-.210** | **-.374** | **-.131** | **-.140** |
| | .7 | **-.126** | **-.191** | **-.138** | **-.214** | -.074 | -.075 |
| | .9 | **-.137** | **-.119** | -.053 | **-.156** | -.029 | -.035 |
| .6 | .1 | **-.883** | **-.851** | **-.585** | **-.814** | **-.433** | **-.380** |
| | .3 | **-.539** | **-.833** | **-.457** | **-.471** | **-.206** | **-.240** |
| | .5 | **-.464** | **-.630** | **-.331** | **-.351** | **-.153** | **-.203** |
| | .7 | **-.332** | **-.312** | **-.188** | **-.246** | **-.119** | **-.112** |
| | .9 | **-.173** | **-.183** | -.099 | **-.120** | -.072 | -.087 |

*Note*. $\rho_\alpha$ is the population alpha, $\rho_{ZY}$ is the population correlation between *Z* and *Y*. $\pi$ is the selection ratio. $\rho_{ZY_T}$ is the population correlation between *Z* and $Y_T$. Biases that are outside the nominal range of $\pm 10\%$ are presented in bold.

Regarding the Case I corrected alpha, most of the conditions (i.e., 524 out of 540) produced a percentage bias within the stringent nominal level of $\pm 5\%$. Only very few minor exceptions could be found when the population alpha was small (i.e., .70), the item number was few (i.e., 5), the selection ratio was stringent (i.e., .10), the restricted sample size was small (i.e., 30) and the correlations between *Z* and *Y* (and *Z* and $Y_T$) were large, as shown in Table 3. These levels are regarded as the most challenging conditions in this study, but the biases are in fact quite small. Similar patterns but slightly larger biases could also be found for the Case III corrected alpha; these are shown in Table 4.

Table 3. *Means of the 1,000 replicated percentage biases obtained by the Case I alpha across 20 selected simulation conditions when the population alpha = .70, item number = 5, and restricted sample size = 30 for continuous responses.*

| | $\rho_{ZY} = .3$ | | $\rho_{ZY} = .6$ | |
|---|---|---|---|---|
| $\pi$ | $\rho_{ZY_T} = .3$ | .6 | $\rho_{ZY_T} = .3$ | .6 |
| .1 | .016 | **-.055** | .012 | **-.057** |
| .3 | .000 | .023 | .039 | -.028 |
| .5 | -.010 | -.045 | .031 | .004 |
| .7 | -.004 | .009 | -.006 | -.035 |
| .9 | .041 | -.026 | -.018 | .004 |

*Note*. $\pi$ is the selection ratio. $\rho_{ZY}$ is the population correlation between $Z$ and $Y$. $\rho_{ZY_T}$ is the population correlation between $Z$ and $Y_T$. Biases that are outside the stringent criterion of $\pm5\%$ are presented in bold.

Table 4. *Means of the 1,000 replicated percentage biases obtained by the Case III alpha across 20 selected simulation conditions when population alpha = .70, item number = 5, and restricted sample size = 30 for continuous responses.*

| | $\rho_{ZY} = .3$ | | $\rho_{ZY} = .6$ | |
|---|---|---|---|---|
| $\pi$ | $\rho_{ZY_T} = .3$ | .6 | $\rho_{ZY_T} = .3$ | .6 |
| .1 | -.062 | **-.114** | -.088 | **-.167** |
| .3 | **-.109** | .048 | .014 | -.066 |
| .5 | -.085 | -.094 | -.018 | -.033 |
| .7 | -.062 | -.024 | .000 | -.069 |
| .9 | .011 | -.083 | -.040 | .002 |

*Note*. $\pi$ is the selection ratio. $\rho_{ZY}$ is the population correlation between *Z* and *Y*. $\rho_{ZY_T}$ is the population correlation between *Z* and $Y_T$. Biases that are outside the nominal criterion of ±10% are presented in bold.

**Dichotomous Responses**

The previous section presents the results based on the continuous responses, providing theoretical support for the adequacy of the bias-corrected alphas when the scores are represented by the original continuous scales. In practice, researchers often use a test or scale with dichotomized responses, i.e., either 0 or 1. Generally, the findings and patterns of relationships are similar to those in continuous data, as shown in the previous section, although less accurate (or more fluctuated) results are obtained.

Because the initial results showed that the selection ratio of .1 often led to a perfectly homogenous sample (i.e., all 1s in any item) with zero variance, this condition was dropped and will not be discussed in the following section. As a result, a total of 432 conditions (i.e., 3 levels of item number by 3 levels of restricted sample size by 3 levels of alpha by 2 levels of correlation between $Z$ and $Y_T$ by 2 levels of correlation between $Z$ and $Y$ by 4 levels of selection ratio) are discussed.

Figure 10 shows the means of the 1,000 replicated uncorrected and two corrected alphas across the 432 simulation conditions, and Figure 11 displays the associated percentage biases. Generally, the uncorrected alphas were inadequate. The percentage biases ranged from -136.2% to -3.3%, with a mean of -43.1%. Of the 432 conditions, only 13 (or 3.0%) were within the stringent level of $\pm 5\%$, and 48 (or 11.1%) were within the lenient level of $\pm 10\%$. The overall MAPE was 43.1% outside the nominal level of 10%. Note that data dichotomization resulted in smaller true alpha values, and the corresponding true population values are presented on the horizontal axis of Figures 10 – 11.

*Figure 10*. Means of the 1,000 replicated uncorrected and corrected alpha estimates across 432 simulation conditions for dichotomous responses.

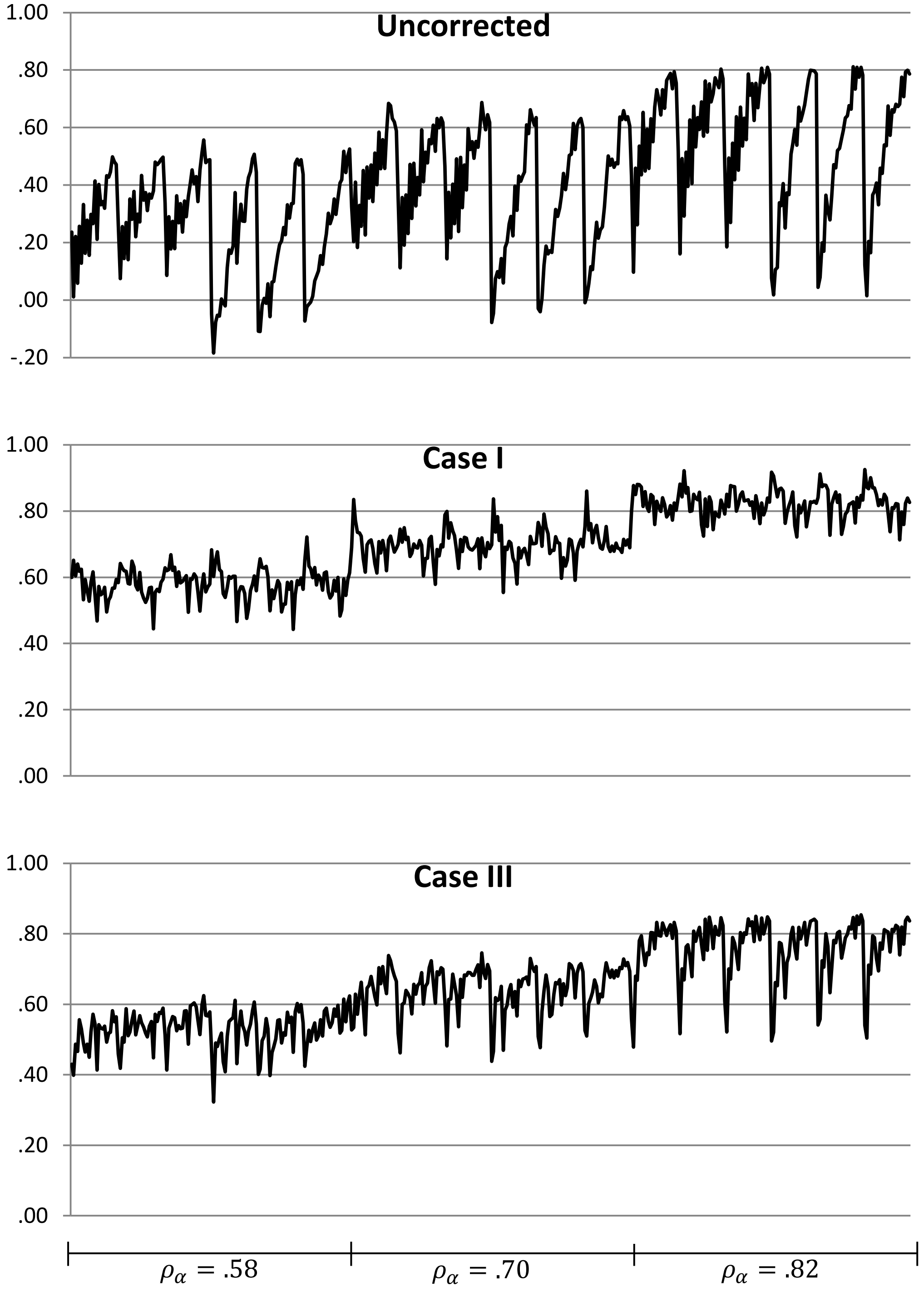
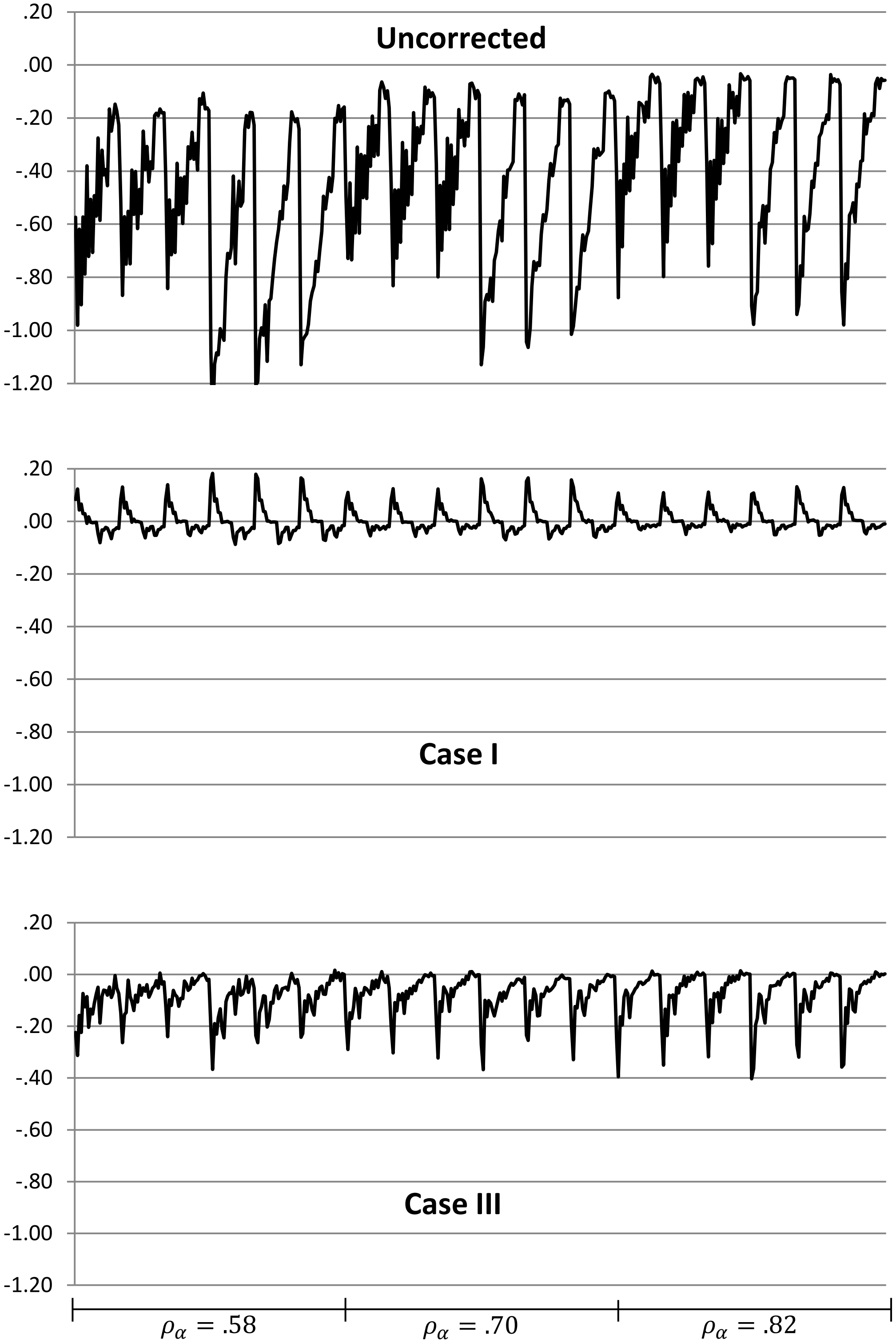


*Figure 11*. Means of the 1,000 replicated percentage biases yielded by the uncorrected and corrected alphas across 432 simulation conditions for dichotomous responses.



As in the continuous responses, the Case I and III corrected alphas improved the performance noticeably, and the former outperformed the latter. For Case I, the percentage biases ranged from -8.7% to 18.2%, with a mean of 4.8%. Of the 432 conditions, 339 (or 78.5%) resulted in a bias within the stringent criterion of ±5%, and 406 (or 94.0%) within the lenient level of ±10%. The overall MAPE was 3.4% within the nominal level of 10%. For Case III, the mean percentage bias was -7.9% with a range of [-40.3%, 1.6%]. Of the 432 conditions, only 190 (or 44.0%) produced a bias within the stringent criterion, and 317 (or 73.4%) yielded a bias inside the lenient region. A summary of the results yielded by the uncorrected and corrected alphas is provided in Table 5.

Table 5. *Summary of the percentage biases obtained by the uncorrected and corrected alphas for dichotomous responses.*

|  |  |  |  |  | # within | # within |
|---|---|---|---|---|---|---|
| Alpha | Mean | SD | [Max, Min] | MAPE | ±5% | ±10% |
| UnCor | -43.1% | 29.1% | [-136.2%, -3.3%] | 43.1% | 13(3.0%) | 48(11.1%) |
| Case I | 0.3% | 4.8% | [-8.7%, 18.2%] | 3.4% | 339(78.5%) | 406(94.0%) |
| Case III | -7.9% | 8.0% | [-40.3%, 1.6%] | 8.0% | 190(44.0%) | 317(73.4%) |

*Note*: UnCor is uncorrected. # indicates number of conditions

This section discusses the specific effect of each factor on the findings. Given that the item number and restricted sample size did not show obvious effects on the uncorrected alpha, the results are presented based on a total of 48 conditions (i.e., 3 population alpha levels by 2 correlation levels between $Z$ and $Y_T$ by 2 correlation levels between $Z$ and $Y$ by 4 levels of selection ratio), as shown in Table 6. The effects of the four factors remained the same as in the results obtained from the continuous response, and hence details are not repeated here.

Table 6. *Means of the 1,000 replicated percentage biases obtained by the uncorrected alpha across 60 selected simulation conditions when the item number was fixed at 10 and restricted sample size was fixed at 50 for dichotomous responses.*

| | | $\rho_\alpha = .58$ | | $\rho_\alpha = .70$ | | $\rho_\alpha = .82$ | |
|---|---|---|---|---|---|---|---|
| $\rho_{ZY}$ | $\pi$ | $\rho_{ZY_T} = .3$ | .6 | $\rho_{ZY_T} = .3$ | .6 | $\rho_{ZY_T} = .3$ | .6 |
| .3 | .3 | **-.572** | **-.750** | **-.471** | **-.728** | **-.392** | **-.667** |
| | .5 | **-.402** | **-.616** | **-.323** | **-.527** | **-.208** | **-.364** |
| | .7 | **-.307** | **-.441** | **-.236** | **-.293** | **-.104** | **-.246** |
| | .9 | **-.192** | **-.166** | -.095 | **-.120** | -.058 | -.075 |
| .6 | .3 | **-1.032** | **-.991** | **-.992** | **-.835** | **-.757** | **-.794** |
| | .5 | **-.880** | **-.796** | **-.639** | **-.564** | **-.438** | **-.461** |
| | .7 | **-.580** | **-.456** | **-.417** | **-.365** | **-.219** | **-.221** |
| | .9 | **-.198** | **-.215** | **-.133** | **-.135** | -.052 | -.062 |

*Note*. $\rho_\alpha$ is the population alpha, $\rho_{ZY}$ is the population correlation between *Z* and *Y*. $\pi$ is the selection ratio. $\rho_{ZY_T}$ is the population correlation between *Z* and $Y_T$. Biases that are outside the nominal range of $\pm10\%$ are presented in bold.

Regarding the Case I corrected alpha, most of the conditions produced a percentage bias within the lenient level of $\pm10\%$, as shown in Table 7. Only some minor exceptions could be found when the population alpha was small (i.e., .70), item number was small (i.e., 5), selection ratio was stringent (i.e., .30), and the correlations between $Z$ and $Y$ (and $Z$ and $Y_T$) were large. As noted above, these levels are regarded as the most challenging conditions in this study, but the biases are still quite reasonable (with a maximum of 17.9%). Note that the bias appeared to be slightly positive when the selection ratio became more stringent. Similar patterns of relationships were found for the Case III corrected alpha, as shown in Table 8; however, the magnitudes of the biases were slightly larger, and the positive biases found in Case I disappeared even when the selection ratio was stringent.

Table 7. *Means of the 1,000 replicated percentage biases obtained by the Case I alpha across 48 selected simulation conditions when the population alpha = .58 and restricted sample size = 50 for dichotomous responses.*

| | | | $k$ = 5 | | $k$ = 10 | | $k$ = 20 | |
|---|---|---|---|---|---|---|---|---|
| $\rho_\alpha$ | $\rho_{SY}$ | $\pi$ | $\rho_{SY_T} = .3$ | .6 | $\rho_{SY_T} = .3$ | .6 | $\rho_{SY_T} = .3$ | .6 |
| .58 | .3 | .3 | **.085** | **.131** | .051 | .073 | .029 | .039 |
| | | .5 | -.004 | .009 | -.001 | .000 | -.003 | -.004 |
| | | .7 | -.040 | -.063 | -.027 | -.035 | -.021 | -.020 |
| | | .9 | -.054 | -.053 | -.032 | -.030 | -.021 | -.021 |
| | .6 | .3 | **.179** | **.162** | .080 | .081 | .038 | .037 |
| | | .5 | .000 | .001 | -.003 | -.003 | -.002 | -.003 |
| | | .7 | -.084 | -.080 | -.043 | -.040 | -.023 | -.023 |
| | | .9 | -.069 | -.050 | -.035 | -.035 | -.021 | -.024 |

*Note*. $\rho_\alpha$ is the population alpha, $\rho_{ZY}$ is the population correlation between $Z$ and $Y$. $\pi$ is the selection ratio. $\rho_{ZY_T}$ is the population correlation between $Z$ and $Y_T$, $k$ is item number. Biases that are outside the nominal range of $\pm 10\%$ are presented in bold.

Table 8. *Means of the 1,000 replicated percentage biases obtained by the Case III alpha across 48 selected simulation conditions when the population alpha = .58 and restricted sample size = 50 for dichotomous responses.*

| | | | $k$ = 5 | | $k$ = 10 | | $k$ = 20 | |
|---|---|---|---|---|---|---|---|---|
| $\rho_\alpha$ | $\rho_{SY}$ | $\pi$ | $\rho_{SY_T}$ = .3 | .6 | $\rho_{SY_T}$ = .3 | .6 | $\rho_{SY_T}$ = .3 | .6 |
| .58 | .3 | .3 | **-.147** | **-.263** | **-.155** | **-.147** | **-.022** | **-.086** |
| | | .5 | **-.104** | **-.139** | -.081 | **-.106** | -.042 | -.080 |
| | | .7 | -.052 | -.065 | -.039 | -.088 | -.074 | -.053 |
| | | .9 | -.045 | -.024 | -.053 | .005 | -.038 | -.029 |
| | .6 | .3 | **-.237** | **-.264** | **-.147** | **-.124** | -.082 | -.085 |
| | | .5 | **-.203** | **-.175** | **-.108** | **-.111** | -.062 | -.044 |
| | | .7 | -.093 | -.079 | -.055 | -.053 | -.050 | -.061 |
| | | .9 | -.025 | .004 | -.015 | -.034 | -.020 | -.052 |

*Note*. $\rho_\alpha$ is the population alpha, $\rho_{ZY}$ is the population correlation between $Z$ and $Y$. $\pi$ is the selection ratio. $\rho_{ZY_T}$ is the population correlation between $Z$ and $Y_T$, $k$ is item number. Biases that are outside the nominal range of ±10% are presented in bold.

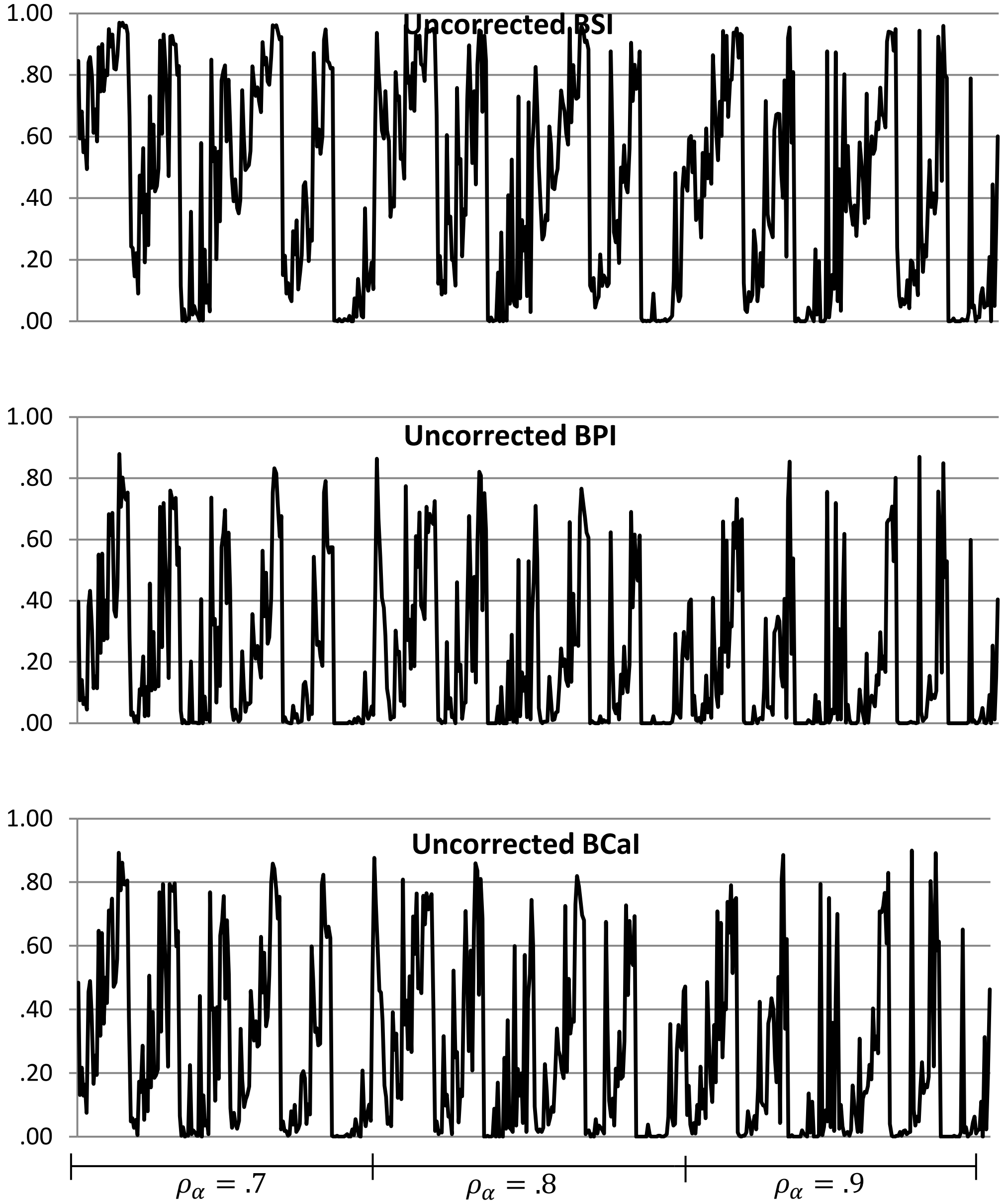
## Summary

The uncorrected alpha was found to be generally inaccurate. The largest percentage bias was found to be -99.7% (or alpha = .002) for continuous responses, and it was -136.2% (or alpha = -.185) for dichotomous responses; this reflects a very poor reliability but it is in fact due to a restricted sample. The bias became less problematic when the true alpha increased gradually and the selection ratio became less stringent. By contrast, the two corrected alphas performed desirably, and comparatively, the Case I correction procedure yielded slightly more accurate estimates.

## Results for Goal 2: Evaluating the Confidence Intervals in Single Study Continuous Responses

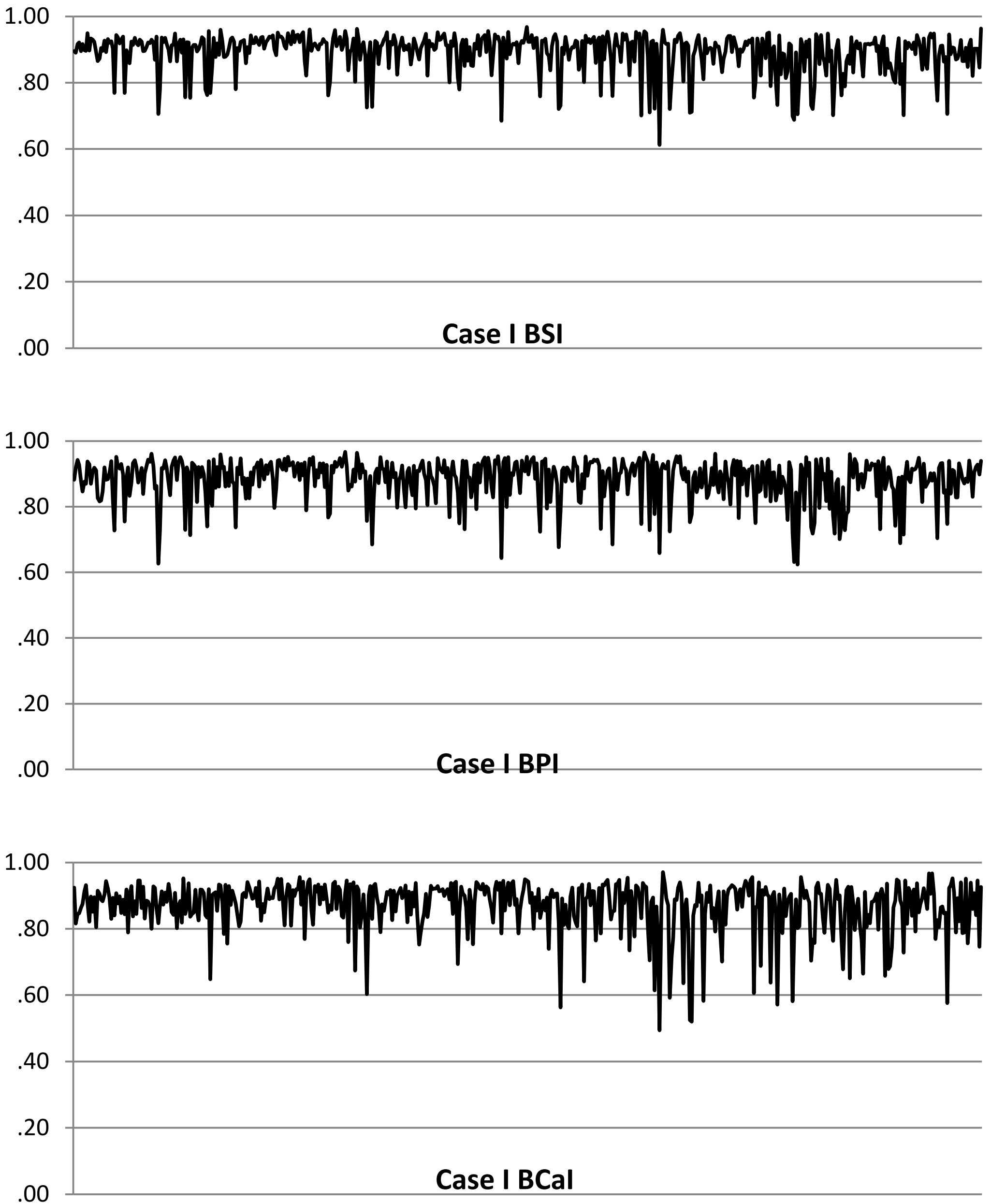
**Uncorrected CIs**. As shown in Figure 12, the bootstrap CIs surrounding the uncorrected alpha were generally inadequate due to an inaccurate point estimate of the uncorrected alpha. The mean coverage was .438 for BSI, .208 for BPI, and .252 for BCaI. The number of conditions that produced a coverage probability within the stringent criterion of [.922, .968] was 47 (or 8.7%) for BSI, and 0 for either BPI or BCaI. Even with the lenient cutoff of .855, the number of conditions became 78 (or 14.4%) for BSI, 3 (or .6%) for BPI, and 8 (or 1.5%) for BCaI.

*Figure 12.* Means of the 1,000 replicated coverage probabilities obtained by the uncorrected confidence intervals across 540 simulation conditions for continuous responses.



**Case I corrected CIs.** The bootstrap CIs surrounding the Case I corrected alpha appear to be more adequate than those built for the uncorrected alpha, as shown in Figure 13. For BSI, the coverage probabilities ranged from .613 to .968, with a mean of .892. For BPI, they ranged from .624 to .966, with a mean of .883. For BCaI, they ranged from .494 to .971, with a mean of .869. These results are similar to the average coverage probability (i.e., .89) yielded by the parametric CI surrounding Hunter et al.'s (2006) Case IV corrected correlation (Schmidt, personal communication, October 18th, 2010). The number of conditions that produced a coverage probability within the stringent criterion [.922, .968] was 201 (or 37.2%) for BSI, 171 (or 31.7%) for BPI, and 129 (or 23.9%) for BCaI. However, if we use a more lenient cutoff of .855, the results seem to behave more reasonably. For BSI, of the 540 conditions, 447 (or 82.8%) produced a coverage probability that exceeded this criterion. For BPI, 410 (or 75.9%) conditions passed this benchmark. For BCaI, 371 (or 68.7%) conditions were beyond this criterion. These findings are similar to Cui and Li's (2012) simulation, in which the BSI surrounding the alpha (when it was not subject to range restriction) achieved more adequate coverage probabilities than the BPI and BCaI. Note that the sample in that study was not subject to range restriction. Hence, the current findings add to the literature that the BSI is also accurate when it is used to resample a restricted sample and apply the correction procedure in order to construct the CI for the bias-corrected alpha.

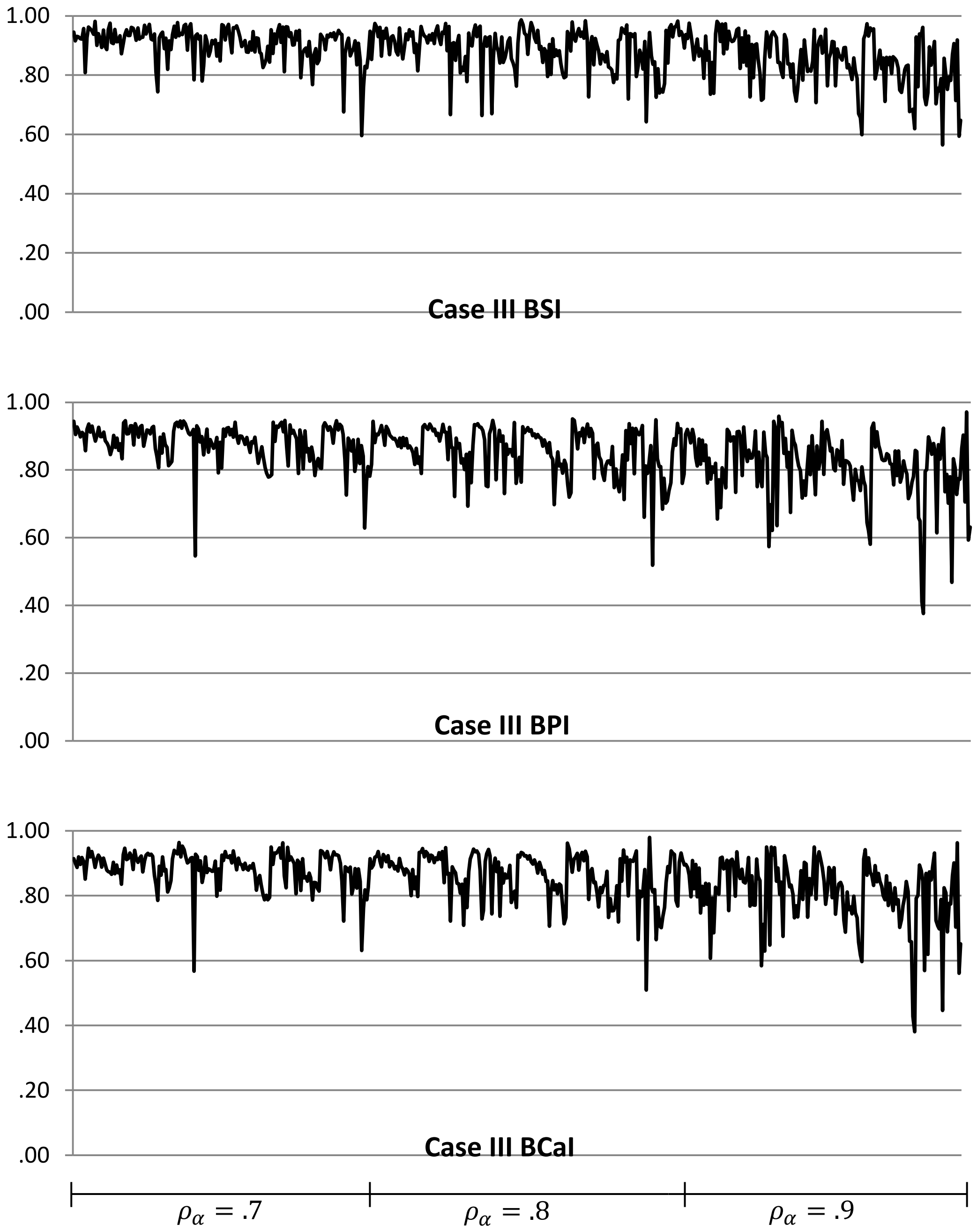
*Figure 13*. Means of the 1,000 replicated coverage probabilities obtained by the Case I corrected confidence intervals across 540 simulation conditions for continuous responses.



**Case III corrected CIs.** The bootstrap CIs built for the Case III corrected alpha also appeared to be more adequate than those for the uncorrected alpha;

however, their performances were slightly poorer than those of the CIs surrounding the Case I corrected alpha, as shown in Figure 14. The mean coverage probability was .883 for BSI, .854 for BPI, and .852 for BCaI. For BSI, of the 540 conditions, 193 (or 35.7%) resulted in a coverage probability within the stringent interval [.922, .968]. For BPI and BCaI, only 85 (or 15.7%) and 92 (or 17.0%) conditions, respectively, produced a coverage probability that fell inside this interval. Likewise, if we use a more lenient criterion (i.e., .855), BSI appears to be more adequate, with 390 (or 72.2%) conditions within this criterion. Moreover, the number of acceptable conditions becomes 321 (or 59.4%) for BPI, and 310 (or 57.4%) for BCaI.

*Figure 14.* Means of the 1,000 replicated coverage probabilities obtained by the Case III corrected alpha across 540 simulation conditions for continuous responses.



To summarize, the bootstrap CIs constructed for the uncorrected alpha were inappropriate because of the inaccurate point estimate of the alpha. By contrast, the bootstrap CIs surrounding the corrected alpha—especially the BSI for the Case I and III corrected alphas—were found to be more adequate across the simulation conditions. Details of these results are presented in Table 9.

Table 9. *Summary of the coverage probabilities yielded by different bootstrap CIs when items were continuously scored.*

| | | | | | # within | # beyond |
|---|---|---|---|---|---|---|
| CI | Alpha | Mean | SD | [Min, Max] | [.922, .968] | [.855] |
| BSI | UnCor | .438 | .330 | [0, .970] | 47(8.7%) | 78(14.4%) |
| | **Case I** | **.892** | **.058** | **[.613, .968]** | **201(37.2%)** | **447(82.8%)** |
| | **Case III** | **.883** | **.076** | **[.565, .986]** | **193(35.7%)** | **390(72.2%)** |
| BPI | UnCor | .208 | .253 | [0, .879] | 0(0%) | 3(0.6%) |
| | Case I | .883 | .064 | [.624, .966) | 171(31.7%) | 410(75.9%) |
| | Case III | .854 | .081 | [.376, .971] | 85(15.7%) | 321(59.4%) |
| BCaI | UnCor | .252 | .272 | [0, .899] | 0(0%) | 8(1.5%) |
| | Case I | .869 | .076 | [.494, .971] | 129(23.9%) | 371(68.7%) |
| | Case III | .852 | .083 | [.380, .979] | 92(17.0%) | 310(57.4%) |

*Note:* UnCor is uncorrected. # is number of conditions. The more adequate CIs for Case I and III are presented in bold.

Regarding the specific effects, some factors did not make an obvious impact on the coverage probability. Given that the Case I corrected BSI yielded the most accurate results, this section presents the findings based on this CI. Specifically, the restricted sample size and selection ratio did not show specific effects on the CI performance. The remaining factors only demonstrated minimal impact on the coverage probability, as shown in Table 10. Three key findings emerged. First, when the item number increased from 5 to 20, the coverage probability tended to increase gradually. This finding is reasonable because more items are expected to improve the precision and accuracy of the reliability when other factors are held constant. Second, when the alpha increased from .7 to .9, the coverage probability decreased slightly. This was probably due to the too-narrow confidence width for a large (or very precise) alpha value, which affected the probability of spanning the true population value. Third, when the correlations between *Z* and *Y* (or *Z* and $Y_T$ or both) increased, the coverage probabilities tended to decrease slightly; this was due to the more stringent or challenging selection procedure. Generally, the average coverage probabilities in each of the 36 aggregated conditions did not differ substantially.

Table 10. *Means of the 1,000 replicated* coverage *probabilities obtained by the Case I corrected BSI in 36 aggregated simulation conditions for continuous responses.*

| | | $k$ = 5 | | $k$ = 10 | | $k$ = 20 | |
|---|---|---|---|---|---|---|---|
| Alpha | $\rho_{ZY}$ | $\rho_{ZY_T} = .3$ | .6 | .3 | .6 | .3 | .6 |
| .70 | .30 | .893 | .897 | .900 | .881 | .915 | .916 |
| | .60 | .886 | .882 | .909 | .907 | .928 | .909 |
| .80 | .30 | .900 | .884 | .888 | .914 | .923 | .919 |
| | .60 | .849 | .846 | .914 | .893 | .931 | .917 |
| .90 | .30 | .871 | .859 | .847 | .885 | .887 | .905 |
| | .60 | .878 | .859 | .869 | .870 | .905 | .894 |

*Note*. $k$ is the item number. $\rho_{ZY}$ is the population correlation between *Z* and *Y*. $\rho_{ZY_T}$ is the population correlation between *Z* and $Y_T$.

**Confidence width**. The fluctuations of the widths were due to the factors evaluated in this dissertation. Specifically, a decrease in the width was due to a) an increased item number, b) an increased sample size, c) an increased selection ratio, and d) a larger coefficient alpha. These patterns remain the same for the dichotomous data, and hence they will not be repeated in the following sections. Regarding the uncorrected CIs, their confidence widths were generally wider than those of the corrected CIs, as shown in Table 11. For example, the widths ranged from .070 to 1.385 for the uncorrected BSI. The exceptionally wide widths were not meaningful in evaluating the sampling error of the alpha, and these results were mainly found in conditions with a small number of items, a small restricted sample size, and a very stringent selection ratio. By contrast, the Case I CIs produced the most precise (or narrowest) widths. For instance, the Case I corrected BSI produced coverage widths ranging from .013 to .179, with a mean of .057.

Table 11. *Summary of the confidence widths yielded by different bootstrap CIs for continuous responses.*

| | | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| BSI | Uncorrected | .420 | .239 | .070 | 1.385 |
| | **Case I** | **.057** | **.033** | **.013** | **.179** |
| | **Case III** | **.223** | **.158** | **.027** | **.928** |
| BPI | Uncorrected | .415 | .234 | .070 | 1.360 |
| | Case I | .059 | .033 | .013 | .178 |
| | Case III | .219 | .152 | .027 | .888 |
| BCaI | Uncorrected | .379 | .206 | .067 | 1.230 |
| | Case I | .059 | .035 | .013 | .191 |
| | Case III | .213 | .144 | .027 | .858 |

*Note*. The more adequate CIs for Case I and III are presented in bold.

### Dichotomous Responses

As in the results obtained from the continuous responses, the bootstrap CIs surrounding the uncorrected alpha were generally inadequate due to an inaccurate point estimate of the uncorrected alpha (see Figure 15). The mean coverage probability was .488 for BSI, .302 for BPI, and .337 for BCaI, and they were not desirable.  By contrast, the two corrected CIs were more accurate than the uncorrected CI. Comparatively, the Case III corrected BSI was the most accurate, as shown in Figures 16 and 17. The coverage probabilities ranged from .575 to .996, with a mean of .905. Of the 432 conditions, 353 conditions (or 81.7%)

were larger than the lenient criterion of .855. However, it may not be realistic for one to know the selection variable *Z* that causes range restriction on the test. The best alternative for the CI surrounding the Case I alpha was its BPI, with a mean coverage probability of .859, and a range of [.548, .997]. Of the 432 conditions, 259 (or 60.0%) produced a coverage probability beyond the lenient criterion. Details of the results for the other methods are presented in Table 12.

*Figure 15.* Means of the 1,000 replicated coverage probabilities obtained by the uncorrected confidence intervals across 432 simulation conditions for dichotomous responses.

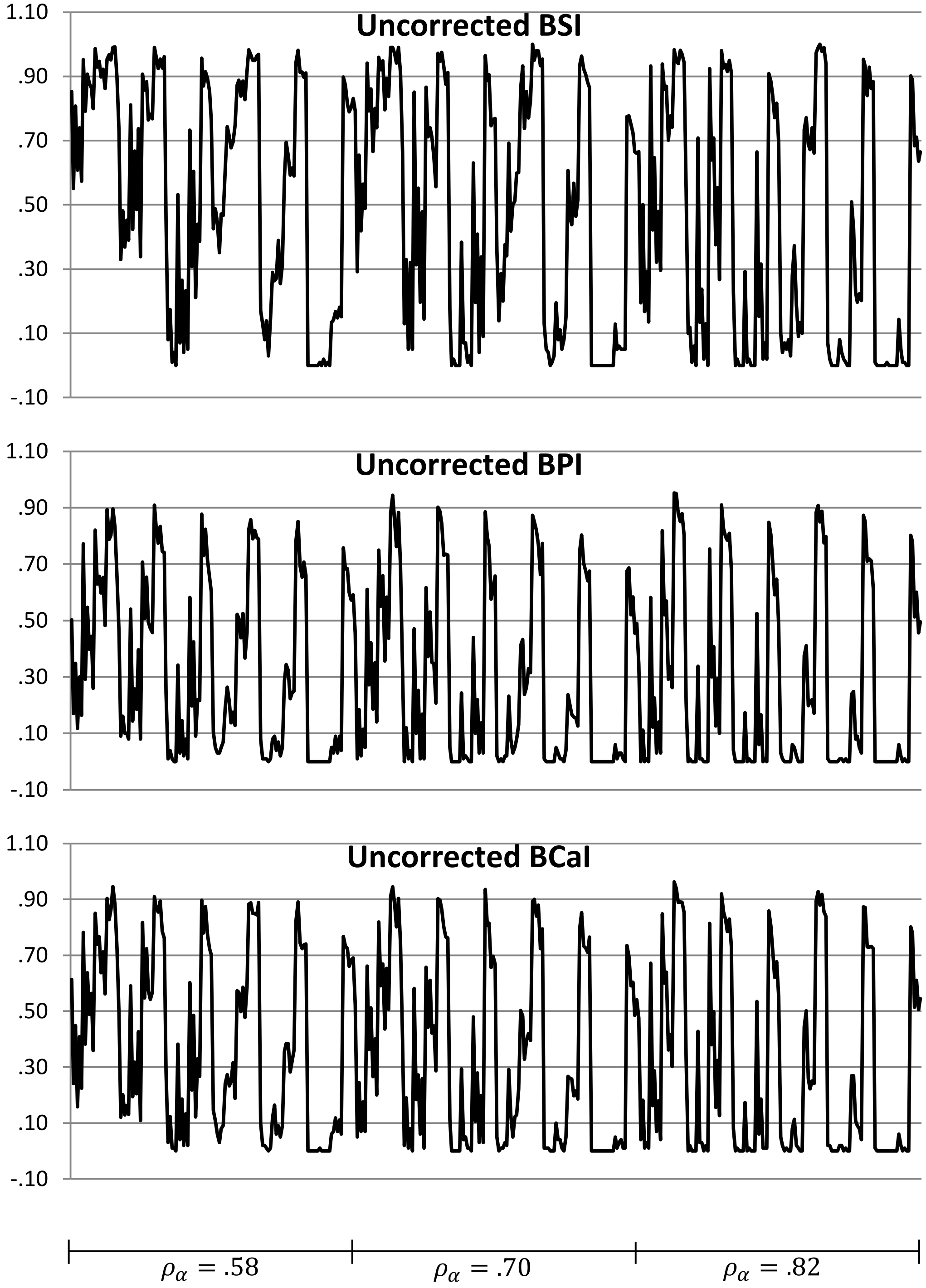


*Figure 16*. Means of the 1,000 replicated coverage probabilities obtained by the Case I corrected confidence intervals across 432 simulation conditions for dichotomous responses.

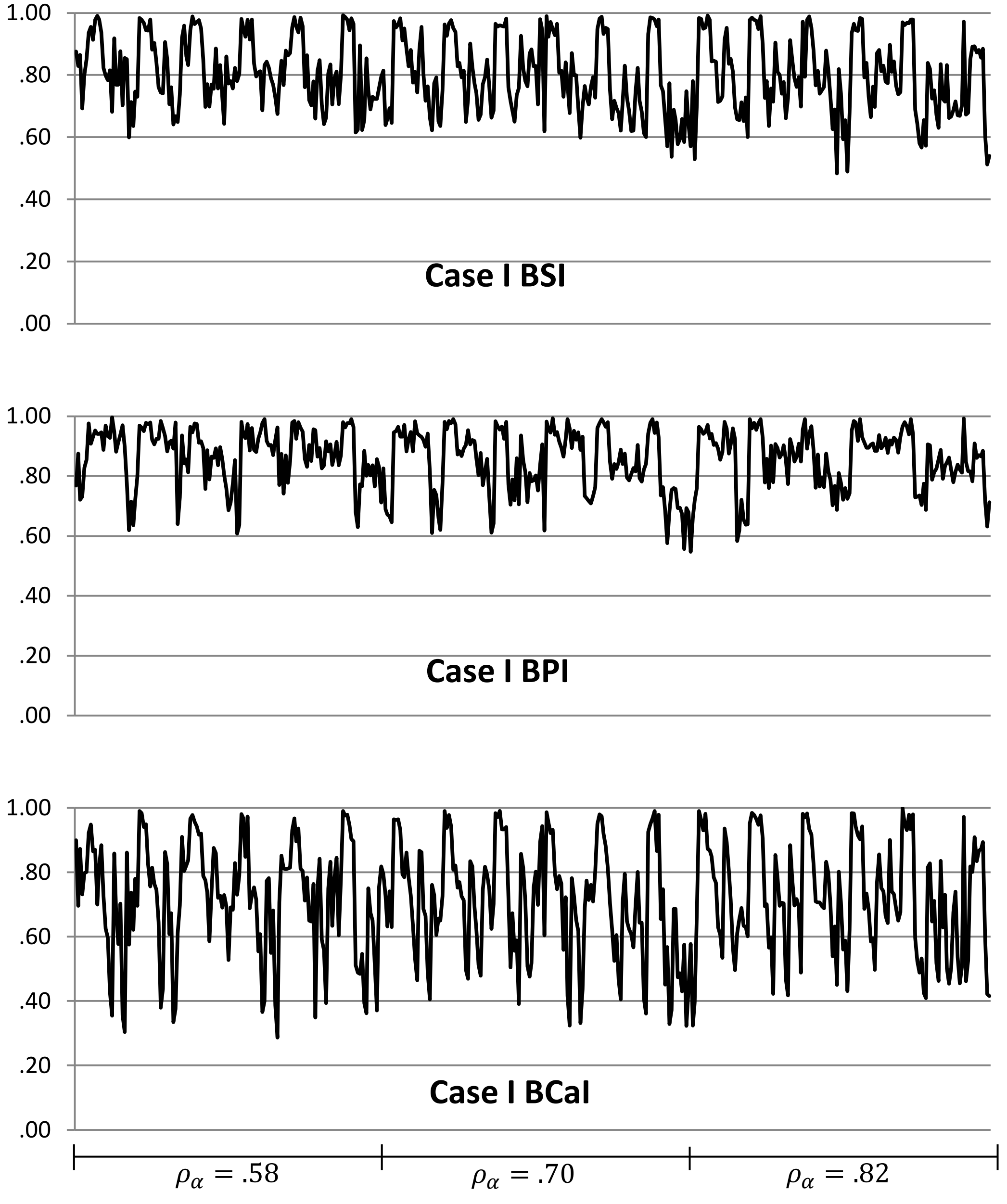


*Figure 17*. Means of the 1,000 replicated coverage probabilities obtained by the Case III corrected confidence intervals across 432 simulation conditions for dichotomous responses.

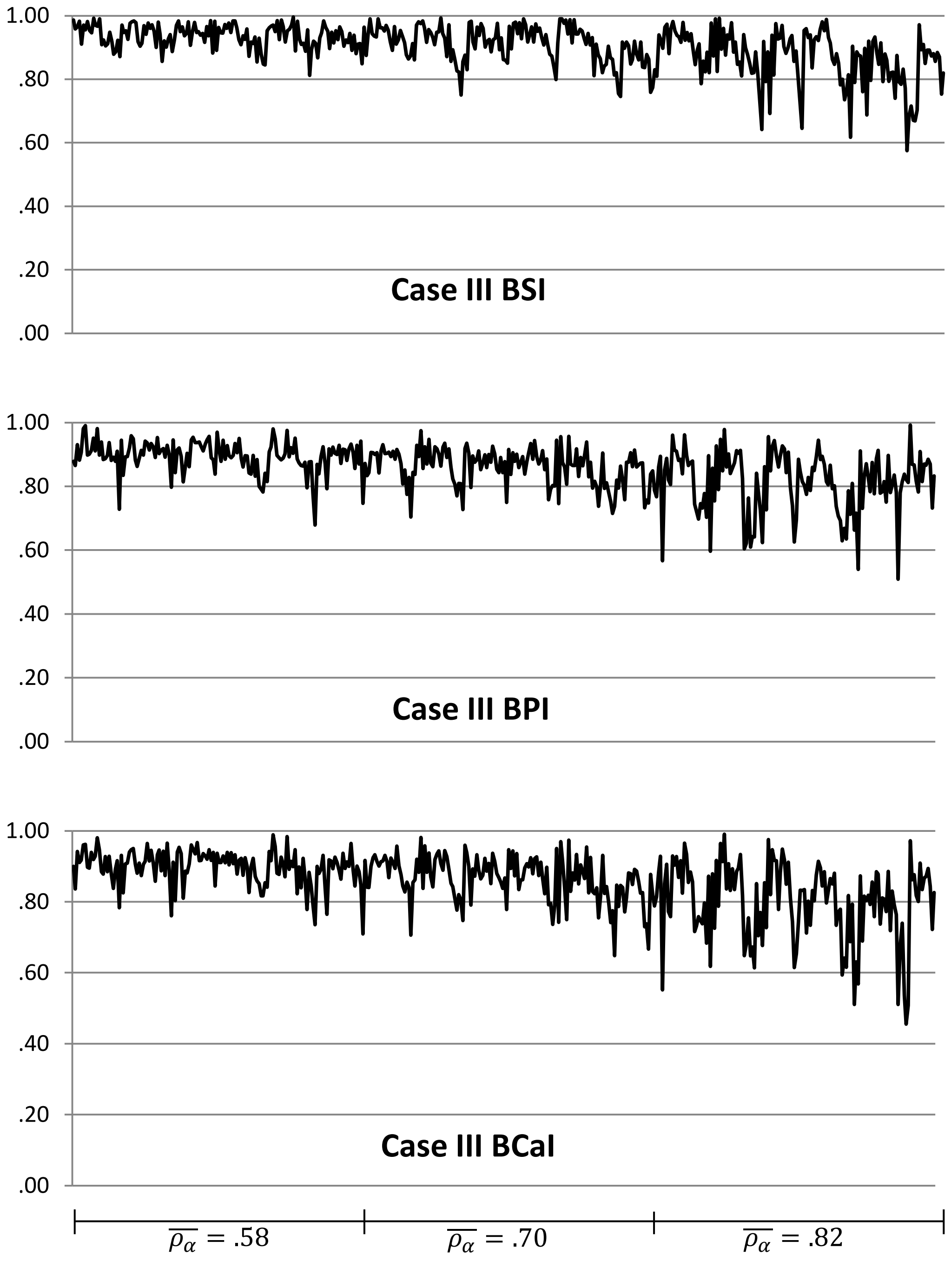


Table 12. *Summary of* c*overage probabilities yielded by different bootstrap CIs when items were dichotomously scored*

| | | | | | # within | # beyond |
|---|---|---|---|---|---|---|
| CI | Alpha | Mean | SD | [Min, Max] | [.922, .968] | [.855] |
| BSI | UnCor | .488 | .370 | [0, 1] | 46(10.6%) | 105(24.3%) |
| | Case I | .803 | .121 | [.484, .993] | 56(13.0%) | 147(34.0%) |
| | **Case III** | **.905** | **.069** | **[.575, .996]** | **140(32.4%)** | **353(81.7%)** |
| BPI | UnCor | .302 | .317 | [0, .953] | 3(0.7%) | 21(4.9%) |
| | **Case I** | **.859** | **.101** | **[.548, .997]** | **93(32.4%)** | **259(60.0%)** |
| | Case III | .857 | .078 | [.509, .992] | 60(13.9%) | 275(63.7%) |
| BCaI | UnCor | .337 | .330 | [0, .963] | 6(1.4%) | 39(9.0%) |
| | Case I | .718 | .177 | [.286, 1] | 41(9.5%) | 111(25.7%) |
| | Case III | .855 | .088 | [.455, .990] | 82(19.0%) | 259(60.0%) |

*Note:* UnCor is uncorrected. # is number of conditions. The results of the more adequate CI for Case I and III, respectively, are presented in bold.

Note that data categorization tended to decrease the accuracy of the bootstrap CIs, especially for the Case I corrected alpha. One possible reason is that the Case I correction depends on the accuracy of the ratio of the restricted to unrestricted SD, but data categorization may lead to a small and unstable restricted SD. As shown in Table 13, the coverage probability became smaller with a more stringent selection ratio, which is regarded as the most influential factor on CI performance for the Case I corrected alpha. For Case III, the selection ratio had an adverse effect on the coverage probability when the true alpha value was large (i.e., .82). To summarize, a more stringent selection ratio decreases the accuracy of the Case I corrected CIs; however, the Case III corrected CI is relatively robust to this factor when the true alpha value is small to moderate.

Table 13. *Means of the 1,000 replicated coverage probabilities obtained by the Case I and III corrected BSI in 32 selected conditions for dichotomous responses.*

| | | | $k = 5$ | | | |
|---|---|---|---|---|---|---|
| | | | $\rho_{SY_T} = .3$ | | $\rho_{SY_T} = .6$ | |
| Alpha | $\rho_{SY}$ | $\pi$ | Case I | Case III | Case I | Case III |
| .58 | .3 | .3 | **.768** | .981 | **.845** | .959 |
| | | .5 | .976 | .956 | .908 | .982 |
| | | .7 | .946 | .967 | .887 | .990 |
| | | .9 | .947 | .948 | .881 | .919 |
| | .6 | .3 | **.686** | .963 | **.712** | .946 |
| | | .5 | .981 | .947 | .939 | .984 |
| | | .7 | .887 | .937 | .880 | .925 |
| | | .9 | .917 | .933 | .902 | .855 |
| .82 | .3 | .3 | **.694** | **.830** | **.679** | **.809** |
| | | .5 | .965 | .973 | .957 | .881 |
| | | .7 | .941 | .934 | .904 | .916 |
| | | .9 | .981 | .943 | .958 | .892 |
| | .6 | .3 | **.810** | **.718** | **.777** | **.646** |
| | | .5 | .952 | .855 | .985 | .925 |
| | | .7 | .912 | .952 | .894 | .989 |
| | | .9 | .884 | .880 | .936 | .851 |

*Note*. $k$ is the item number. $\rho_{ZY}$ is the population correlation between $Z$ and $Y$. $\rho_{ZY_T}$ is the population correlation between $Z$ and $Y_T$, $\pi$ is the selection ratio. Coverage probabilities that are smaller than the lenient criterion of .855 are presented in bold.

**Confidence width**. As in the continuous data, similar patterns of confidence widths were obtained for the dichotomous data, as shown in Table 14. Regarding the uncorrected CIs, their confidence widths were generally wider than the corrected CIs, as shown in Table 7. For example, the widths ranged from .105 to 1.423 for the uncorrected BSI. The exceptionally wide widths were not meaningful in evaluating the sampling error of the alpha. By contrast, the Case I CIs produced the most precise (or narrowest) widths. For instance, the Case I corrected BPI produced a mean width of .070, and they ranged from .018 to .190.

Table 14. *Summary of the confidence widths yielded by different bootstrap CIs when items were dichotomously scored.*

| | | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| BSI | Uncorrected | .522 | .262 | .105 | 1.423 |
| | Case I | .070 | .042 | .018 | .190 |
| | **Case III** | **.351** | **.203** | **.050** | **1.133** |
| BPI | Uncorrected | .519 | .258 | .105 | 1.376 |
| | **Case I** | **.070** | **.041** | **.018** | **.190** |
| | Case III | .349 | .199 | .050 | 1.091 |
| BCaI | Uncorrected | .484 | .232 | .102 | 1.282 |
| | Case I | .067 | .041 | .018 | .186 |
| | Case III | .344 | .195 | .050 | 1.093 |

*Note*. The more adequate CIs for Case I and III are presented in bold.

## Summary

The BSIs constructed for the Case I and III corrected alphas appeared to be reasonable when the items were continuously scored. The mean coverage probability was .892 for Case I, and it was .883 for Case III. These results were better than the bootstrap CIs surrounding the uncorrected alpha with a maximum mean coverage probability of .438 only yielded by the BSI. When the items were dichotomously scored, the Case III BSI and Case I BPI seemed to be reasonable.

## Monte Carlo 2 – Meta-Analysis

This section presents the results for the uncorrected and corrected coefficient alphas, as well as for their CIs, in a meta-analytic study. These results seek to evaluate biases that come from the uncorrected alpha, and to examine whether or not the two corrected alphas and their associated CIs can improve the accuracy of these measures. As in Study 1, given that the patterns of the results did not differ across the three data conditions (i.e., parallel, tau-equivalent, and congeneric data), the results based on the most relaxed congeneric condition are discussed. Thus, a total of 480 conditions (i.e., 2 levels of item number by 2 levels of restricted sample size by 5 levels of selection ratio by 3 levels of alpha by 2 levels of correlation between $Z$ and $Y_T$ by 2 levels of correlation between $Z$ and $Y$ by 2 levels of number of studies) are presented in the following sections. Moreover, the findings are discussed based on the continuous and dichotomous responses, respectively.

## Results for Goal 3: Evaluating the Mean Alpha in Meta-Analysis

### Continuous Responses

Figure 18 shows the means of the 1,000 replicated percentage biases yielded by the uncorrected and corrected mean alphas across 480 simulation conditions. As in Study 1, the biases of the uncorrected mean alpha were generally undesirable. They ranged from -86.4% to -3.7%, with a mean of -26.6%. Of the

480 conditions, only 103 (or 21.5%) were within the lenient level of ±10%. By contrast, the two corrected mean alphas were highly accurate, and their biases were close to zero. Specifically, the Case I corrected mean alpha yielded a mean bias of -0.4%, with a range of -1.2% to 0.1%. All of the 480 conditions yielded a bias that fell within the stringent level of ±5%. The overall MAPE was 0.4%, which was much smaller than the nominal level of 10%. The Case III corrected mean alpha also achieved highly accurate results, and all of the 480 conditions produced a percentage bias within the stringent level of ±5%. These biases ranged from -1.2% to 0.7%, with a mean of -0.3%. The overall MAPE was 0.4%, which was smaller than the nominal level of 10%. Details of these results are presented in Table 16.

*Figure 18*. Means of the 1,000 replicated alphas and percentage biases obtained by the uncorrected and corrected alpha in meta-analysis across 480 simulation conditions for continuous responses.

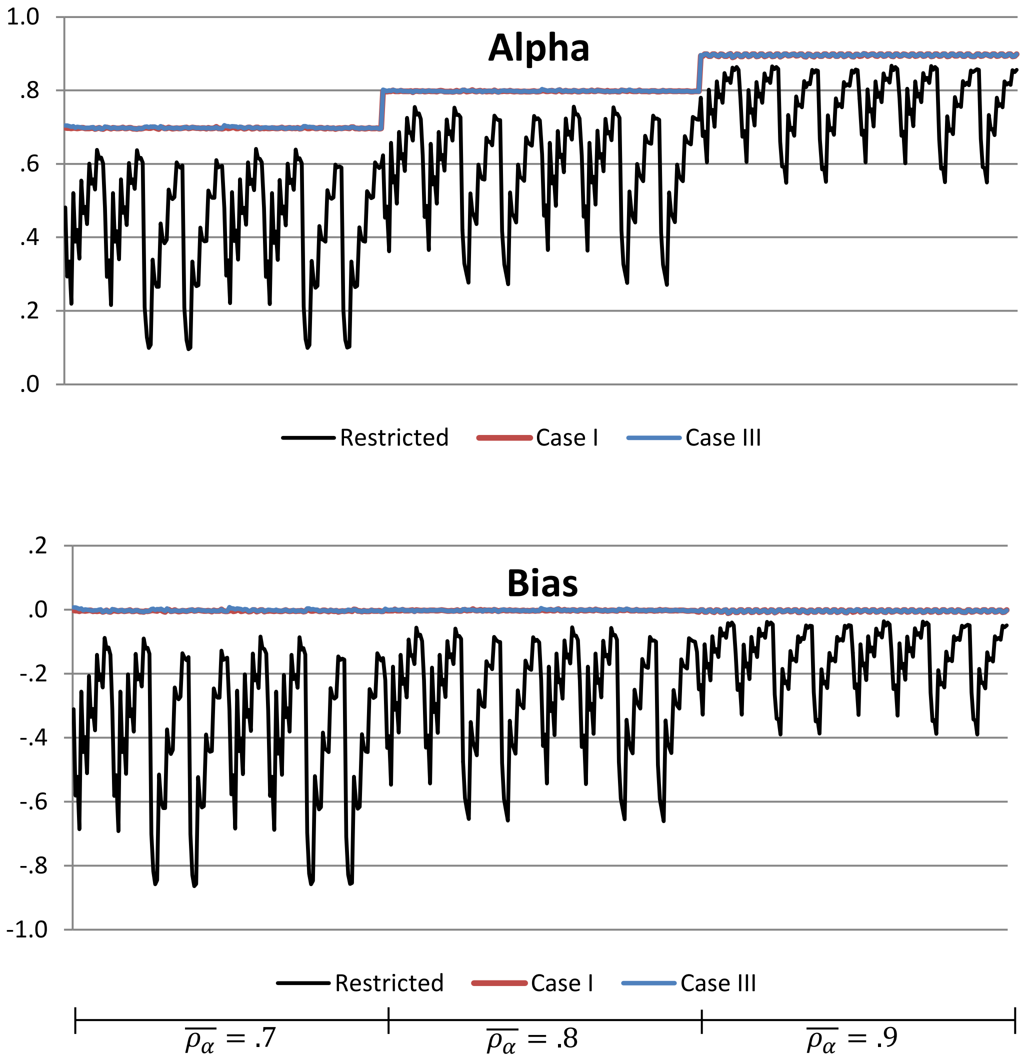


Table 15. *Summary of the percentage biases obtained by the uncorrected and corrected mean alphas in meta-analysis for continuous responses.*

| Alpha | Mean<br>Bias | SD | [Max, Min] | MAPE | # within<br>±5% | # within<br>±10% |
|---|---|---|---|---|---|---|
| UnCor | -26.6% | 19.0% | [-86.4%, -3.7%] | 13.5% | 24(5%) | 103(21.5%) |
| Case I | -0.4% | 0.2% | [-1.2%, 0.1%] | 0.4% | 480(100%) | 480(100%) |
| Case III | -0.3% | 0.3% | [-1.2%, 0.7%] | 0.4% | 480(100%) | 480(100%) |

*Note*: UnCor is uncorrected. # is number of conditions.

Regarding the specific effects of each factor on the biases, given that the restricted sample size and the number of studies did not show any obvious relationships with the uncorrected alpha, the results are discussed based on restricted sample size = 100 and number of studies = 15, as shown in Table 16. Moreover, the patterns of relationships repeated for increasing levels of alpha, and thus the results are fixed at alpha = .7. Three key findings were obtained. First, when the selection ratio became more stringent (or smaller), the uncorrected mean alpha became smaller (i.e., more downward biased). As in a single study, a more stringent selection process led to a more homogenous sample, which could not represent the scores in the original unrestricted sample appropriately. Second, when the correlations between $Z$ and $Y$ (or $Z$ and $Y_T$ or both) became stronger, the uncorrected mean alpha became smaller. This was due to the stronger range-restriction effect from $Z$ to the true and observed scores, and hence a more homogenous sample. Third, the downward bias became slightly more severe with an increasing number of items. This result was unexpected at first because an increase in this number might help improve the accuracy of reliability evaluation.

However, upon further examination, one may find that the item correlation matrix tends to be more adversely affected by range restriction for twenty items than five items, when the alpha is held constant. For instance, when the alpha value was .70, selection ratio was .10, stringent correlations (i.e., .60) between $Z$ and $Y$ (and $Y_T$), restricted sample size was 100, and item number was fixed at 5, the item correlations could range from .24 to .39 for an unrestricted sample, and they ranged from -.05 to .33 for a restricted sample. By contrast, given the same conditions except with the item number equaled 20, the item correlations ranged from .09 to .33 for an unrestricted sample, and they ranged from -.27 to .27 for a restricted sample. In this sense, the low item correlations in a longer test (e.g., from .09 to -.27) were more adversely affected by range restriction than those in a shorter test (e.g., from .24 to -.05). This resulted in a smaller alpha value for a long test than a short test, when a sample is subject to range restriction. This situation seemed to be more problematic in a meta-analytic study than a single study, given that the item-correlation matrix in a long test in each single study can have a chance to be much reduced. Further examination is needed to investigate this pattern.

Table 16. *Means of the 1,000 replicated uncorrected mean alpha values across 40 selected simulation conditions when the population alpha = .7, restricted sample size = 100, and number of study = 15 for continuous responses.*

| | | $k = 5$ | | $k = 20$ | |
|---|---|---|---|---|---|
| $\rho_{ZY}$ | $\pi$ | $\rho_{ZY_T} = .3$ | .6 | $\rho_{ZY_T} = .3$ | .6 |
| .3 | .1 | .482 | .293 | .334 | .219 |
| | .3 | .520 | .388 | .421 | .342 |
| | .5 | .555 | .465 | .483 | .436 |
| | .7 | .601 | .546 | .558 | .530 |
| | .9 | .638 | .614 | .617 | .601 |
| .6 | .1 | .209 | .128 | .099 | .108 |
| | .3 | .339 | .283 | .266 | .265 |
| | .5 | .438 | .396 | .384 | .393 |
| | .7 | .530 | .510 | .504 | .509 |
| | .9 | .604 | .594 | .589 | .596 |

*Note*. $k$ is the item number. $\rho_{ZY}$ is the population correlation between $Z$ and $Y$. $\rho_{ZY_T}$ is the population correlation between $Z$ and $Y_T$, $\pi$ is the selection ratio.

**Dichotomous Responses**

As in Study 1, the selection ratio of .10 was too challenging for the execution of the simulation study, so this level was dropped, resulting in 384 simulation conditions. Given that the patterns for the uncorrected and corrected mean alphas across 384 simulation conditions were similar to those for the continuous data (i.e., Figure 18), the results obtained from the dichotomous scores can be simplified and summarized in Table 17. The uncorrected mean alpha estimates, likewise, were inaccurate. The percentage biases ranged from -93.65% to 4.0%, with a mean of -37.1%. The overall absolute bias (MAPE = 37.1%) was also inappropriate outside the nominal level of 10%. Comparing the two corrected cases, the Case I mean alpha yielded slightly better results. The percentage biases ranged from -8.0% to 11.8%, with a mean of 1.8%. The overall absolute MAPE was 3.7%, which was within the nominal level of 10%. Of the 384 conditions, 308 (or 80.2%) were within the stringent nominal level of ±5%. The Case III mean alpha also produced good results, with the percentage biases ranging from -34.3% to 4.7%, and the mean was -5.9%. The overall absolute bias MAPE was 6.1%, which was also inside the nominal level of 10%. Of the 384 conditions, 231 (or 60.2%) were within the stringent nominal level of ±5%.

Table 17. *The performance of the uncorrected and corrected mean alpha estimates in meta-analysis for dichotomous responses.*

| Alpha | Mean<br>Bias | SD | [Max, Min] | MAPE | # within<br>±5% | # within<br>±10% |
|---|---|---|---|---|---|---|
| UnCor | -37.1% | 22.4% | [-93.6%, -4.0%] | 37.1% | 4(1.0%) | 36(9.4%) |
| Case I | 1.8% | 0.6% | [-8.0%, 11.8%] | 3.7% | 308(80.2%) | 348(90.6%) |
| Case III | -5.9% | 1.7% | [-34.3%, 4.7%] | 6.1% | 231(60.2%) | 311(81.0%) |

*Note*: UnCor is uncorrected. # is number of conditions.

When evaluating the impact of each factor on the alpha estimate, the selection ratio was found to be the most influential factor, as shown in Table 18. As predicted, increasing the selection ratio decreased the restricted alpha value. In addition, when the correlation between the selection construct $Z$ and $Y_T$ became stronger, the downward bias became more severe. This can be explained by a stronger relationship between the selection variable and the variable of interest should lead to a more homogeneous sample, leading to a smaller mean alpha estimate. Regarding the performance of the Case I and III corrected mean alphas, both of them became more accurate when the selection ratio became less stringent or the correlation between $Z$ and $Y_T$ became weaker, or both.

Table 18. *Means of 1,000 replicated percentage biases of the uncorrected and corrected mean alphas for dichotomous data with number of studies fixed at 15, alpha = .7, item number = 5, and mean restricted sample size of 100 for dichotomous responses.*

| Alpha | Ratio | $\rho_{ZY_T} = .3$ | $\rho_{ZY_T} = .6$ |
|---|---|---|---|
| Uncorrected | .3 | **-.441** | **-.789** |
| | .5 | **-.313** | **-.559** |
| | .7 | **-.235** | **-.315** |
| | .9 | -.097 | **-.143** |
| Case I | .3 | **.106** | **.127** |
| | .5 | .034 | .035 |
| | .7 | -.040 | -.011 |
| | .9 | -.013 | -.012 |
| Case III | .3 | **-.126** | **-.230** |
| | .5 | -.045 | -.107 |
| | .7 | -.045 | -.027 |
| | .9 | .009 | .014 |

*Note*. $\rho_{ZY_T}$ is the population correlation between $Z$ and $Y_T$. Biases that are outside the nominal range of $\pm10\%$ are presented in bold.

**Summary**

The uncorrected mean alpha was generally undesirable. The largest percentage bias was found to be -86.4% for continuous responses, and it was

-93.6% for dichotomous responses. The bias became less problematic when the true alpha increased gradually and the selection ratio became less stringent. Moreover, it tended to be slightly smaller when the number of items decreased from 20 to 5; this pattern needs further examination. By contrast, the two corrected alphas were accurate, and the Case I correction procedure yielded slightly better results when the items were dichotomously scored.

## Results for Goal 4: Evaluating the Confidence Intervals in Meta-Analysis

### Continuous Responses

**Coverage probability**. The bootstrap CIs surrounding the Case I and III corrected mean alphas appeared to be more adequate than the non-parametric CIs, as shown in Figures 19 and 20. For Case I, the BSI appeared to be more adequate. The coverage probabilities yielded by the Case I BSI ranged from .702 to .970, with a mean of .897. Of the 480 conditions, 189 (or 39.4%) were within the stringent criterion of [.922, .968], and 332 (or 69.2%) were larger than the lenient level of .855. For Case III, the BSI produced the coverage probabilities ranging from .732 to .989, with a mean of .893. Of the 480 conditions, 192 (or 40.0%) were within the stringent criterion, and 390 (or 81.3%) were higher than the lenient criterion. Details of the results yielded by the other methods are presented in Table 19.

*Figure 19*. Means of the 1,000 replicated coverage probabilities yielded by the Case I bootstrap CIs across 480 conditions for continuous responses.

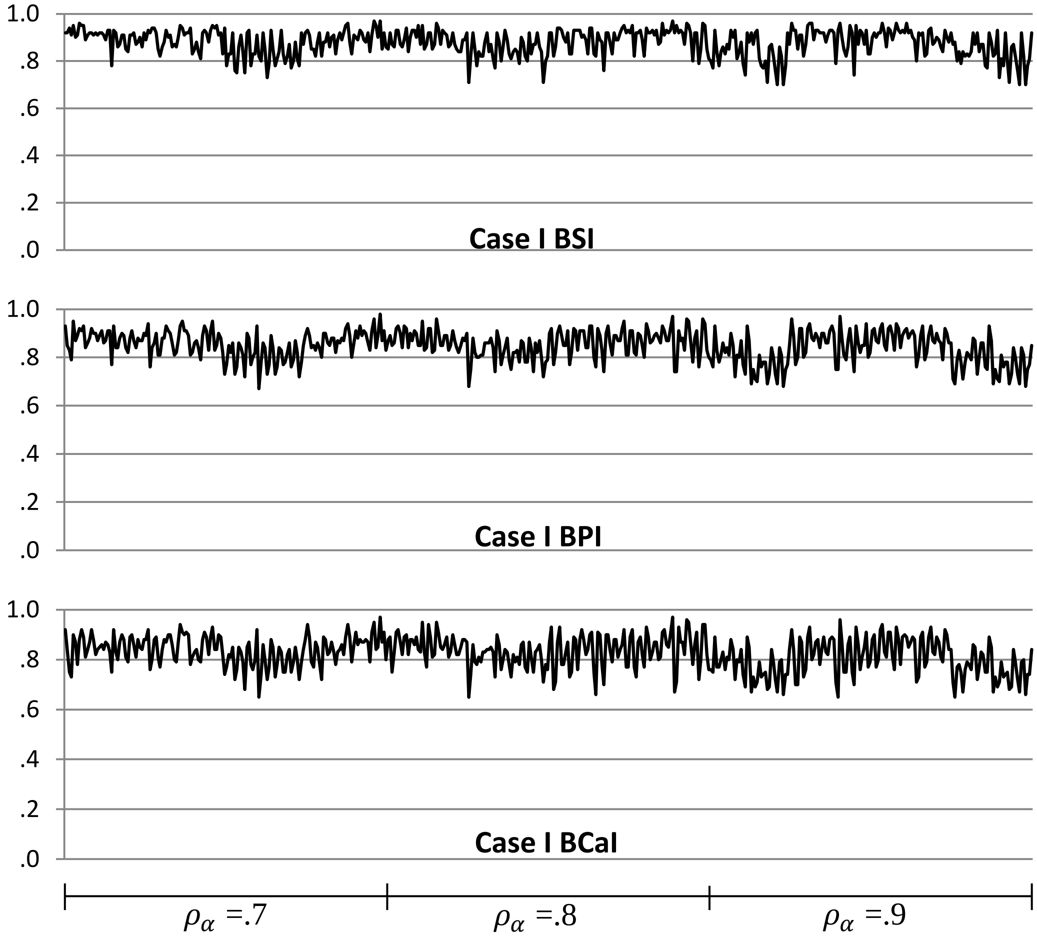


*Figure 20.* Means of the 1,000 replicated coverage probabilities yielded by the Case III bootstrap CIs across 480 conditions for continuous responses.

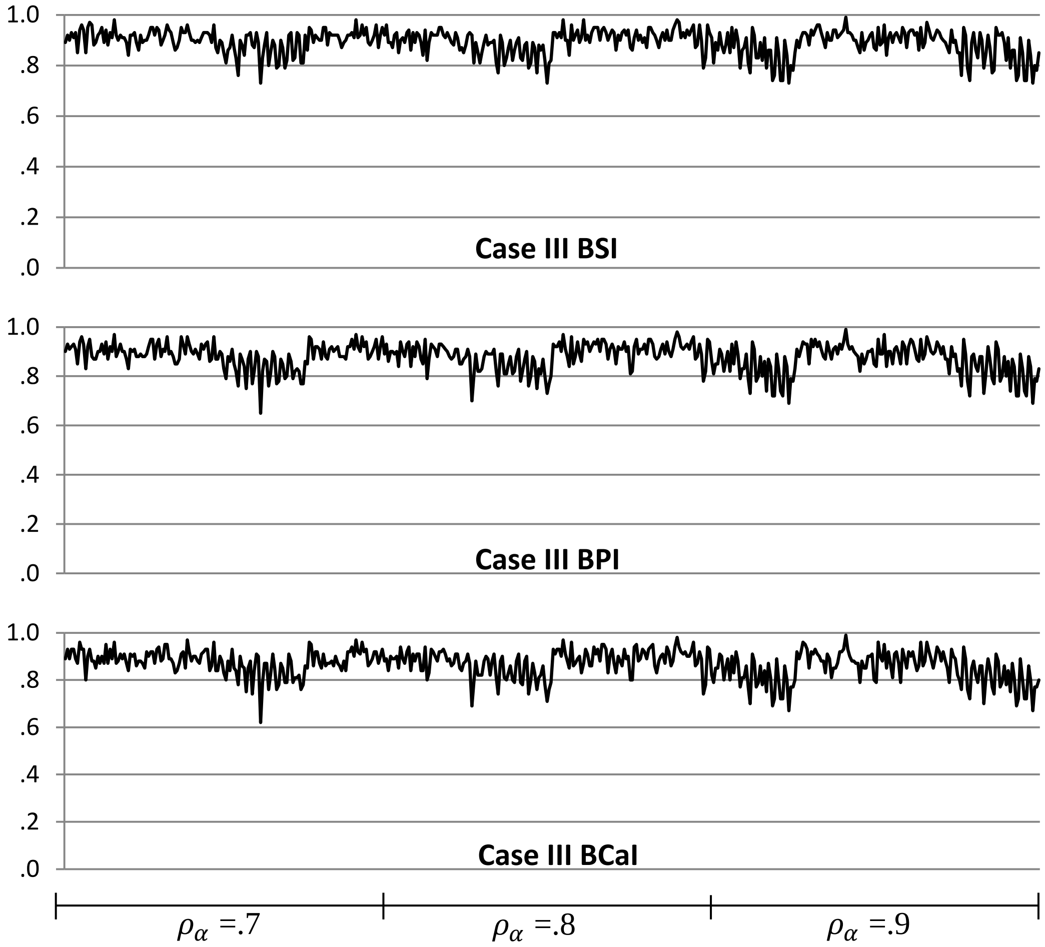


Table 19. *Means of 1,000 replicated coverage probabilities of the CIs surrounding the uncorrected and corrected mean alphas in meta-analysis for continuous responses.*

| | | Mean | SD | [Min, Max] | # within<br>[.922, .968] | # beyond<br>[.855] |
|---|---|---|---|---|---|---|
| BSI | UnCor | .005 | .024 | [0, .320] | 0(0%) | 0(0%) |
| | **Case I** | **.897** | **.150** | **[.702,.970]** | **192(40.0%)** | **332(69.2%)** |
| | **Case III** | **.893** | **.110** | **[.732, .985]** | **192(40.0%)** | **390(81.3%)** |
| BPI | UnCor | .003 | .018 | [0, .260] | 0(0%) | 0(0%) |
| | Case I | .851 | .148 | [.660, .980] | 89(18.5%) | 255(53.1%) |
| | Case III | .879 | .111 | [.651, .989] | 136(28.3%) | 353(73.5%) |
| BCaI | UnCor | .003 | .016 | [0, .230] | 0(0%) | 0(0%) |
| | Case I | .828 | .147 | [.651,.971] | 50(16.4%) | 194(40.4%) |
| | Case III | .868 | .111 | [.623, .987] | 107(22.3%) | 320(66.7%) |

*Note*: UnCor is uncorrected. # is number of conditions. The more adequate CI for Case I and III are presented in bold.

To summarize, the BSIs surrounding the Case I and III corrected mean alphas appeared to be more adequate than other methods, given that their mean coverage probabilities (i.e., .897 and .893) were closer to the ideal value (i.e., .95), and the percentage of conditions within the lenient criterion were acceptable (i.e., 69.2% and 81.3%). Moreover, these methods outperformed the CIs surrounding the uncorrected mean alpha with 0% of the conditions that fell within the criterion. These results are also comparable to a Monte Carlo study on the performance of the CIs surrounding the corrected mean correlation (Li et al., in press). In that study, the percentage of conditions that fell within the stringent criterion was 45.4% for the BSI, whereas these values are 40.0% for both the Case I and III BSIs in the present study. If we evaluate the performance based on a more lenient criterion, both BSIs seem to be more reasonable, with 69.2% acceptable conditions for Case I and 81.3% for Case III.

**Confidence width.** Given that a meta-analysis involves a much larger sample size than a single study, the sampling error and the confidence width are expected to be narrower, and this may affect the coverage of the true reliability. As shown in Table 20, the Case I BSI produced a mean width of .0027, with a range (.0005, .0060). The Case III BSI yielded a mean width of .0032, with a range (.0008, .0082). Because the Case I and III BSIs produced reasonable coverage probabilities of the true alpha values, these widths provided highly precise evaluation of the sampling error of the mean alpha level of a test in a research domain. Regarding the specific effects, as in Study 1, a decrease in the width was due to a) an increased item number, b) an increased sample size, c) an increased selection ratio, and d) a larger coefficient alpha.

Table 20. *Summary of the confidence widths of the CIs surrounding the uncorrected and corrected mean alphas for continuous responses.*

| | | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| BSI | Uncorrected | .0073 | .0027 | .0023 | .0176 |
| | **Case I** | **.0026** | **.0012** | **.0005** | **.0060** |
| | **Case III** | **.0032** | **.0015** | **.0008** | **.0082** |
| BPI | Uncorrected | .0073 | .0027 | .0023 | .0176 |
| | Case I | .0026 | .0012 | .0005 | .0060 |
| | Case III | .0032 | .0015 | .0008 | .0082 |
| BCaI | Uncorrected | .0744 | .0279 | .0227 | .1817 |
| | Case I | .0259 | .0124 | .0048 | .0613 |
| | Case III | .0327 | .0153 | .0079 | .0839 |

*Note:* The more adequate CI for Case I and III are presented in bold.

## Dichotomous Responses

**Coverage probability**. As in the continuous responses, the uncorrected mean alpha estimates produced very poor results, as shown in Table 21. Generally, the bootstrap CIs surrounding the uncorrected mean alpha led to a zero percentage of the conditions within the nominal range of [.922, .968]. These findings were even worse than those in the continuous data, given that the dichotomous data led to even narrower widths, which adversely affected the coverage of the true reliability. By contrast, the bootstrap CIs surrounding the Case I and III corrected mean alphas were better than the uncorrected CIs; however, they were less accurate than those found in the continuous data. The mean coverage probabilities were .816 for the Case I BSI, and .805 for the Case III BSI. The proportions of conditions within the nominal range of [.922, .968] were 27.1% and 26.6%, respectively; they became 66.4% and 67.2% when the lenient cutoff of .855 was used. As noted above, this was due to an extremely narrow width that came from the dichotomous scores and the large sample size in meta-analysis; this will be discussed in the next section.

Table 21. *Coverage probabilities of the CIs surrounding the uncorrected and corrected mean alphas in meta-analysis for dichotomous responses.*

| | | Mean | SD | [Min, Max] | # within<br>[.922, .968] | # beyond<br>[.855] |
|---|---|---|---|---|---|---|
| BSI | UnCor | .003 | .012 | [0, .531] | 0(0%) | 0(0%) |
| | **Case I** | **.816** | **.108** | **[.210, .941]** | **104(27.1%)** | **255(66.4%)** |
| | **Case III** | **.805** | **.126** | **[.203, .944]** | **102(26.6%)** | **258(67.2%)** |
| BPI | UnCor | .004 | .045 | [0, .532] | 0(0%) | 0(0%) |
| | Case I | .693 | .113 | [.202, .937] | 93(24.2%) | 215(56.0%) |
| | Case III | .682 | .128 | [.213, .938] | 88(22.9%) | 217(56.5%) |
| BCaI | UnCor | .004 | .032 | [0, .530] | 0(0%) | 0(0%) |
| | Case I | .660 | .112 | [.252, .867] | 25(6.5%) | 48(12.5%) |
| | Case III | .673 | .111 | [.221, .871] | 12(3.1%) | 50(13.0%) |

*Note*: UnCor is uncorrected. # is number of conditions. The more adequate CI for Case I and III are presented in bold.

**Confidence width.** The relatively poor coverage probabilities can be explained in part by the associated confidence widths in meta-analysis. Again, the width of the uncorrected mean alpha was wider than the corrected mean alpha. The non-parametric bootstrap CIs surrounding the Case I and III corrected mean alphas produced the narrowest widths. Given that the BSIs for the Case I and III corrected mean alphas yielded acceptable coverage probabilities (see Table 22), these narrow confidence widths appeared to provide precise evaluation of the sampling error of the mean alpha in a meta-analytic study. However, one still needs to note that these results were still less desirable than those obtained from the continuous responses, and their adequacy still needs further examination.

Table 22. *Confidence widths yielded by the CIs surrounding the uncorrected and corrected mean alphas in meta-analysis for dichotomous responses.*

| | | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| BSI | Uncorrected | .0034 | .0012 | .0000 | .0075 |
| | **Case I** | **.0014** | **.0006** | **.0005** | **.0028** |
| | **Case III** | **.0018** | **.0007** | **.0005** | **.0036** |
| BPI | Uncorrected | .0035 | .0011 | .0015 | .0075 |
| | Case I | .0014 | .0006 | .0005 | .0027 |
| | Case III | .0018 | .0007 | .0005 | .0036 |
| BCaI | Uncorrected | .1182 | .0368 | .0494 | .2565 |
| | Case I | .0462 | .0193 | .0156 | .0913 |
| | Case III | .0609 | .0252 | .0184 | .1238 |

*Note*: The more adequate CI for Case I and III are presented in bold.

### Summary

The bootstrap CIs, especially the BSI, constructed for the corrected mean alpha were more reasonable than those built for the uncorrected mean alpha. In fact, the coverage probability was highly inaccurate for the CIs surrounding the uncorrected mean alpha, and the number of conditions within the lenient criterion was 0%. On the other hand, the CIs surrounding the corrected mean alpha in meta-analysis were slightly less accurate than those in single study. This was in part due to the more precise (or narrower) confidence width in meta-analysis than single study, and hence the coverage of the ideal alpha value became more problematic. Better procedures for constructing the CIs are needed especially when the items are dichotomously scored.

## Conclusion

The results showed that the uncorrected alphas and the associated CIs could be highly inaccurate in either a single or meta-analytic research situation. In this sense, the alpha values reported in the published articles may be larger than researchers thought; this affects the accuracy of reliability evaluation of a test or scale. Alternatively, the results showed that the Case I and III correction procedures yielded accurate alpha estimates. Moreover, the bootstrap CIs constructed for the corrected alpha led to good results in single study and reasonable results in meta-analysis. In sum, these procedures can be summarized in Table 23.

Table 23. *Summary of the procedures for correcting the alpha for range restriction and for constructing the confidence interval.*

| Type of study | Item response | Point estimate | Confidence interval |
|---|---|---|---|
| Single | Continuous | Case I | BSI |
| | | Case III | BSI |
| | Dichotomous | Case I | BPI |
| | | Case III | BSI |
| Meta-analysis | Continuous | Case I | BSI |
| | | Case III | BSI |
| | Dichotomous | Case I | BSI |
| | | Case III | BSI |

The next chapter considers how to apply the corrected procedures in a real single and meta-analytic study using reliability assessment data from the Spence’s Children Anxiety Scale (Spence, 1997). After that, discussion, conclusions, and recommendations of the findings are presented.

# Chapter 5 – A Real Study: Spence Children's Anxiety Scale (SCAS; Spence, 1997)

This section uses data from a real study (Spence, 1997)—which developed a scale which measures children's anxiety (i.e., Spence Children's Anxiety Scale; or SCAS)—to demonstrate how to use the (Case I) correction procedures for reliability. In the literature, anxiety disorder is regarded as one of the most commonly diagnosed psychological disorders for children (Merikangas & Avenevoli, 2002). If anxiety disorders are not treated properly, they tend to persist and interfere with daily functioning (Saavedra & Silverman, 2002). Anxiety is a subjective cognitive and emotional experience, and it can be measured by using a self-report scale.

Spence Children's Anxiety Scale (SCAS; Spence, 1997) is regarded as one of the most commonly used self-report scales. It was developed to measure the six anxiety dimensions specified in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). These dimensions include separation anxiety disorder (SAD; sample item: "I worry about being away from my parents"), social phobia (SP; "I feel scared when I have to take a test"), obsessive-compulsive disorder (OCD; "I get bothered by bad or silly thoughts or pictures in my mind"), panic attack and agoraphobia (PAA; "My heart suddenly starts to beat too quickly for no reason"), physical injury fears (PIF; "I am scared of the dark"), and generalized anxiety disorder (GAD; "I worry about things"). SCAS contains 38 anxiety symptom items and six positive filler items to reduce negative response bias. Children are asked to report the frequency of each item on a 4-point scale (never, sometimes, often, and always). The conceptual and theoretical framework of SCAS has been discussed in detail (e.g., Spence, 1997, 1998; Li, Chan, & Au, 2011b; Nauta et al., 2004; the *Spence Children's Anxiety Scale* website, 2010).

Spence (1997, 1998) provided the psychometric properties of SCAS with a normative sample, which consisted of 698 children in six urban primary schools in Australia. She evaluated the reliability of each dimension based on Cronbach's coefficient alpha. Specifically, the alpha was .70 for separation anxiety disorder, .70 for social phobia, .73 for obsessive-compulsive disorder, .82 for panic attack and agoraphobia, .60 for physical injury fears, and .73 for generalized anxiety disorder. In addition, she provided the mean and standard deviation (SD) of each item, as well as the variance-covariance matrix for the 44 items (see Table A2 in Spence, 1997).

The data presented in Spence (1997, 1998) were based on a sample of typical children, who attended ordinary primary schools. Good psychometric properties (e.g., coefficient alpha) have been documented with the original validation samples in Australia (Spence, 1997, 1998). However, the psychometric properties appear to vary to some extent across cultures (Whiteside & Brown, 2008). They differed, for example, between Japanese and German samples (Essau, Sakano, Ishikawa, & Sasagawa, 2004), and between Hellenic (i.e., Greek-speaking) and other cultural groups (e.g., Australian, Japanese, German; Mellon & Moutavelis, 2007) (Li et al., 2011b). One possible reason for this difference is related to the sample the researchers used to evaluate the psychometric properties. Some studies used a sample of ordinary children in primary schools, whereas other studies examined the properties according to a clinical sample, which tends to be a restricted sample of the general population.

## Single Study

**Mellon and Moutavelis (2007).** Mellon and Moutavelis (2007) evaluated the psychometric properties of SCAS according to a sample of Hellenic (Greek-speaking) children. They found that the alphas of some of the anxiety dimensions were noticeably smaller than those reported in Spence’s (1997, 1998) validation study (Table 1; Mellon & Moutavelis, 2007). For instance, the alpha was .56 for obsessive-compulsive disorder (OCD; .73 in Spence’s study), and.78 for panic attack and agoraphobia (PAA; .82 in Spence’s study). These findings suggest that their sample may be restricted in range compared with the sample in Spence’s normative study. Moreover, the SDs support this claim. They were 2.8 for OCD and 3.9 for PAA in Mellon and Moutavelis’ (2007) study. In comparison, the SDs were 3.67 and 4.24, respectively, in Spence’s study.

The Case I corrected coefficient alphas for Mellon and Moutavelis’ study are presented in Table 24, and they are found to be .74 for OCD and .81 for PAA, respectively. These values are highly comparable with those in Spence’s (1998) normative sample (i.e., .73 and .82). This means that the reduced alpha values were probably due to the artifact of range restriction. If the authors could recruit a more heterogeneous sample, they would obtain the comparable reliability levels to those in the normative sample. Because the selection variable Z was given in Mellon and Moutavelis’ study, the Case III corrected alpha cannot be estimated. Note that the standard errors of measurement (SE) are quite comparable between the two studies (i.e., OCD: 1.89 in Spence and 1.86 in Mellon and Moutavelis; PAA: 1.80 in Spence and 1.83 in Mellon and Moutavelis), and this satisfies the assumption of equal standard errors between the two samples.

Table 24. *Coefficient alphas for the dimensions OCD and PAA in Spence (1997) and Mollen and Moutavelis (2007).*

| Dimension | Study | SD | Type | Alpha |
|---|---|---|---|---|
| OCD | Spence | 3.64 | Normative | .73 |
|  | Mellon & Moutavelis | 2.80 | Uncorrected | .56 |
|  |  |  | Case I | .74 |
| PAA | Spence | 4.24 | Normative | .82 |
|  | Mellon & Moutavelis | 3.90 | Uncorrected | .78 |
|  |  |  | Case I | .81 |

**Li, Lau, and Au (2011)**. The aforementioned section evaluates only the Case I corrected alpha in a published article. In conducting a research study, researchers should have the scores, and hence they can use the non-parametric bootstrap procedure to construct the confidence intervals (CIs). This dissertation uses one of our studies (Li et al., 2011b)—which evaluated SCAS in a Hong Kong Chinese sample—to estimate the Case I corrected alpha and the associated bootstrap CIs. Likewise, it is not easy to identify the selection variable *Z* in this example, and hence the Case III correction procedure is not demonstrated here.

For the purpose of illustration, I selected a homogenous group of 10-year-old children (*N* = 62), and calculated the uncorrected and Case I corrected alphas, and CIs, as displayed in Table 25. The coefficient alpha for OCD was only .65 (.73 in Spence, 1997). When I used the correction procedures, the Case I corrected alpha was .73, which is the same as in Spence's study. In addition, the BSI, BPI, and BCaI are presented in Table 25 for researchers who seek to understand the precision of the alpha estimate. The Case I bootstrap CIs are more precise (narrower) than the restricted CI, thereby providing more accurate evaluation of the associated sampling error.

Table 25. *Coefficient alpha for the dimension OCD in Spence's (1997) and Li et al. (2011b).*

| | Type | SD | Alpha | BSI | BPI | BCaI |
|---|---|---|---|---|---|---|
| Spence | Normative | 3.64 | .73 | n.a. | n.a. | n.a. |
| Li et al. | Uncorrected | 3.23 | .65 | (.45, .81) | (.44, .76) | (.45, .77) |
| | Case I | | .73 | (.67, .79) | (.67, .79) | (.66, .78) |

**Meta-analysis**

This section presents an application of the correction procedure to a meta-analytic study of SCAS coefficient alphas. Note that this section focuses on the statistical aspect of the meta-analytic technique, and hence it demonstrates the possibility of using the correction procedure in meta-analysis. Hence the true mean alpha value of the scale needs further investigation. Moreover, due to the data complexity in a meta-analytic study, it is quite subjective for a meta-analyst to determine whether some of the alphas need adjustments.

A small-scale meta-analytic study was conducted (see Table 26). To determine which studies had used the SCAS, I searched the PsycINFO database using the keywords Spence's children anxiety scale and SCAS between 1995 and June 2012. Conceptual and theoretical papers were eliminated. The literature search resulted in six studies that used the SCAS with eight independent samples. The dimension *physical injury fears* (PIF) was chosen for illustrative purposes. Note that Mellon and Moutavelis' (2007) study discussed in the previous section was excluded because it reported the alpha of seven items (rather than five) based on their exploratory factor analysis. This alpha included two items from another dimension, PAA, and hence its value could not be synthesized in the meta-analysis.

Table 26. *Previous studies that reported statistics for the physical injury and fear (PIF) dimension*.

| Study | Sample | Age group | Region | SD | Alpha | N |
|---|---|---|---|---|---|---|
| Spence (1997) | Community | 8 - 12 | Australia | 2.68 | .60 | 698 |
| Li et al. (2011b) | Community | 6 - 11 | Hong Kong | 3.55 | .63 | 207 |
| Spence et al. (2003) | Community | 13 - 14 | Australia | 2.34 | .60 | 875 |
| Muris et al. (2002) | Regular School | M = 15.1 | Belgium | 2.20 | .54 | 521 |
| *Nauta et al. (2004) | Clinic | 6 - 18 | Australia, Netherlands | 2.30 | .58 | 261 |
| Whiteside 2008 (sample 1) | Community | 9 - 18 | US | 2.40 | .53 | 82 |
| *Whiteside 2008 (sample 2) | Clinic | 9 - 18 | US | 2.10 | .47 | 80 |

* sample suspected of range restriction

In the literature on children’s anxiety, participants are often recruited from two populations: general and anxious groups. The majority of the studies collected data from a regular school or community sample, which represents a population of typical children. By contrast, other studies appear to have recruited children from a clinical sample. Generally, a clinical sample is a restricted sample of the general population, and hence the SD and alpha may be biased. If one performs a *bare-bone* meta-analysis (i.e., a strategy without considering any range-restriction artifacts, so called by Hunter and Schmidt, 2004), the uncorrected mean alpha is found to be .586, and the 95% CI is [.560, .610]. On the other hand, if one follows Hunter and Schmidt’s method of adjusting the restricted alpha for range restriction, she or he can obtain a sample estimate of the unrestricted SD from a large sample or technical manual. In this case, the pooled SD estimated from the remaining studies (i.e., 2.50) or the SD from the normative sample in Spence’s (1997) study (i.e., 2.68) can be used. By using the first pooled SD, the two corrected alphas become .645 for Nauta et al.’s (2004) sample and .626 for Whiteside’s (2008) sample 2. The mean corrected alpha becomes .596 with the 95% CI [.572, .620], which may indicate that the reliability level is more adequate for the general population of children and adolescents. The bootstrap CIs appear to be more adequate as evidenced by the Monte Carlo results. Specifically, the 95% BSI is [.571, .621], BPI is [.568, .619], and BCaI is [.570, .620]. Although the difference between the two methods appears to be small in this example, it can be more substantial when one includes more samples in a typical meta-analysis, which often includes more than 20 study samples.

Details are shown in Table 27. Note that the standard errors of measurement are comparable between Spence's and Whiteside's studies, i.e., 1.69 for Spence's sample, and 1.65 and 1.53 for Whiteside's samples.

Table 27. *The uncorrected and corrected mean alphas and their bootstrap CIs*

| | Mean alpha | BSI | BPI | BCaI |
|---|---|---|---|---|
| Uncorrected meta-analysis | .586 | [.560, .609] | [.553, .601] | [.550, .600] |
| Bias-corrected meta-analysis | .596 | [.571, .621] | [.568, .619] | [.570, .620] |

# Chapter 6 – Discussion and Conclusions

Range restriction has long been a common methodological problem in social sciences research, and it often leads to a downward-biased estimate of a statistic (e.g., Pearson’s correlation *r*, reliability coefficient alpha, etc.). The problem appears to be more prominent in psychological measurement than one thinks, because it can sometimes occur without researchers being aware. For instance, one may incorrectly reach a conclusion of an unreliable test without even knowing that the sample used (e.g., college, clinic, etc.) is actually range-restricted.

Although two conventional correction procedures for reliability have been proposed and discussed in the literature (e.g., Gulliksen 1950, 1987), much attention has been focused on the four correction procedures for Pearson’s correlation. Hence, this dissertation sought to examine the accuracy of the two correction procedures for reliability, so as to provide empirical evidence for their usefulness in practice. In addition, researchers may be interested in assessing the CI surrounding the alpha, and this is regarded as the best strategy in many publication manuals (e.g., APA, 2010). Thus, this dissertation also sought to evaluate the performance of the CI surrounding the uncorrected and corrected alphas.

The first Monte Carlo study shows that the uncorrected alpha can be very misleading. In some adverse data situations (e.g., a stringent selection ratio and a large correlation between the selection variable and test), the uncorrected alpha can be as small as .002, which is 99.7% below its true population value. This means that one may mistakenly conclude that a test is highly unreliable without being aware of the problem of range restriction. To adjust for the bias, researchers can use either the Case I or III correction procedure for reliability, depending on whether the unrestricted SD of the variable *Y* or the unrestricted SD of the selection variable *Z* is known. Both methods are found to result in adequate alpha estimates across the simulation conditions, including the selection ratio, number of items, sample size, etc. Moreover, the real world example shows that, when the alphas reported in Mellon and Moutavelis (2007) are corrected for the Case I restriction, they are highly comparable to the values found in Spence's (1997) normative sample. This suggests that the bias-corrected alpha is not only theoretically supported, but it is also practical in a real research situation.

When one intends to make a statistical inference about the bias-corrected alpha, she or he can use the non-parametric bootstrap procedure in order to build the confidence interval (CI). Assuming that one seeks to examine whether there is a significant difference between the reliability level of a test between two populations (e.g., male and female), she or he can construct the CIs surrounding the corrected alpha, if the samples are actually range-restricted. Suppose the 95% CIs are [.71, .78] and [.80, .88], respectively, for the two samples. The researcher can conclude that there is a significant difference at the .05 level. Hence the CI construction for reliability is important for comparing different tests, scoring rubrics, or training procedures for raters or observers (Haertel, 2006). Unfortunately, such an evaluation may be biased when the sample is actually range-restricted; thus researchers are encouraged to construct the CI for the corrected alpha as an alternative. In particular, three bootstrap CIs (i.e., BSI, BPI, BCaI) for Case I and III, respectively, are examined in this dissertation. When the items are continuously scored, the BSI is found to be the most adequate method for both Case I and III. When the items are dichotomously scored, the BPI is recommended for Case I, and the BSI is suggested for Case III. To demonstrate the application of these procedures, I also constructed the 95% BSI, BPI, and BCaI for the uncorrected and bias-corrected alphas for the dimension OCD, as originally investigated in Li et al.'s (2011b) study. The CIs surrounding the bias-corrected alpha appear to be more reasonable, given that they are more precise in terms of their widths, and they are more likely to span the alpha value reported in Spence's (1997) normative sample.

As discussed in Bonett (2010), the confidence width for coefficient alpha in a single study, however, may be too wide for an accurate sampling error evaluation. An alternative is to evaluate the alpha of a test from multiple studies, and this technique is known as meta-analysis. Meta-analysis is a statistical procedure that synthesizes the quantitative findings provided in multiple studies conducted by independent researchers. Rodriguez and Maeda (2006) have proposed and developed a framework specific to the meta-analysis of coefficient alpha, and they called this concept *reliability generalization*. Since then, various studies—e.g., Vassar and Bradley (2010), Vassar and Crosby (2008), and Warne (2011)—have evaluated the mean or typical alpha reliability level of different psychological scales.

Despite its recent popularity, the meta-analysis of coefficient alpha is also biased when the alpha reported in each single study is subject to range restriction. Hence the second Monte Carlo study sought to examine the performance of the uncorrected and corrected mean alpha in meta-analysis. The results show that the smallest uncorrected mean alpha is .095, which is 86.4% below its true value, for continuous responses. Moreover, it can be as small as .037 for dichotomous responses, which is -93.6% below its true value. This means that evaluating the typical reliability level of a test in a meta-analytic study based on the uncorrected mean alpha is undesirable. Rather, both the Case I and III corrected mean alphas can improve the accuracy substantially. In addition, the bootstrap CIs—especially BSI—surrounding the Case I and III corrected mean alphas appear to be more reasonable than the uncorrected mean alpha. The benefits of reporting the bias-corrected mean alpha and its CI have also been shown in the real world example. For instance, some single studies used a clinical sample to evaluate the reliability alphas, and they should be adjusted for range restriction. I found that the bias-corrected mean alpha was .596 with the 95% BSI of (.571, .621); by contrast, the uncorrected mean alpha was .586 with the 95% BSI of (.560, .609). Although the difference between the two alphas was small, one can predict that the discrepancy will be larger when more samples or studies are included in a meta-analysis. Importantly, this example shows that the bias-corrected mean alpha and its CI are practical in real research situations.

In sum, the corrected alpha estimates, as well as the associated bootstrap CIs, have been found to be adequate in both single and meta-analytic research scenarios, thereby providing a useful and trustworthy method of reliability evaluation when a sample is subject to range restriction. The following sections will first discuss the benefits of applying the correction procedures according to the four goals respectively, and will then present the limitations and directions for future research.

### Implications of Goal 1: Evaluating the Corrected Alpha in Single Study

As noted above, the problem of range restriction seems to be more prominent in psychological measurement than one thinks, because it can sometimes occur without researchers being aware. One may incorrectly conclude that a test is unreliable without even knowing that the sample in hand (e.g., college, clinic, etc.) is actually range-restricted. As shown in the Monte Carlo results, one may obtain a severely downward-biased estimate of reliability—i.e., -99.7% for continuous scores and -136.2% for dichotomous scores—when the data conditions are highly challenging (e.g., a stringent selection ratio, or a large correlation between the selection variable and the variable of interest). Even when the data conditions become less stringent, one is expected to obtain a noticeable negative-biased reliability estimate, with an average of -30.9% bias for continuous scores, and an average of -43.1% bias for dichotomous scores.

Fortunately, the aforementioned biases can be adjusted based on the Case I or III correction procedure depending on whether the unrestricted SD of the variable *Y* or the unrestricted SD of the selection variable *Z* is known. Here I note Schmidt’s advice that researchers and practitioners can trace the sources of these estimates from the existing literature (Li et al., 2011a). For instance, “most measures have a manual that contains the SD for a norm sample, which is usually an unrestricted sample. Alternatively, practitioners can find research manuscripts where the SD is presented for a sample that is not suspected of range restriction" (Li et al., 2011a, p. 371). Applying the Case I correction procedure, one can obtain a more accurate alpha with minimal biases, i.e., a mean of -0.4% for continuous responses, and a mean of -2.4% for dichotomous responses. Likewise, the Case III adjusted alpha results in adequate estimates, with a mean of -2.4% bias for continuous responses and a mean of -7.9% bias for dichotomous responses.

## Implications of Goal 2: Constructing the CIs Surrounding the Corrected Alpha in a Single Study

When a researcher seeks to evaluate the sampling error and to make statistical inferences about the alpha, she or he is encouraged to examine the effect of range restriction on the alpha estimate. As shown in the simulation results, given that the point estimate (i.e., approximately the middle point) of the CI is inaccurate with a restricted sample, the constructed CI becomes inadequate, with mean coverage probabilities ranging from .208 to .488 obtained by BSI, BPI, or BCaI (as opposed to the ideal value of .95 or the lenient criterion of .855). Applying either the Case I or III correction procedure in building the CIs can improve the accuracy of estimation. Generally, the mean coverage probabilities range from .859 to .905 given these correction procedures, meaning that the accuracy improves substantially.

### Implications of Goal 3: Evaluating the Corrected Mean Alpha in a Meta-Analysis

As an extension, evaluating reliability based on a single study may not be sufficient. An additional method is to conduct a meta-analysis of coefficient alpha in order to evaluate the typical reliability level of a test or scale; this involves using all of the empirical findings provided in the literature. As in a single study, however, the mean alpha can be misleading when the alphas reported in single studies are subject to range restriction. The second Monte Carlo study shows that the uncorrected mean alpha estimate was found to be highly inadequate. The largest percentage bias was -86.4% (or mean alpha = .095) for continuous responses and it was –93.6% (or mean alpha = .037) for dichotomous responses, and this may lead a researcher to conclude an unreliable test incorrectly. By contrast, the Case I and III corrected mean alphas could adjust for this bias adequately. Both methods could result in a mean estimate which was very close to its true unrestricted value, with percentage biases ranging from -5.9% to 1.8%; this provides a useful method for evaluating the mean or typical alpha reliability level across published studies in a research domain.

### Implications of Goal 4: Constructing the CIs Surrounding the Corrected Mean Alpha in a Meta-Analysis

The problem of constructing the CI for the uncorrected mean alpha has been demonstrated in the second Monte Carlo study, in which the mean coverage probability is close to zero, meaning that the CI cannot span the true alpha value adequately. Therefore, one should correct the biased alpha for range restriction before they are used to estimate the mean alpha value. The benefits of applying the corrections procedures can be seen in terms of their improved coverage probability. For instance, the coverage probability yielded by the Case I method increases to .897 for continuous responses and .816 for dichotomous responses. Likewise, the Case III mean alpha produces a mean coverage of .893 for continuous responses, and .805 for dichotomous responses.

Based on the empirical findings obtained in the two Monte Carlo studies, a summary of how to correct the alpha coefficient for range restriction and construct the associated confidence intervals is presented in Table 28. Researchers and applied users can follow these guidelines to evaluate the alpha level of the scores in their study when the sample is suspected of range restriction. This provides a useful and trustworthy method for examining the reliability level of the scores whenever the correction factors (e.g., the sample estimate of the unrestricted SD) are available.

Table 28. *Summary of the procedures for correcting the alpha for range restriction and building the corresponding confidence interval.*

| Type of study | Item response | Point estimate | Confidence interval |
|---|---|---|---|
| Single | Continuous | Case I | BSI |
| | | Case III | BSI |
| | Dichotomous | Case I | BPI |
| | | Case III | BSI |
| Meta-analysis | Continuous | Case I | BSI |
| | | Case III | BSI |
| | Dichotomous | Case I | BSI |
| | | Case III | BSI |

## Limitations and Directions for Future Research

**Investigating other reliability coefficients under range restriction**. This dissertation provides a comprehensive assessment of coefficient alpha under range restriction in both single and meta-analytic studies. Coefficient alpha is selection because it is one of the most commonly used reliability coefficients in the literature. Moreover, most of the recent studies about reliability generalizations involve meta-analysis of coefficient alpha (Vassar & Bradley, 2010; Vassar & Crosby, 2008; Warne, 2011). Despite its common use, coefficient alpha is only one of the lowest bound estimates of the internal consistency reliability. In fact, some studies (e.g., Sijtsma, 2009) found that it becomes less accurate when its underlying assumption (i.e., congeneric data) is violated. Some researchers, such as Sijtsma (2009) and Revelle and Zinbarg (2009), suggested that researchers should also provide other lower bound estimates (e.g., McDonald’s omega) in addition to coefficient alpha. Future research can extend the current framework to other reliability coefficients.

**Examining the assumption of the correction procedures for reliability**. Some authors have questioned the need to meet the assumption of homoscedasticity (or equal standard errors). As shown in the Monte Carlo results, when the data meet (at least) the congeneric data assumption and undergo a selection process, the standard errors of measurement for the restricted and unrestricted groups appeared to be similar. This means that if the scores meet the congeneric assumption in the original unrestricted sample, the assumption of homoscedasticity may be met. On the other hand, the present Monte Carlo results showed that when the unrestricted sample contains scores that met (at least) the congeneric condition, and then underwent an indirect range restriction process, the standard errors appeared to be similar between the restricted and unrestricted samples. This may provide empirical evidence about the adequacy of the correction procedures. Further research can examine the need for this assumption more comprehensively.

**Evaluating the correction procedures for reliability under different data situations**. As in other simulation studies, some controlled environments may not reflect the complex data situations that exist in real practice. On one hand, the current design follows most other simulation studies in this area assuming that a sample estimate of the unrestricted SD can be estimated. On the other hand, the present study seeks to be more realistic than other studies (e.g., Fife et al., 2012; Li et al., 2011a) in that only a sample estimate of the unrestricted SD from a simulated sample was obtained. It thus differed from other studies which assumed that the true population unrestricted SD is known. For instance, Fife et al. (2012) stated that "[t]his simulation assumes that the unrestricted variance of *Y* is known…otherwise, it could not be corrected" (p. 25), meaning that the true population value could be known in practice for whatever reasons. Further research can investigate the possibility of the published studies in different areas (e.g., educational, governmental, and organizational settings), thus providing a good estimate or proxy of the unrestricted SD, so that researchers can use them for the correction procedures.

Another area of research lies in examinations of the effects of data complexity on the alphas and their CIs associated with real meta-analyses. As noted by Cheung (2008), “data in meta-analyses are usually more complex in terms of conceptualization and data collection than data in primary studies” (p. 195). An example is related to missing data in primary studies, in which only a portion of the samples provides the essential statistics for performing a correction. Another example is where only some but not all of the reported alphas are subject to range restriction. Further studies may be designed to mimic different complex scenarios for meta-analyses, and to evaluate their impacts on the uncorrected and corrected alphas as well as their CIs.

# References

Alexander, R. A., Hanges, P. J., & Alliger, G. M. (1985). Correcting for restriction of range in both *X* and *Y* when the unrestricted variances are unknown. *Applied Psychological Measurement, 9*, 317-323. doi:10.1177/014662168500900310

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, D.C.

Andre, T., & Hegland, S. (1998). Range restriction, outliers, and the use of the graduate record examination to predict graduate school performance. *American Psychologist*, *53,* 574-575. doi:10.1037/0003-066X.53.5.574

Banks, G. C., Batchelor, J. H., & McDaniel, M. A. (2010). Smarter people are (a bit) more symmetrical: A meta-analysis of the relationship between intelligence and fluctuating asymmetry. *Intelligence*, *38,* 393–401. doi: 10.1016/j.bbr.2011.03.031.

Barchard, K. A., & Hakstian, A. R. (1997a). The effects of sampling model on inference with coefficient alpha. *Educational and Psychological Measurement, 57,* 893-905. doi:10.1177/0013164497057006001

Barchard, K. A., & Hakstian, A. R. (1997b). The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivariate Behavioral Research*, *32,* 169-191. doi:10.1207/s15327906mbr3202_4

Beasley , W. H., & Rodgers, J. L. (2009). Resampling methods. In R. E. Millsap A. Maydeau-Olivares (Eds.), The Sage handbook of quantitative methods in Psychology. CA: Sage.

Blixt, S. L., & Shama, D. D. (1986). An empirical investigation of the standard error of measurement at different ability levels. *Educational and Psychological Measurement*, *46*, 545–550. doi:10.1177/0013164486463007

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, *15*, 368-385. doi:10.1037/a0020142

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods*, *15*, 386-397. doi:10.1037/a0019626

Brannick, M. T., Yang, L., & Cafri, G. (2011). Comparison of weights for meta-analysis of *r* and *d* under realistic conditions. *Organizational Research Methods, 14,* 587 - 607. doi:10.1177/1094428110368725

Brennan, R. L. (2006). *Educational Measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322. doi:10.1111/j.2044-8295.1910.tb00207.x

Burke, M. J., Normand, J., & Doran, L. I. (1989). Estimating unrestricted population parameters from restricted sample data in employment testing. *Applied Psychological Measurement, 13,* 161-166. doi:10.1177/014662168901300206

Cass, M. H., Siu, O. L., Faragher, E. B., & Cooper, C. L. (2003). A meta-analysis of the relationship between job satisfaction and employee health in Hong Kong. *Stress and Health*, *19*, 79-95. doi:10.1002/smi.959

Chan, W., & Chan, D. W.-L. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods*, *9*, 369–385. doi:10.1037/1082-989X.9.3.369

Cheung, M.W. L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. *Psychological Methods*, *13*, 182-202. doi:10.1037/a0013163

Christian, M. S., Bradley, J. C., Wallace, J. C., & Burke, M. J. (2009). Workplace safety: A meta-analysis of the roles of person and situation factors. *Journal of Applied Psychology, 94,* 1103-1127. doi:10.1037/a0016172

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of a test. *Psychometrika*, *16*(3), 297-334.

Cronbach, L. J. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. NY: Wiley.

Crook, T. R., Todd, S. Y., Combs, J. G., Woehr, D. J., Ketchen, D. (2011). Does human capital matter? A meta-analysis of the relationship between human capital and firm performance. *Journal of Applied Psychology*, *96*, 443-456. doi:10.1037/a0022147

Cui, Y. & Li., J. C.-H. (2012). Evaluating the performance of parametric and non-parametric procedures of constructing confidence interval for coefficient alpha: A simulation study. *British Journal of Mathematical and Statistical Psychology*, *65*, 467-498. doi:10.1111/j.2044-8317.2012.02038.x

Darlington, R. B. (1998). Range restriction and the graduate record examination. *American Psychologist, 53*(5), 572-573.

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

Elliot, E.S., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, *72*, 218-232. doi:10.1037/0022-3514.72.1.218

Essau, C. A., Sakano, Y., Ishikawa, S., & Sasagawa, S. (2004). Anxiety symptoms in Japanese and in German children. *Behaviour Research and Therapy*, *42*, 601–612. doi:10.1016/j.bbr.2011.03.031.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, *30*, 357-370. doi:10.1007/BF02289499

Feldt, L. S., & Qualls, A. L. (1999). Variability in reliability coefficients and the standard error of measurement from school district to district. *Applied Measurement in Education*, *12*, 367–381. doi:10.1207/S15324818AME1204_3

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*, 444-467. doi:10.1037/1082-989X.10.4.444

Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The assessment of reliability under range restriction: A comparison of $\alpha$, $\omega$, and test-retest reliability for dichotomous data. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164411430225

Flanagan, J. C (1937). A proposed procedure for increasing the efficiency of objective tests. *Journal of Educational Psychology*, *28*(1), 17-21.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532. doi:10.1007/BF02293811

Flores, B. E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega-International Journal of Management Science*, *14*(2), 93–98.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8. doi:10.3102/0013189X005010003

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*, 930-944. doi:10.1177/0013164406288165

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Gulliksen, H. (1987). *Theory of mental tests.* Hillsdale, NJ: Erlbaum.

Guttman, L. (1945). A basis for analyzing test-retest reliability, *Psychometrika*, *10*, 255-282. doi:10.1007/BF02288892

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger

Hakstian, A. R. & Whalen, T. E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika*, *41*, 219-231. doi:10.1007/BF02291840

Henson, R. K. (2001). Understanding internal consistency reliability estimates. A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177-189.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: correcting error and bias in research findings* (2nd ed.). Thousand Oaks, California: Sage.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*, 594-612. doi:10.1037/0021-9010.91.3.594

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109-133. doi:10.1007/BF02291393

Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology*, *95*, 54 – 78. doi:10.1037/a0017286

Le, H., & Schmidt, F. L. (2006). Correcting indirect range restriction in meta-analysis: Testing a new analytic procedure. *Psychological Methods*, *11*, 416-438. doi:10.1037/1082-989X.11.4.416

Lent, R. H., Aurbach, H. A., & Levin, L. S. (1971). Research design and validity assessment. *Personnel Psychology*, *24*, 247-274. doi:10.1111/j.1744-6570.1971.tb02475.x

Li, J. C.-H., Chan, W., & Cui, Y. (2011a). Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. *British Journal of Mathematical and Statistical Psychology*, *64*, 367-387. doi:10.1348/2044-8317.002007

Li, J. C.-H., Cui, Y., & Chan, W. (in press). Bootstrap confidence intervals for the mean correlation corrected for Case IV range restriction – A more adequate procedure for meta-analysis. *Journal of Applied Psychology*. Advance online publication. doi:10.1037/a0029946

Li, J. C.-H., Cui, Y., Gierl, M. J., & Chan, W. (2012). *Bootstrap confidence intervals for range-restricted coefficient alpha*. Poster presented at the 2012 Annual Meeting, American Educational Research Association (AERA), Vancouver, Canada.

Li, J. C.-H., Lau, W.-Y., & Au, T. K.-F. (2011b). Psychometric properties of the Spence Children's Anxiety Scale in a Hong Kong Chinese community sample. *Journal of Anxiety Disorders*, *25*, 584 – 591. doi:10.1016/j.janxdis.2011.01.007

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, *21*, 239–243. doi:10.1111/j.1745-3984.1984.tb01031.x

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

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution free (ADF) interval estimation of coefficient alpha. *Psychological Methods, 12*, 157-176. doi:10.1037/1082-989X.12.4.433

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

Mellon, R. C., & Moutavelis, A. G. (2007). Structure, developmental course, and correlates of children's anxiety disorder-related behavior in a Hellenic community sample. *Journal of Anxiety Disorders*, *21*, 1–21. doi:10.1016/j.janxdis.2006.03.008

Mendoza, J. L., Hart, D. E., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. *Multivariate Behavioral Research, 26*, 255-269. doi:10.1207/s15327906mbr2602_4

Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, *12*, 282-293. doi:10.3102/10769986012003282

Merikangas, K. R., & Avenevoli, S. (2002). Epidemiology of mood and anxiety disorders in children and adolescents. In M. T. Tsuang, & M. Tohen (Eds.), Textbook in psychiatric epidemiology. New York: Wiley-Liss

Mooney, C. Z. (1997). *Monte Carlo simulation* (Sage University Paper series on Quantitative Applications in the Social Sciences, Series No. 07–116). Thousand Oaks, CA: Sage.

Muris, P., Merckelbach, H., Ollendick, T., King, N., & Bogie, N. (2002). Three traditional and three new childhood anxiety questionnaires: their reliability and validity in a normal adolescent sample. *Behaviour Research & Therapy, 40*(7), 753–772.

Nauta, M. H., Scholing, A., Rapee, R. M., Abbott, M., Spence, S. H., Waters, A., et al. (2004). A parent-report measure of children's anxiety: psychometric properties and comparison with child-report in a clinic and normal sample. *Behaviour Research & Therapy, 42*, 813–839. doi:10.1016/S0005-7967(03)00200-6

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Olson, C. A., & Becker, B. E. (1983). A proposed technique for the treatment of restriction of range in selection validation. *Psychological Bulletin, 93,* 137–148. doi:10.1037/0033-2909.93.1.137

Pearson, K. (1903). Mathematical contributions to the theory of evolution-XI: On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, London, Series A, 200,* 1-66. doi:10.1098/rsta.1903.0001

Revelle, W., & Zinbarg, R.E. (2009). Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika, 74,* 145-154. doi:10.1007/s11336-008-9102-z

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322. doi:10.1037/1082-989X.11.3.306

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9*, 99-103.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion reliability: Implications for validation research. *Personnel Psychology*, *55*, 807-826. doi:10.1111/j.1744-6570.2002.tb00130.x

Salgado, J. F. (1998). Sample size in validity studies of personnel selection. *Journal of Occupational and Organizational Psychology*, *71*, 161-164. doi:10.1111/j.2044-8325.1998.tb00669.x

Schmidt, F. L., Hunter, J. E., & Urry, V. E. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*, 473-485. doi:10.1037/0021-9010.61.4.473

Schmidt, F. L., & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540. doi:10.1037/0021-9010.62.5.529

Schmidt, F. L., On, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59*, 281-305. doi:10.1111/j.1744-6570.2006.00037.x

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach’s alpha. *Psychometrika, 74,* 107-120. doi:10.1007/s11336-008-9101-0

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35,* 137 – 167. doi:10.1207/S15327906MBR3502_1

Spearman, C. (1904). “General intelligence” objectively determined and measured. *American Journal of Psychology*, *15*(2), 201–293.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295. doi:10.1111/j.2044-8295.1910.tb00206.x

Spence, S. H. (1997). Structure of anxiety symptoms among children: a confirmatory factor-analytic study. *Journal of Abnormal Psychology*, *106*(2), 280–297.

Spence, S. H. (1998). A measure of anxiety symptoms among children. Behavior Research and Therapy, 36, 545–566. doi:10.1016/S0005-7967(98)00034-5

Spence, S. H., Barrett, P. M., & Turner, C. M. (2003). Psychometric properties of the Spence Children’s Anxiety Scale with young adolescents. *Journal of Anxiety Disorders*, *17*, 605 – 625. doi:10.1016/S0887-6185(02)00236-0

Spence Children’s Anxiety Scale. (2010). *The Spence children’s anxiety scale website*. Retrieved from http://www.scaswebsite.com

Stauffer, J. M., & Mendoza, J. L. (2001). The proper sequence for correcting correlation coefficients for range restriction and unreliability. *Psychometrika*, *66,* 63 – 68. doi:10.1007/BF02296198

Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *The MMPI-2 Restructured Clinical Scales: Development, validation, and interpretation*. Minneapolis, MN2). *An MMPI handbook: Vol. I. Clinical interpretation*. Minneapolis: University of Minnesota Press.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20. doi:10.1177/0013164498058001002

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. doi:10.1007/BF02296146

Vassar, M., & Bradley, G. (2010). A reliability generalization study of coefficient alpha for the life orientation test. *Journal of Personality Assessment*, *92*, 362 – 370. doi:10.1080/00223891.2010.482016

Vassar, M., & Crosby, J. W. (2008). A reliability generalization study of coefficient alpha for the UCLA loneliness scale. *Journal of Personality Assessment*, *90*, 601 – 607. doi:10.1080/00223890802388624

Warne, R. T. (2011). An investigation of measurement invariance across genders on the Overexcitability questionnaire – Two. *Journal of Advanced Academics*, *22*(4), 578 - 593.

Whiteside, S. P., & Brown, A.M. (2008). Exploring the utility of the Spence Children's Anxiety Scales parent- and child-report forms in a North American sample. *Journal of Anxiety Disorders*, 22, 1440-1446. doi:10.1016/j.janxdis.2008.02.006

Yang, H., Sackett P. R., & Nho, Y. (2004). Developing a procedure to correct for range restriction that involves both institutional selection and applicants' rejection of job offers. *Organizational Research Methods, 7*, 442-455. doi:10.1177/1094428104269054

Ziegler, M., Dietl, E., Danay, E., Vogel, M., & Bhner, S. (2011). Predicting training success with general mental ability, specific ability tests, and (un)structured interviews: A meta-analysis with unique samples. *International Journal of Selection and Assessment*, *19*, 170 – 182. doi: 10.1111/j.1468-2389.2011.00544.x